

Abstract of “Advancing Attention Mechanisms in Multimodal Deep Learning Models” by Michal Golovanevsky, Ph.D., Brown University, May 2026.

Multimodal models are central to modern Artificial Intelligence systems, with growing relevance in domains such as healthcare. Their success is often credited to attention mechanisms, which enable rich cross-modal interactions. However, attention scales poorly when more than two modalities are integrated, and its internal functions remain poorly understood. This thesis addresses the scalability and interpretability of attention in multimodal models by introducing novel integration mechanisms, interpretability pipelines, and targeted interventions. In the first chapter, I introduce One-Versus-Others attention, a scalable alternative to self- and cross-attention that reduces computational cost while preserving accuracy in high-modality settings such as clinical data integration. In the second chapter, I introduce NOTICE, a vision–language causal mediation framework that enables the discovery of human-interpretable functions in attention mechanisms, revealing distinct roles for cross- and self-attention. In the third chapter, I analyze the internal competition between memorized priors and new perceptual input in multimodal models, demonstrating that steering vectors can reallocate attention to favor either prior knowledge or new visual evidence. In the final chapter, I examined how specific groups of attention heads induce prompt-copying behavior in vision–language models, leading to systematic prompt-induced hallucinations on numerical and semantic tasks. Through targeted ablation, I showed that these hallucinations can be reduced without retraining, and that the mechanisms by which copying is suppressed differ across models while producing a shared shift toward visual grounding. The findings presented in this thesis deepen our understanding of attention in multimodal models, offering concrete tools for scalability and interpretability. More broadly, they establish foundations for building multimodal systems that are trustworthy and reliable in real-world applications.

Advancing Attention Mechanisms in Multimodal Deep Learning Models

by

Michal Golovanevsky

B.S., California Polytechnic State University, San Luis Obispo, 2020

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

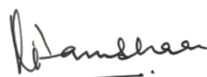
Providence, Rhode Island

May 2026

© Copyright 2026 by Michal Golovanevsky

This dissertation by Michal Golovanevsky is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

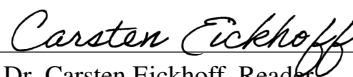
Date 02/10/2026



Dr. Ritambhara Singh, Director

Recommended to the Graduate Council

Date 02/11/2026



Dr. Carsten Eickhoff, Reader

Date 02/11/2026



Dr. Chen Sun, Reader

Approved by the Graduate Council

Date _____

David P. Lindstrom
Dean of the Graduate School

Contents

List of Tables	vii
List of Figures	ix
1 Scaling Multimodal Attention	3
1.1 Introduction & Background	3
1.2 Related Work	5
1.3 Methods	6
1.3.1 One-Versus-Others (OvO) attention	6
1.3.2 Multi-headed OvO Attention	7
1.3.3 Model Complexity	7
1.3.4 Illustration through simulation	8
1.4 Experiments	8
1.4.1 Dataset descriptions	9
1.4.2 Baselines	10
1.4.3 Implementation details	10
1.4.4 Hyperparameter Tuning	10
1.5 Results	11
1.6 Conclusion	13
2 Interpreting Multimodal Attention	14
2.1 Introduction	14
2.2 Related Work	16
2.3 Methods	17
2.4 Results	19
2.4.1 Module-wise Activation Patching	20
2.4.2 Universal Attention Heads Exist Across Corruption Schemes and Tasks	21
2.4.3 Universal Attention Heads Perform Distinct Functions	23
2.5 Conclusion	24

3 Controlling Factual Recall in Attention	26
3.1 Introduction	26
3.2 Related Works	27
3.3 Creating Visual CounterFact	29
3.4 Methods	30
3.4.1 Early Decoding	30
3.4.2 Pixel Versus Prior Steering	31
3.5 Results	31
3.5.1 MLLMs are Distracted By Counterfactual Images	31
3.5.2 Localizing Visual Perception Shifts through Early Decoding	32
3.5.3 Controlling World Knowledge Associations with PvP Steering Vectors	34
3.5.4 Impact of Prompting and Steering Vectors on Attention Mass	35
3.6 Conclusion	37
4 Reducing Hallucinations through Attention Heads	38
4.1 Introduction	38
4.2 Related Work	40
4.2.1 Prompt-Induced Hallucinations in LLMs	40
4.2.2 Modality Conflicts in VLMs	40
4.3 Prompt-Induced Hallucinations in Object Counting	41
4.3.1 Experimental Setup	41
4.3.2 PIH at Different Object Counts	41
4.3.3 PIH Increase as Confidence Decreases	42
4.4 Locating PIH Mechanisms	43
4.4.1 Method: Attention Head Knockouts	43
4.4.2 Attention Knockouts for PIH Mitigation	43
4.4.3 Early-Layer LM Heads Mediate PIH	44
4.5 Generalization Beyond Counting	45
4.6 Analyzing the Function of PIH-heads	45
4.6.1 Ablation Impacts Copying Form	46
4.6.2 PIH-Ablation Shifts Attention Mass to Image Tokens	47
4.6.3 PIH Heads Mediate Model-Specific Copying Mechanisms	48
4.7 Conclusion	49
5 Conclusion and Outlook	51
Appendix	70
A.1 Computational Complexity Analysis for Multimodal Integration Schemes	70
A.1.1 Early Fusion	70
A.1.2 Cross-modal Attention	71

A.1.3 One-Versus-Others (OvO) Attention Complexity	72
A.2 OvO Simulation Dataset Details	73
A.3 Compute Resources for OvO Attention	73
A.4 Significance Testing for OvO Attention	73
A.5 NOTICE VQA-Prompts	74
A.6 NOTICE Datasets Details	75
A.7 BLIP and LLaVA Performance on NOTICE Datasets	75
A.8 SIP with Stable Diffusion	76
A.9 NOTICE Result Details	76
A.9.1 SIP vs. Gaussian Noise Image Corruption.	76
A.9.2 Module-wise Activation Patching	76
A.10 LLaVA Self-Attention Patching Details	80
A.10.1 Comparing Cross-Attention and Self-Attention Head Functions	82
A.10.2 Cross-Attention Head Function Classes	83
A.11 Creating Visual CounterFact	85
A.12 Visual Counterfact Dataset Statistics and Examples	87
A.13 PvP Attention Mass Details	88
A.14 PvP Early Decoding Details	90
A.15 PIH Experimental Details	91
A.15.1 Accuracy on CountBench	91
A.16 PIH with Large Offsets	92
A.17 Base PIH Probabilities for All Models	92
A.18 PIH Attention Head Knockout Details	93
A.19 Full List of PIH-heads	93
A.19.1 Attention Patterns of PIH-heads	94
A.20 Output Examples After PIH Knockouts	96
A.21 PIH Color Task	96
A.22 Layer-wise Attention Mass Shifts After PIH Ablation	97

★ Each chapter of this thesis has been published elsewhere.

List of Tables

1.1	Per-layer complexities of model paradigms. Let k be the number of modalities, n the feature length of a modality, and d the representation dimension.	8
1.2	MIMIC IV+CXr results. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 93.73% compared to self and cross-attention.	11
1.3	TADPOLE results. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.45% compared to self and cross-attention.	12
1.4	eICU results. We report the average of 10 random seeds for AUROC and AUPRC, along with standard deviations. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.12% compared to self and cross-attention.	12
2.1	Universal attention heads in LLaVA and BLIP, identified as having consistently high logit differences across tasks and modality corruptions. In LLaVA, most critical self-attention heads serve multimodal or vision-specific roles. In contrast, BLIP’s dual-stream design results in distinct vision, text, and multimodal heads, supporting more modular cross-modal integration.	22
3.1	Accuracy (%) for color and size tasks under “this” (e.g., “What color is this strawberry?”) and “most” (e.g., “What color are most strawberries?”) questions with counterfactual (CF) and world knowledge (WK) images. Models perform well when grounded in the current image, but accuracy drops sharply in the “most + CF” setting, indicating that MLLMs are overly influenced by misleading visual input.	33
3.2	Flip statistics for size and color attributes with counterfactual images and prompts designed to elicit world knowledge responses. A “flip” occurs when the initially less probable response later exceeds the alternative by at least 5%.	33
3.3	Flip performance (%) under Color and Size tasks for two directions (WK $\rightarrow$ CF, CF $\rightarrow$ WK). Layer ranges indicate where flips occur.	34

4.1 Response accuracies on baseline and misaligned prompts before and after PIH ablations.	
The baseline prompt is “How many [objects] are there in the image?”. Misaligned prompts	
take the form “Describe the $N + k$ [objects] in the image” ($k \in \{1, 2, \dots, 5\}$) with results	
averaged across all k . Lower prompt match indicates reduced PIH, while a higher true-count	
match reflects stronger correction toward visual evidence. Prompt and true-count matches do	
not sum to 100%, as some responses contain no numerical value.	43
4.2 Distribution of PIH and no PIH cases before and after PIH attention head ablation for	
the color task and semantic response types (%). NC, FC, SC, and EC denote no-, format-,	
soft-, and exact-copying, respectively. Correct denotes the ground-truth image color, while	
incorrect denotes the color in the misaligned prompt. Summed rows highlight the shift from	
PIH to image-grounded behavior after ablation.	45
4.3 Definition of copying forms under misaligned prompts for count and color tasks.	46
4.4 Response format and correctness before and after PIH attention head ablation. Digit	
responses indicate copying format used in the prompt, while word responses correspond to	
spelled-out number expressions.	47
A1 Average runtimes for different tasks and model types using one NVIDIA GeForce RTX 3090	
GPU.	74
A2 Results of t-tests comparing the performance metrics between OvO attention and the next	
best-performing multimodal models across different datasets.	74
A3 Comparison of BLIP and LLaVA accuracy (%) by dataset.	76
A4 Universal cross-attention heads identified via head patching.	82
A5 Distribution of object categories in the dataset.	88
A6 Max change in image attention mass (Δ) under intervention and prompt changes for each	
model and task.	90
A7 Details of evaluated models, including underlying language models, vision encoders, sizes,	
and Hugging Face model paths.	91
A8 Performance across different numbers of objects on CountBench.	91
A9 Top-10 attention heads per model ranked by individual knockout success. Success denotes	
the percentage of samples switching from PIH to the correct image-grounded count when the	
head is ablated.	93
A10 Global mean success rates (%) under different PIH head selection strategies. Best-m denotes	
ablating the top-, PIH heads ranked by individual knockout success.	94
A11 Qualitative examples of generations before and after PIH attention head knockout. Before	
intervention, models not only copy the incorrect count implied by the misaligned prompt, but	
also hallucinate non-existing details. After ablating PIH heads, generations more consistently	
reflect the true number of objects present in the image across all three models.	96

List of Figures

1.1	Integration scheme comparison. (a) Early fusion to self-attention with scaled dot product attention [152], and (b) Pairwise cross-attention integration with scaled dot product attention [152]. (c) Our proposed method, One-Versus-Others (OvO), does not rely on pairwise interactions or long concatenated sequences but rather captures all modalities in a single attention score. A modality embedding is represented by m_i and W is a learnable parameter (see Section 1.3.1).	4
1.2	The impact of using OvO attention to fuse simulated data. Using FLOPs as a measure of compute, we demonstrate that OvO grows linearly with respect to the number of modalities, while self and cross-attention grow quadratically.	8
2.1	NOTICE applied to SVO-Probes, MIT-States, and Facial Expression Recognition. NOTICE involves creating Semantic Image Pairs for image corruption and Symmetric Token Replacement for text corruption.	15
2.2	Activation Patching using SIP corruption. The image of the puppy is the clean image, I , and the goat is the corrupt image, I^* . Patching the correct answer token “puppy” at $M_I(I, T)$, from the clean image to the “puppy” token at $M(I^*, T)$ creates the patched states $M'(I^*, T)$ shown as orange diamonds.	16
2.3	Module-wise activation patching results for BLIP and LLaVA on “objects” from SVO-Probes. We visualize the restoration probability after patching for MLP, self-attention, and cross-attention layers in the image-grounded text-encoder. The y-axis denotes which token we patch, and the x-axis denotes which layer we patch.	18
2.4	Module-wise activation patching results for SIP, and Gaussian-Noise corruption on SVO-Probes on BLIP. SIP corruption produces activation patterns that align with Stable Diffusion results and highlight the importance of middle layers, while Gaussian noise fails to reveal meaningful attention layers, emphasizing the effectiveness of SIP for probing vision-language models.	20
2.5	Logit difference demonstrating the impact of patching the correct answer for each LLaVA self-attention head and BLIP cross-attention head on the SVO-Probes, MIT States, and the Facial Expressions datasets. Many key attention heads overlap in importance across both modalities.	21

2.6	Universal cross-attention heads in BLIP implement object detection and object suppression. We visualize the average of all attention heads compared to the universal heads performing each function.	23
2.7	Universal heads in BLIP and LLaVA perform outlier suppression. We visualize the average of all attention heads compared to the universal heads performing outlier suppression.	24
3.1	Visual CounterFact. A new benchmark to study how VLMs utilize world knowledge compared to visual inputs. (Left) images created using color relations, (right) images created using size relations.	29
3.2	Pixel Versus Prior steering vectors are created by contrasting representations of prompts that emphasize pixel-level information (“this”) versus priors (“most”), using the last hidden state.	32
3.3	Early decoding results on LLaVA-Next show a conflict between answering “world knowledge” using priors or answering “counterfact”.	34
3.4	Effect of prompt changes and interventions on attention mass across layers for LLaVA-Next in the color and size tasks. Solid lines show changes when applying the steering vector; dashed lines show the effect of modifying the prompt. Green and purple lines represent attention shifts toward image and text tokens, respectively. The red shaded region highlights the layers where the intervention was applied (corresponding to Table 3.3). We see that intervention has a much stronger effect than changing the prompt.	35
3.5	Impact of intervention on attention distribution. We observe more attention on text before intervention, and see a shift in attention towards the image and away from the text after intervention.	36
4.1	Example of prompt-induced hallucination (PIH) in numerical reasoning. The model correctly answers the baseline counting question. However, it conforms to the prompt and hallucinates additional waterlilies when asked to describe a mismatched number of objects. As the mismatch between the prompt and visual evidence grows, the model increasingly favors the image and recovers the correct count.	39
4.2	PIH rates across different numbers of ground-truth objects. Blue lines show the percentage of outputs matching the prompted number for different offsets, k . Orange lines show matches to the true count. As the object count increases, models stop correcting the prompt.	41
4.3	Base probabilities on the object counting task. We plot the probability assigned to the true count N and neighboring counts $N + k$ ($k \in \{1, \dots, 5\}$) given the prompt “How many [objects] are in the image?”. As object counts increase, confidence in N decreases and probability mass spreads to neighboring counts.	42
4.4	Image attention mass at the layer with the largest intervention-induced change for each model. Namely, LLaVA-OneVision (layer 2, $\Delta=0.121$), Qwen-VL (layer 27, $\Delta=0.053$), and Janus-pro (layer 22, $\Delta=0.037$). Gray bars indicate the baseline attention mass for image tokens, while green bars show an increase after PIH-knockout.	48

4.5 Impact of knockout on the probability of the correct count in digit (blue) and word form (orange).	
We measure the probabilities given the prompt “Describe the N_{digit} [objects]”. The knockout effects are shown in green (increase) and red (decrease).	49
A1 Successful examples of images generating using <i>stable-diffusion-2-inpainting</i> that showcase corrupt images close enough to the clean image yet different in (1) Subject (2) Verb (3) Object	77
A2 Poor examples of images generating using <i>stable-diffusion-2-inpainting</i> that showcase corrupt images which deviate from the clean image in (1) Subject (2) Verb (3) Object	77
A3 Comprehensive view of Gaussian noise effects across all tasks in the SVO-Probes dataset, providing a contrast to SIP effects and highlighting the differential impact of noise on model accuracy and layer activation.	78
A4 Module-wise activation patching results on BLIP for the SVO-Probes dataset.	79
A5 Module-wise activation patching results on LLaVA for the SVO-Probes dataset.	80
A6 Module-wise activation patching results on BLIP for the Facial Expressions dataset.	80
A7 Module-wise activation patching results on LLaVA for the Facial Expressions dataset.	81
A8 Module-wise activation patching results on BLIP for the MIT States dataset.	81
A9 Module-wise activation patching results on LLaVA for the MIT States dataset.	81
A10 Logit difference when patching self-attention heads in LLaVA on the correct multiple choice option in SVO-Probes, MIT States and Facial Expressions. Top row shows SIP image corruption, the bottom row shows STR text corruption.	82
A11 Cross-Attention patterns for universal cross-attention head L0.H11 on a sample of images from each dataset.	83
A12 Cross-Attention patterns for universal cross-attention head L3.H0 on a sample of images from each dataset.	84
A13 Cross-Attention patterns for universal cross-attention head L5.H3 on a sample of images from each dataset.	84
A14 Examples from our dataset. (Left) images created using color relations, (right) images created using size relations.	85
A15 Construction pipeline for the Visual Counterfactual Dataset. We identify typical traits using semantic knowledge sources, retrieve realistic visual exemplars, and apply transformations to create perceptually plausible counterfactuals that conflict with language priors.	86
A16 Attention mass difference across layers for all models on the color task. Solid lines show changes from PvP steering vectors; dashed lines show prompt-only changes. Green represents attention to image tokens, purple to text tokens. Each subplot shows one model and intervention direction (WK $\rightarrow$ CF or CF $\rightarrow$ WK).	89
A17 Attention mass difference across layers for all models on the size task. Solid lines show changes from PvP steering vectors; dashed lines show prompt-only changes. Each subplot shows one model and intervention direction (WK $\rightarrow$ CF or CF $\rightarrow$ WK).	89

A18	Early decoding results on Qwen2.5-VL and Janus-Pro show a conflict between answering “world knowledge” using in-weight memorization or answering “counterfact” using visual perception.	91
A19	PIH rates across different number of ground-truth objects. Blue lines show the percentage of outputs matching the prompted number for different offsets k , while orange lines show matches to the true count. Bias grows with object count, with four objects marking the approximate threshold at which models stop correcting the prompt. Prompt match and true count match do not sum to 100% because some model responses do not contain a number. .	92
A20	Base probabilities on the object counting task. We plot the probability assigned to the true count N and neighboring counts $N + k$ ($k \in \{1, \dots, 5\}$) given the prompt “How many [objects] are in the image?”. As object counts increase, confidence in N decreases and probability mass spreads to neighboring counts, matching the increased susceptibility to PIH.	94
A21	Attention patterns for the top-ranked PIH attention head in three multimodal models, averaged over all samples and computed over the full input-generation sequence. Left: LLaVA-OneVision (L0H3), Center: Qwen-VL (L0H3), Right: Janus-Pro (L0H20). Query tokens are shown on the y-axis and key tokens on the x-axis, with each row summing to 1. Image pad tokens are largely aggregated into a single row and column (<i>IMAGE_PADS</i>), with boundary tokens retained.	95
A22	Impact of knockout on the probability of the ground-truth answer in both digit (blue) and word form (orange) given the prompt “Describe the $[N + k]_{\text{digit}}$ [objects]” for count. Note that when $k = 0$, there is no modality conflict in the prompt.	97
A23	Layer-wise change in attention mass after PIH head ablation. For each model, we plot the change in attention mass to image tokens (blue) and text tokens (orange) as a function of transformer layer. Positive values indicate increased attention allocation relative to the baseline. Across all models, PIH ablation consistently shifts attention from text to image tokens, with LLaVA-OneVision exhibiting the strongest and earliest effect.	97

Introduction

Multimodal learning has emerged as a promising approach, which enables joint learning from multiple modalities of data (e.g., text and images). Combining different modalities allows for a more comprehensive and accurate understanding of tasks such as image and video captioning [168, 132], audio-visual speech recognition [142], sentiment analysis [116], and medical decision support [96].

Multimodal learning has been explored through various backbone methods in machine learning (e.g., Decision Trees) and various Neural Network architectures. While feature-level integration was mostly used in more traditional machine learning algorithms, Neural Networks have allowed for the intermediate fusion of modalities through layers and late fusion at the decision-making stage. However, both fusion paradigms lack a key component - capturing explicit interaction between modalities. For example, in detecting hateful social media posts [147], imaging features help reinforce and ground the textual information and thus lead to more robust decision-making. The success of the Transformer [152] on various NLP tasks and, consequently, the Vision Transformer’s (ViT [24]) success on vision tasks motivated the extension of the Transformer to the multimodal case.

The success of the Transformer [152] in various NLP tasks, and consequently the Vision Transformer’s (ViT [24]) success on vision tasks, motivated the extension of the Transformer to the multimodal case. Multimodal Transformers, such as LXMERT [145], ViLBERT [84], and later BLIP [75], introduced a fusion method that captures interactions between modalities using *cross-attention*. Cross-attention operates by using the representation of one modality as queries and the representation of another modality as keys and values, thereby enabling each modality to attend to and incorporate information from the other. In parallel, another popular approach used in models such as VisualBERT [77], VL-BERT [143], and later LLaVA [80], used early fusion, where vision and language inputs are concatenated early to learn multimodal through *self-attention*. In early fusion, self-attention is applied jointly over the concatenated sequence of tokens from all modalities, allowing each token to attend to every other token regardless of modality. These approaches established attention as the central mechanism for modeling cross-modal interactions.

Despite this progress, fundamental challenges remain for attention-based multimodal models. A central issue is **scalability**: Both self-attention and cross-attention incur quadratic computational cost with respect to the number of modalities, which limits their practicality in settings that go beyond the typical two modalities of vision–language tasks. While popular benchmarks rarely exceed four modalities (e.g., images, text, audio, and video), other domains demand far richer integration. Healthcare exemplifies this challenge: a single clinical task may involve radiology scans, pathology slides, genetic screenings, genomic profiles, and textual clinical

notes, all of which must be fused to make reliable predictions. As the number of modalities increases, the quadratic cost of attention creates a severe computational bottleneck, restricting the applicability of current approaches in high-modality domains.

Parallel to scalability, a second major challenge is **interpretability**. Even within the widely studied vision–language setting, the internal behavior of attention is not well understood. While probing methods have provided some insights, they are ultimately limited: they rely on correlations between model representations and external labels, and cannot identify causal mechanisms. In language models, mechanistic interpretability (MI) has shown that attention can be decomposed into circuits responsible for specific functions such as factual recall and syntactic processing [156, 91]. Yet in multimodal architectures, we do not know whether self- and cross-attention similarly specialize into human-interpretable roles. In particular, it is not known whether multimodal attention mechanisms implement distinct functions in self- and cross-attention, or whether they play interchangeable roles in multimodal integration. Addressing this gap is especially important as recent architectures increasingly adopt early fusion, integrating vision directly into large language models where self-attention dominates the interaction paradigm.

Beyond questions of scalability and interpretability, a further challenge lies in **controlling attention**. Even if we can describe how attention mechanisms operate, we lack tools for steering them in ways that reliably influence model behavior. This limitation becomes clear in cases where attention mediates a competition between different information sources, such as memorized priors versus new and conflicting perceptual input. Understanding how to reallocate attention in such settings is critical both for interpretability and for mitigating systematic failure modes. In general, failures often arise when attention disproportionately favors one source of information—such as prior knowledge or linguistic context—over perceptual evidence. A particularly clear case is confirmation bias in vision–language models, where prompt-induced tokens are propagated through attention heads and override visual grounding, leading to hallucinations in numerical reasoning. Developing methods to control attention dynamics is therefore essential for building multimodal systems that are not only scalable and interpretable but also reliable.

In response to these challenges, this thesis develops methods to scale, interpret, and control attention in multimodal models. The contributions are as follows:

- **Chapter 1** introduces One-Versus-Others attention, a scalable integration mechanism that reduces computational cost while preserving accuracy in high-modality domains such as healthcare.
- **Chapter 2** presents NOTICE, a causal mediation framework for vision–language models, and uses it to discover human-interpretable functions implemented by attention heads, revealing systematic differences between self- and cross-attention.
- **Chapter 3** analyzes how attention mediates the competition between memorized priors and perceptual input in multimodal models, demonstrating that steering vectors can reallocate attention to control reliance on knowledge versus evidence.
- **Chapter 4** demonstrates that prompt-induced hallucinations arise from specific attention head groups in vision–language models and that ablating these heads reduces hallucinations across tasks.

Chapter 1

Scaling Multimodal Attention

In this chapter, we present findings from our Pacific Symposium on Biocomputing 2025 paper, *One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data* [38]. Multimodal models have become increasingly important as they surpass single-modality approaches on diverse tasks ranging from question-answering to disease diagnosis. Despite the importance of multimodal learning, existing efforts focus on vision-language applications, where the number of modalities rarely exceeds four (images, text, audio, video). However, data in healthcare domain, may include many more modalities like X-rays, PET scans, MRIs, genetic screening, genomic data, and clinical notes, creating a need for both efficient and accurate data integration. Many state-of-the-art multimodal models rely on cross-attention or self-attention for effective data integration, which do not scale well for applications with more than two modalities. The complexity per layer of computing attention in either paradigm is, at best, quadratic with respect to the number of modalities, posing a computational bottleneck that impedes broad adoption. To address this, we propose a new attention mechanism, One-Versus-Others (OvO) attention, that scales *linearly* with the number of modalities, thus offering a significant reduction in computational complexity compared to existing multimodal attention methods. Using three clinical datasets with multiple diverse modalities, we show that our method decreases computation costs while maintaining or increasing performance compared to popular integration techniques. Across all clinical datasets, OvO reduced the number of required floating point operations (FLOPs) by at least 91.98%, demonstrating its significant impact on efficiency and enabling multi-modal predictions in healthcare.

1.1 Introduction & Background

Multimodal learning enables joint reasoning across diverse data types and has driven advances in applications ranging from image captioning and speech recognition to clinical decision support [96, 36, 47, 168, 132, 142, 116]. A key factor in this success has been the attention mechanism, which explicitly models interactions between modalities. Popular vision-language models such as LXMERT [145] and BLIP [75] employ cross-attention, while models like VisualBERT [77] and LLaVA [81] rely on early fusion through self-attention.

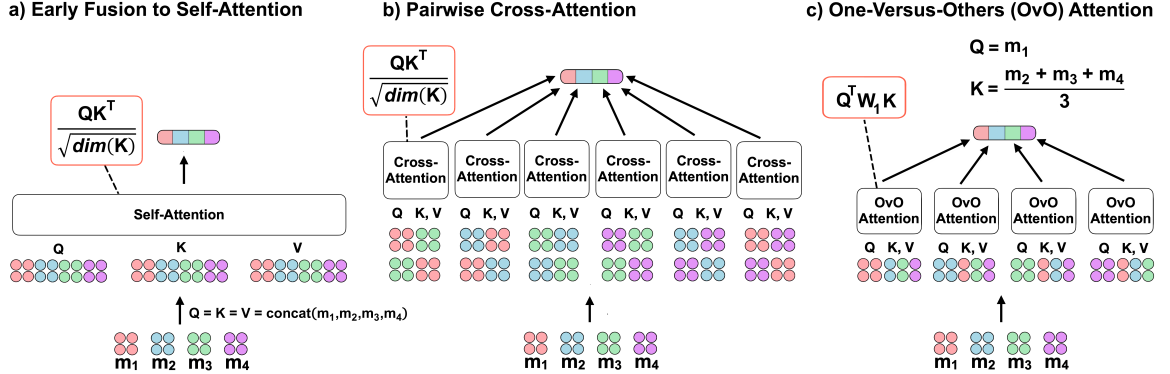


Figure 1.1: **Integration scheme comparison.** (a) Early fusion to self-attention with scaled dot product attention [152], and (b) Pairwise cross-attention integration with scaled dot product attention [152]. (c) Our proposed method, One-Versus-Others (OvO), does not rely on pairwise interactions or long concatenated sequences but rather captures all modalities in a single attention score. A modality embedding is represented by m_i and W is a learnable parameter (see Section 1.3.1).

Similar designs have been adopted in healthcare, with models such as MedFuseNet [134] and ARMOUR [82] using cross-attention for medical vision–language tasks, and BioViL-T [6] and MMBERT [68] applying self-attention for disease prediction and report generation.

However, both self- and cross-attention scale quadratically with the number of modalities, posing a critical bottleneck outside of vision–language tasks. In healthcare, for example, a single prediction may require fusing radiology scans, pathology slides, genomic profiles, genetic data, and clinical notes. As the number of modalities increases, the computational burden of attention becomes prohibitive, limiting the applicability of current architectures in high-modality domains.

To address this gap, we propose a new attention mechanism, One-Versus-Others (OvO) attention. OvO attention is calculated by comparing *one* modality against a combined representation of all *other* modalities (hence the name, One-Versus-Others). Our approach significantly reduces computational complexity as it grows linearly with the number of modalities (see Section 1.3.3). Figure 1.1 sketches a four-modality example to demonstrate the difference between our approach (scales linearly) and self-attention/cross-attention (scales quadratically). OvO is a general attention scheme that can be integrated into existing clinical multimodal architectures instead of cross-attention or self-attention.

We first, present a complexity analysis and validate it through a simulated dataset. Our simulation results show scalability gains in an extreme multimodal setting (20 modalities). Next, we use three diverse clinical datasets that vary in modalities, encoder types (pre-trained and not), number of samples, and tasks to show our model’s improved scalability in different clinical multimodal settings. Our results demonstrate that our method dramatically decreases computation costs (offering at minimum a 91.98% reduction in computations), compared to self-attention and cross-attention while maintaining or even exceeding performance.

Overall, OvO is a novel attention scheme for multimodal integration that scales linearly to the number of modalities, enabling the practical application of deep learning models in healthcare, where computational efficiency and accuracy are vital for deployment.

1.2 Related Work

Multimodal attention-based models are increasingly pivotal in clinical decision support systems, paralleling their widespread use in vision-language applications. In the medical domain, these models have shown remarkable utility in diverse scenarios, such as cancer classification [73], biomarker discovery [10, 57], prognosis prediction [131, 138], and more. These applications highlight the versatility and potential of multimodal learning in handling complex and rich medical data. The attention mechanism serves as a core component in these models. Attention measures the similarity among individual representations, like word embedding vectors or, in the multimodal scenario, modality-specific embeddings. Each input embedding can assume one of three roles: (1) Query (Q), representing the current focus of attention when compared against other input embeddings; (2) Key (K), signifying an input embedding being compared to the Query; and (3) Value (V), which contributes to computing the output for the Query.

Commonly, the representations from each modality in the multimodal models are passed through one of two paradigms - early fusion followed by self-attention or fusion through cross-attention. The early fusion group (e.g., Transformer-based models like UNITER [15], VisualBERT [77], LLaVA [81], BioViL-T [6], MedViLL [99], etc.) concatenates the visual embeddings and the textual embeddings as a single input, before passing the inputs through attention (see Figure 1.1 (a)). Given modalities m_1 and m_2 , queries (Q), keys (K), and values (V) are computed from their concatenated sequence (e.g., $Q_{1,2} = \text{concat}(m_1, m_2)$). The final output from a standard Transformer block is denoted by Z , Equation 1.2.1 shows the early fusion paradigm.

$$\begin{cases} Z_{1,2} = \text{Multiheaded Attention}(Q_{1,2}, K_{1,2}, V_{1,2}) \\ Z = \text{Transformer}(Z_{1,2}) \end{cases} \quad (1.2.1)$$

The cross-attention scheme (used in Transformer-based models like ViLBERT [84], LXMERT [145], MedFuseNet [134], MADDi [36], etc.) inputs each modality into its own Transformer, the outputs of which are fed to a cross-modal Transformer (see Figure 1.1 (b)). For such models, the cross-modal interactions are captured in a pairwise manner through cross-attention, where queries (Q), keys (K), and values (V) are computed from the modality inputs (m_1 and m_2), and then the keys and values from each modality are fed to the multi-headed attention block of the other modality. The output, Z , is shown in Equation 1.2.2.

$$\begin{cases} Z_1 = \text{Multiheaded Attention}(Q_2, K_1, V_1) \\ Z_2 = \text{Multiheaded Attention}(Q_1, K_2, V_2) \\ Z = \text{Transformer}(\text{concat}(Z_1, Z_2)) \end{cases} \quad (1.2.2)$$

While the early fusion and cross-attention paradigms could be extended to three modalities, seen in TriBERT [119] and VATT [1], these models face scalability challenges for more than three modalities. Cross-attention methods can leverage joint representations formed from cross-attention but do not scale well to larger numbers of modalities as they are computed in a pairwise fashion. Thus, if there are k modalities, computing pairwise fusion between each pair will result in $\binom{k}{2}$ matrix computations. Moreover, attention is not a symmetric calculation, which means that most commonly, it is computed bi-directionally (e.g., image to text and text to image), leading to an even greater computational burden. Early fusion involves the concatenation of modalities before the Transformer layer, which similarly does not scale well with the number

of modalities. Self-attention is quadratic with respect to sequence length [152], and since early fusion methods concatenate inputs before attention, the computational complexity will increase quadratically as the number of modalities increases (see Section 1.3.3). Furthermore, concatenation is not order invariant, making the ordering of modalities an important consideration, potentially requiring similar bi-directional computations as cross-attention. Our integration method, OvO, addresses the limitations mentioned above in a scalable and domain-agnostic manner.

1.3 Methods

1.3.1 One-Versus-Others (OvO) attention

We propose a new attention mechanism, One-Versus-Others (OvO) Attention, which grows linearly with the number of modalities rather than quadratically, as would be the case for cross-attention or self-attention (see Section 1.3.3). OvO computes attention between one modality at a time with respect to all other modalities. Given modality m_i , which is an embedding obtained from a dedicated encoder (e.g., CNN, ClinicalBERT, etc.) and $i \in 1, 2, \dots, k$ where k is the number of modalities, OvO takes in one modality and computes the dot product against all the other modalities with a weight matrix W_i . W_i is a learnable parameter that can help scale the importance of each attention calculation (see Figure 1.1(c)) and can learn interactions throughout training. The modality embeddings m_i in OvO attention function like queries, while the weighted sum of the other modalities behaves like keys and values, akin to the dot-product attention mechanisms, with the naming choice of m_i adapted for multimodal applications. This dot product The similarity score function, representing the degree of alignment between the chosen modality and others, calculated for modality m_i with respect to a set of other modalities ($m_j : j \neq i$) is shown in Equation 1.3.1. This produces a vector of scores which measure the relevance of m_i with respect to the other modalities. The context vector in OvO for modality m_i , which is a combined representation of information from the other modalities, is shown in Equation 1.3.2

$$\text{score}(m_i, \{m_j : j \neq i\}) = m_i^T W_i \frac{\sum_{j \neq i}^k m_j}{k-1} \quad (1.3.1)$$

$$\text{OvO}(m_i, \{m_j : j \neq i\}) = \text{softmax}(\text{score}(m_i, \{m_j : j \neq i\})) \cdot m_i \quad (1.3.2)$$

In Equation 1.3.2, the softmax is applied across the input dimension of the embeddings, normalizing across the attention scores. The result is then multiplied by the original modality embeddings to compute the final output. We chose to sum over the “other” modalities instead of concatenation because: (1) the concatenation vector will continue to increase in length with the number of modalities, which will result in a less scalable framework; (2) concatenation is not invariant to the order of modalities, which could affect the model prediction, whereas a sum provides position invariance.

Unlike cross-attention and self-attention, OvO provides a more interpretable mechanism for analyzing interactions between multiple modalities. In cross-attention, interactions are captured in isolated pairs (e.g., m_1 compared to m_2 or m_1 compared to m_3), limiting the ability to see how one modality integrates information from all others. Similarly, self-attention condenses modality interactions into a single operation, which can

obscure explicit cross-modal interactions and make it difficult to disentangle their contributions. OvO, however, creates one attention matrix per modality, with each modality interacting with all others through the learnable weight matrix W_i , which ensures a flexible and adaptive scaling mechanism. In future work, W_i will be studied to better understand the relative importance of each modality’s contributions to the final prediction.

1.3.2 Multi-headed OvO Attention

We extend OvO attention to the multi-headed attention framework to directly compare with early fusion through self-attention and pairwise cross-attention. Multi-headed attention allows the model to attend to the input embeddings in different ways simultaneously. This is achieved by splitting the input embeddings into multiple linear projections, each processed independently through a self-attention mechanism. The outputs of each attention head are then combined to obtain the final output of the multi-headed attention layer. Formally, taking the input modality m_i with respect to a set of other modalities ($m_j : j \neq i$), the multi-headed attention layer for OvO attention is defined as follows:

$$\begin{cases} \text{MultiheadedOvO}(m_i, \{m_j : j \neq i\}) \\ = \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O \\ \text{head}_k = \text{OvO}(m_i W_k^{m_i}, \{m_j W_k^{m_j} : j \neq i\}) \end{cases} \quad (1.3.3)$$

Here, h is the number of attention heads, W_k is a learnable weight matrix for the k -th attention head, W^O is a learnable weight matrix that projects the concatenated outputs of the attention heads back to the original dimension, and Equation 1.3.2 defines OvO Attention.

1.3.3 Model Complexity

This section highlights the complexities of the two commonly used paradigms: early fusion followed by self-attention and pairwise cross-attention, as well as our One-Versus-Others (OvO) attention. Table 1.1 summarizes the complexity per layer. Let k represent the number of modalities, n be the feature-length of each modality (assuming equal), and d be the representation dimension of the respective weight matrices. As established in [152], self-attention has complexity of $\mathcal{O}(n^2 \cdot d)$. In the multimodal case, self-attention concatenates modalities before attention, leading to a sequence length of $k \cdot n$, influencing the quadratic term. Thus, the complexity of self-attention is $\mathcal{O}((k \cdot n)^2 \cdot d) = \mathcal{O}(k^2 \cdot n^2 \cdot d)$. Cross-attention computes attention over all pairwise permutations of modalities: ${}_k P_2 = \frac{k!}{(k-2)!} = k(k-1)$. Thus, the number of operations required by cross-attention is $\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d) = \mathcal{O}((k^2 - k) \cdot n^2 \cdot d)$. When focusing on the fastest-growing terms in big O notation, the final complexity per layer simplifies to $\mathcal{O}(k^2 \cdot n^2 \cdot d)$. One-Versus-Others (OvO) Attention requires one attention calculation per modality, making it linear with respect to k . Thus, the complexity per layer for OvO is $\mathcal{O}(k \cdot n^2 \cdot d)$. See Appendix A.1 which provides step-by-step details for the complexity calculations.

Table 1.1: **Per-layer complexities of model paradigms.** Let k be the number of modalities, n the feature length of a modality, and d the representation dimension.

Model	Complexity Per Layer
Self-Attention	$\mathcal{O}(k^2 \cdot n^2 \cdot d)$
Cross-Attention	$\mathcal{O}(k^2 \cdot n^2 \cdot d)$
One-Versus-Others (OvO) Attention	$\mathcal{O}(k \cdot n^2 \cdot d)$

1.3.4 Illustration through simulation

To illustrate the linearity of OvO compared to the other integration paradigms, we simulated 20 artificial modalities. We consider two classes: (1) 20 random feature values that sum up to 1.0, and, (2) 20 random feature values that are each less than 0.15. These classes were created such that the correct label can only be inferred after inspecting all features. For example, 0.14 is less than 0.15, but it could also be a value that adds to 1. For more details on the simulation dataset and how the threshold was chosen, see Appendix [A.2](#).

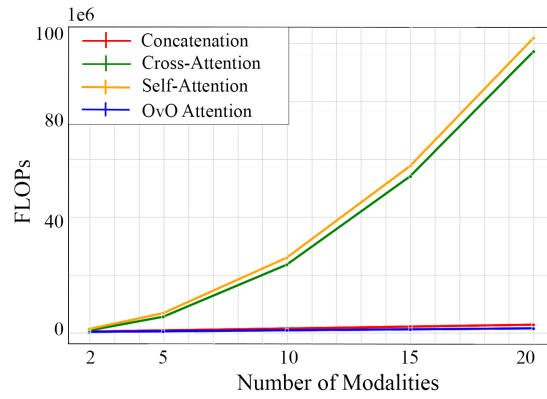


Figure 1.2: **The impact of using OvO attention to fuse simulated data.** Using FLOPs as a measure of compute, we demonstrate that OvO grows linearly with respect to the number of modalities, while self and cross-attention grow quadratically.

Each value was then vectorized by sampling randomly around the selected number, such that each modality is a vector of size 20 rather than a single number, leading to a combined total of 400 features. Overall, the dataset contains 2,000 samples (1,000 for each class). Our constructed simulation dataset tests the scaling capabilities of our method to an extent that real-world datasets do not usually reach.

We examine the computation cost across the three integration methods using 2, 5, 10, 15, and 20 simulated modalities. Most notably, while self-attention and cross-attention grow quadratically with respect to the number of modalities, k , ($\mathcal{O}(k^2 \cdot n^2 \cdot d)$), our method scales linearly ($\mathcal{O}(k \cdot n^2 \cdot d)$), as shown in Figure [1.2](#).

1.4 Experiments

We used three diverse clinical datasets to examine our method against three standard integration techniques: concatenation with no attention (baseline), early fusion with self-attention, and pairwise cross-attention. These

clinical tasks feature a range of rich modalities that, despite their high integration costs, remain essential to solve.

1.4.1 Dataset descriptions

MIMIC-IV and CXR data

MIMIC-IV [65] covers 431K visits for 180K patients admitted to the ICU in the Beth Israel Deaconess Medical Center. MIMIC Chest X-ray (MIMIC-CXR) [64] contains chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at Beth Israel Medical Center. We follow the pre-processing of MedFuse [98] to extract the clinical time-series data from MIMIC-IV along with the associated chest X-ray images in MIMIC-CXR. We further expand the number of modalities by adding a demographics table and discharge notes, resulting in four modalities. We also follow MedFuse in the construction of the phenotyping task. The goal of this multi-label classification task is to predict whether a set of 25 chronic, mixed, and acute care conditions are assigned to a patient in a given ICU stay. This is a 25-class multi-label task with four modalities.

The Alzheimer’s Disease Prediction Of Longitudinal Evolution (TADPOLE) data

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) [113] database provides neuroimaging data, cognitive test scores, biomarker profiles, and genetic information for Alzheimer’s disease (AD), mild cognitive impairment (MCI), and normal patients. We use the processed data from the Alzheimer’s Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge [88]. We focus on a one-time diagnosis prediction task, using the most recent available data point for each patient across all modalities. This ensures that each data sample includes information from the same time point, which aligns with our goal of evaluating cross-modal integration rather than longitudinal progression. We utilize six modalities that have the least missing information per patient: cognitive tests - neuropsychological tests administered by a clinical expert; MRI ROIs (generated from Freesurfer) - measures of brain structural integrity; FDG PET ROI averages - measure cell metabolism, where cells affected by AD show reduced metabolism; AV45 PET ROI averages - measures amyloid-beta load in the brain; demographic information (e.g., age, gender, education); and CSF biomarkers - amyloid and tau levels in the cerebrospinal fluid. The preprocessing provided by TADPOLE turned every modality into a tabular form (including imaging). After removing patients with missing modalities, we had 767 MCI patients, 493 normal patients, and 143 AD patients. Thus, this is a three-class classification task with six modalities.

eICU data

The eICU collaborative database includes data from ICUs across the continental United States between 2014 and 2015 [115]. It consists of tables linked through a patient unit stay ID. For our mortality prediction task, we focus on six tabular modalities: patient, diagnosis, treatment, medication, lab, and apacheApsVar tables. The apacheApsVar table contains numerical variables used to calculate the Acute Physiology Score (APS), an established method within the Acute Physiology Age Chronic Health Evaluation (APACHE) system for

summarizing patient’s severity of illness on ICU admission and predicting outcomes. The patient table includes demographic, admission, and discharge details, and is used to determine mortality status. The diagnosis table lists active diagnoses for each patient, the treatment table includes active treatments, and the medication table contains active medication orders. We extract the features from these tables by one-hot encoding the relevant conditions, treatment types, and drug names, respectively. The lab table includes lab results, with features extracted by summing commonly recorded lab types. Our dataset includes 75,845 unique patients with 93,784 ICU stays, 86,012 recorded as alive and 7,772 as dead. This is a 2-class classification task with six modalities.

1.4.2 Baselines

Our multimodal baselines include a conventional concatenation fusion with no attention, early fusion followed by self-attention, and pairwise cross-attention fusion. The architectures of all models are identical except for their integration stage. For example, since modality-specific encoders can produce different dimension sizes, we add a linear layer before integration to create the same input dimensions. Although this step is not strictly necessary for concatenation, we still add the layer there so that no additional factors influence computation costs and performance. While there are many multimodal Transformers available for the vision-language domain, our focus is on examining the underlying fusion mechanism and creating a general integration paradigm for any application, especially ones outside of vision-language.

1.4.3 Implementation details

For the MIMIC dataset, we follow the established train, validation, and test split in Hayat *et al* [47]. Similarly, for the TADPOLE task, we use the provided data splits but add a constraint that repeating patients cannot appear across data splits to avoid information leakage. In the other datasets, for consistency, we randomly sampled 80% of the data for the training set and 10% each for test and validation sets, as there was not an established split. To evaluate our model against other integration techniques, we use the domain-accepted metrics for each task: For MIMIC and eICU, we use area under the receiver operating characteristic (AUROC) and area under Precision-Recall (PR) curve (AUPRC) as established in past works [47, 136]; For TADPOLE we use the multi-class area under the receiver operating curve (mAUC) and the overall balanced classification accuracy (BCA), as established by the competition creators [88]. For all datasets, we used the number of floating-point operations (FLOPs) as the measure of runtime complexity. FLOPs were measured per sample and reported as the difference between concatenation, the simplest integration setting, and multimodal attention (Δ FLOPs).

1.4.4 Hyperparameter Tuning

Our hyperparameter tuning scheme was consistent for each dataset and each model. For each experiment, we used the evaluation metrics on the validation set to determine the best hyperparameters. We tuned the learning rate ($0.01 - 1 \times 10^{-8}$, dividing by 10 for each interval), batch size (16, 32, 64, 128), epochs (200 epochs with early stopping if validation performance did not increase for 5 epochs), and number of attention heads for OvO, self-attention, and cross-attention models (1, 2, 4, 8, 16). For the neural network encoders, we

tuned the number of linear layers ranging from 1 to 4. Similarly, for the convolutional neural network, we tuned the number of convolution layers ranging from 1 to 4. For compute times and GPU details used for hyperparameter tuning, see Appendix A.3. Lastly, we randomly picked 10 random seeds for every experiment - once the best hyperparameters were picked, ten models initialized with those seeds and parameters were run. Then, using the trained ten models, we evaluated on the test set and took the average of the 10 runs along with the standard deviation, which is reported in Section 2.4.

1.5 Results

Using three real-world clinical datasets, diverse in terms of the number of modalities, feature space, and classification tasks, we demonstrate that our method consistently and drastically reduces computational costs compared to early fusion and pairwise fusion while simultaneously maintaining or enhancing performance. This is demonstrated on one four-modality dataset and two six-modality datasets.

For the four-modality MIMIC task, we used pre-trained ClinicalBERT model for the text modality and fine-tuned it for the unimodal baseline and the multimodal task, separately, ensuring adaptation in each setting. For all other modalities, we used the appropriate neural network architecture (i.e., CNN for images, LSTM for time series, and a multi-layer perception for all tabular data).

Table 1.2: **MIMIC IV+CXr results.** (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 93.73% compared to self and cross-attention.

Model	Modalities	$\downarrow \Delta$ FLOPs	$\uparrow$ AUROC	$\uparrow$ AUPRC
LSTM	Time Series	-	58.8 ± 0.6	28.5 ± 0.4
CNN	Images	-	56.9 ± 0.3	26.7 ± 0.2
Neural Net	Demographics	-	64.1 ± 0.4	32.4 ± 0.3
ClinicalBERT	Text	-	79.3 ± 0.4	58.7 ± 0.3
Concatenation	All	*	82.7 ± 0.6	65.1 ± 1.8
Cross-Attention	All	52,723,712	78.2 ± 2.1	54.1 ± 2.7
Self-Attention	All	67,633,152	78.5 ± 2.0	55.7 ± 3.1
OvO Attention	All	4,227,072	83.6 ± 1.1	66.2 ± 2.6

The results on MIMIC are presented in Table 1.2, clearly demonstrating the scalability and performance advantages of OvO attention. OvO’s 4,227,072 FLOPs notably reduce computational costs compared to cross-attention (52,723,712 FLOPs) and self-attention (67,633,152 FLOPs), achieving reductions by **91.98%** and **93.75%**, respectively, thus highlighting OvO’s superior efficiency. The unimodal results show that the textual modality is most valuable in phenotype prediction, and ClinicalBERT alone performs better than self-attention and cross-attention. This indicates that the added complexity and forced interactions are not necessarily conducive to result quality. However, OvO attention can extract information from the other modalities for a significant performance increase rather than a decrease.

For the six-modality Alzheimer’s detection task from TADPOLE, we show our results in Table 1.3. OvO’s 405,504 FLOPs significantly undercut cross-attention (8,921,088 FLOPs) and self-attention (9,633,792 FLOPs), achieving reductions of **95.45%** and **95.79%**, respectively, highlighting OvO’s remarkable efficiency.

Table 1.3: **TADPOLE results.** (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.45% compared to self and cross-attention.

Model	Modalities	$\downarrow \Delta$ FLOPs	$\uparrow$ MAUC	$\uparrow$ BCA
Neural Net	AV45 PET ROI	-	63.5 ± 3.1	56.4 ± 3.8
Neural Net	CSF Biomarkers	-	64.4 ± 1.1	53.6 ± 2.7
Neural Net	MRI ROIs	-	67.0 ± 1.3	57.2 ± 1.0
Neural Net	FDG PET ROI	-	66.6 ± 0.3	60.8 ± 0.7
Neural Net	Demographics	-	74.6 ± 0.9	62.0 ± 0.6
Neural Net	Cognitive Tests	-	97.8 ± 0.2	88.6 ± 0.7
Concatenation	All	*	97.7 ± 0.8	91.9 ± 1.9
Cross-Attention	All	8,921,088	97.1 ± 0.6	90.7 ± 1.7
Self-Attention	All	9,633,792	94.8 ± 1.1	86.6 ± 2.6
OvO Attention	All	405,504	98.3 ± 0.4	93.0 ± 1.4

Similarly to the MIMIC results, the unimodal results show that the cognitive tests modality is most valuable in disease prediction, and performs on its own better than self-attention and cross-attention. However, OvO attention can extract information from the other modalities for a significant performance increase rather than a decrease (p-value <0.01).

Lastly, the results on the six-modality eICU mortality prediction task are shown in Table [1.4](#) demonstrating the scalability and performance advantages of OvO attention.

Table 1.4: **eICU results.** We report the average of 10 random seeds for AUROC and AUPRC, along with standard deviations. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.12% compared to self and cross-attention.

Model	Modalities	$\downarrow \Delta$ FLOPs	$\uparrow$ AUROC	$\uparrow$ AUPRC
Neural Net	Demographics	-	50.2 ± 0.6	91.8 ± 0.2
Neural Net	Medication	-	56.3 ± 1.3	93.1 ± 0.3
Neural Net	Diagnosis	-	58.2 ± 2.1	93.3 ± 0.4
Neural Net	Treatment	-	66.1 ± 0.5	94.8 ± 0.1
Neural Net	APACHE APS	-	77.6 ± 0.2	97.0 ± 0.1
Neural Net	Laboratory	-	81.5 ± 0.4	97.0 ± 0.1
Concatenation	All	*	81.7 ± 1.6	97.5 ± 0.3
Cross-Attention	All	129,957,888	77.6 ± 1.6	95.4 ± 0.3
Self-Attention	All	151,781,376	80.2 ± 2.0	96.8 ± 0.4
OvO Attention	All	6,340,608	82.5 ± 0.9	97.8 ± 0.2

OvO’s 6,340,608 FLOPs significantly undercut those of cross-attention (129,957,888 FLOPs) and self-attention (151,781,376 FLOPs), achieving reductions of approximately **95.12%** and **95.82%**, respectively, thereby highlighting OvO’s efficiency. Mirroring the trends observed in the MIMIC and TADPOLE datasets, we note a dominant unimodal modality, specifically Lab modality, in this experiment as well. While concatenating modalities does enhance performance, this improvement is not seen in self and cross-attention models. In contrast, OvO attention not only reflects these performance gains but does so significantly (p-value <0.01).

We hypothesize that this is due to the overfitting of more complex integration frameworks of self and cross-attention on relatively smaller clinical datasets. OvO, in its simplicity akin to concatenation, manages to strike a balance by maintaining flexibility and capturing inter-modal interactions through its attention mechanism, thus offering an edge in performance without excessive complexity.

In summary, across diverse clinical datasets and modalities, OvO attention consistently outperforms traditional fusion techniques in both predictive performance and computational efficiency, underlining its robustness in handling complex multimodal healthcare data.

1.6 Conclusion

In this chapter, we present One-Versus-Others (OvO), a new scalable multimodal attention mechanism. The proposed formulation significantly reduces the computational complexity compared to the widely used early fusion through self-attention and cross-attention methods. Notably, OvO achieves, at minimum, a reduction of 91.98% in FLOPs when benchmarked against self and cross-attention methods across a range of clinical datasets containing up to six modalities. We provide both a detailed theoretical complexity analysis and empirical evidence from a simulated experiment, illustrating that OvO’s computational demand scales linearly with the number of modalities, in contrast to the quadratic scaling observed in other methods. Our proposed method provides a way to overcome one of the major challenges associated with multimodal datasets - computational resource demand and cost, thus enabling adoption in resource-constrained domains, such as clinical decision support. Overall, the results unequivocally establish that OvO not only significantly reduces computational expenses but also exceeds the performance of existing state-of-the-art fusion methodologies.

Chapter 2

Interpreting Multimodal Attention

In this Chapter, we present key results from our NAACL 2025 paper, *What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation* [37]. In this chapter, we focus on understanding the attention mechanisms within Vision–Language Models (VLMs), which play a central role in their success on complex cross-modal tasks. Despite the prominence and success of VLMs, their internal mechanisms, particularly the roles of cross-attention and self-attention in multimodal integration, are not fully understood. To address this gap, we introduce NOTICE, a Gaussian-**NO**ise-free **Text-Image Corruption and Evaluation** pipeline for mechanistic interpretability in VLMs. NOTICE introduces Semantic Image Pairs (SIP) corruption, the first visual counterpart to Symmetric Token Replacement (STR) for text. Through NOTICE, we uncover a set of “universal attention heads” in BLIP and LLaVA that consistently contribute across different tasks and modalities. In BLIP, cross-attention heads implement object detection, object suppression, and outlier suppression, whereas important self-attention heads in LLaVA only perform outlier suppression. Notably, our findings reveal that cross-attention heads perform image-grounding, while self-attention in LLaVA heads do not, highlighting key differences in how VLM architectures handle multimodal learning.

2.1 Introduction

Vision-Language Models (VLMs) have become central to both computer vision and natural language processing (NLP) due to their ability to complete complex multimodal tasks like visual recognition [45, 92], visual question answering [81, 5, 75, 78], and image captioning [84, 165, 174]. Many foundational VLMs, such as BLIP (Bootstrapping Language-Image Pre-training) [75], use cross-attention to capture interactions between modalities [145, 84, 75], while others, like LLaVA [81], successfully use self-attention for multimodal integration. Despite recent works investigating the differences in cross-attention and self-attention in VLMs [55, 110, 23], the specific role of different attention heads in vision-language is not fully understood, particularly in large VLMs such as LLaVA [170].

In NLP, mechanistic interpretability (MI) has proven effective for unraveling the complex structure of

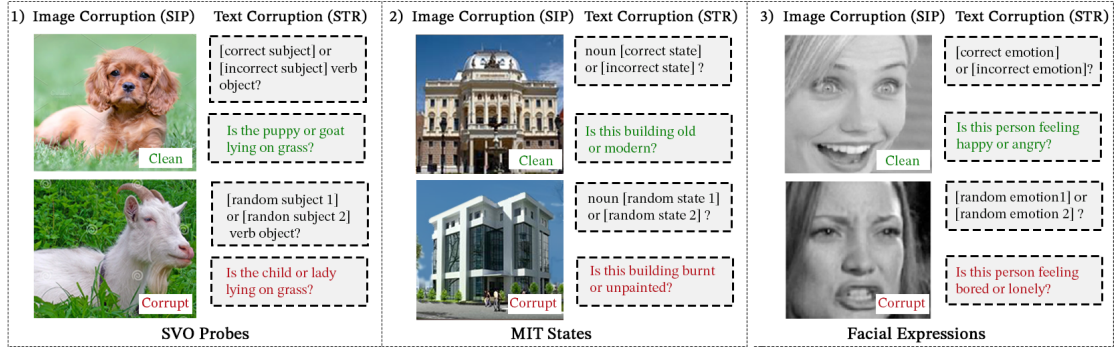


Figure 2.1: NOTICE applied to SVO-Probes, MIT-States, and Facial Expression Recognition. NOTICE involves creating Semantic Image Pairs for image corruption and Symmetric Token Replacement for text corruption.

LLMs by identifying pathways, or *circuits*, that enable behaviors like logical reasoning and indirect object identification [46, 156]. A core method in MI is *causal mediation analysis* (CMA), which measures output changes by intervening on inputs and tracking effects. To apply CMA in Transformers, prior works have used *activation patching* [91], which identifies key model components for a given task by replacing the hidden states of a *corrupt* run with those from a *clean* run. A “clean run” refers to a standard forward pass on an input. A “corrupt” run either alters the input by swapping input tokens with semantically related tokens or applies Gaussian noise to token embeddings so that the model can no longer produce the correct answer from the clean pass. While successful in NLP, mechanistic interpretability methods are understudied in VLMs due to the need for an effective semantic-based corruption scheme for images.

Previous works applying activation patching to VLMs only study BLIP, focus on Multi-Layer Perceptron (MLP) layers in BLIP’s decoder, and apply Gaussian noise to image embeddings to create corrupt runs [108]. However, multimodal interaction occurs in BLIP’s image-grounded text encoder through the cross-attention layers, meaning prior findings [108] do not provide insights on how VLMs integrate modalities. Further, recent works have shown that creating corrupt runs by applying Gaussian noise to token embeddings can produce illusory results [171]. The research [171] advocates for *symmetric token replacement* (STR) that swaps tokens in a clean prompt with semantically related tokens to create a corrupt run as it keeps the corrupted run within distribution and provides more reliable insights. Current approaches to image corruption lack an equivalent method to STR that does not use Gaussian noise for image corruption, leaving a gap in reliable multimodal analysis.

We propose *Semantic Image Pairs* (SIP) corruption, which leverages existing datasets to create clean and corrupt images that vary by one semantic property (Figure 2.1). For instance, in the SVO-Probes dataset [48], subject-verb-object triplets are already curated to vary by just one element (e.g., ‘puppy, lying, grass’ vs. ‘goat, lying, grass’), allowing us to create “clean” and “corrupt” image pairs from the existing image labels (Figure 2.1 panel 1). SIP is not limited to curated datasets and can be extended to any image dataset with easily interchangeable attributes, such as cars, shapes, animals, etc. Further, we show in Section 2.4.1 that SIP corruption can be performed using generative models to create corrupt images.

Our framework, NOTICE (Gaussian-**NO**ise-free **T**ext-**I**mage Corruption and **E**valuation), is a mechanistic interpretability pipeline for VLMs. NOTICE uses our novel SIP method to corrupt images and STR to corrupt text. Using NOTICE, we uncover “universal cross-attention heads” in BLIP and “universal self-attention heads” in LLaVA that consistently produce large patching effects across all tasks. We show that universal cross-attention heads implement human-interpretable functions categorized into three classes: object detection, object suppression, and outlier suppression. We find that both self-attention and cross-attention heads implement outlier suppression. However, only cross-attention heads in BLIP perform the image-grounding functions of object detection and object suppression, highlighting key differences in how VLM architectures handle multimodal learning.

2.2 Related Work

Multimodal Interpretability. Despite the successes of VLMs, the role of attention in multimodal integration remains under-studied [161]. Most works either design probing tasks to identify what vision-language concepts models struggle to capture [137, 48, 85] or assess whether models rely more on one modality over the other [12, 129, 49]. Probing tasks evaluate whether a model encodes a concept by training a linear classifier on its hidden states. Others [129], create visual, text, and multimodal probing tasks to show that models heavily rely on text priors and struggle with object, size, and counting. Gradient-based methods analyze gradients to understand how changes in input features influence outputs, assigning importance scores to features and providing insights into the decision-making process. Liang et al. [79] use second-order gradients, building on LIME [123] and Shapley values [93], to disentangle cross-modal interactions and determine how much each modality contributes to decisions. While these works have provided insights into multimodal models, recent research shows that 1) probing methods can produce misleading results [7] and 2) gradient-based methods for feature attributions can fail to predict behavior [9]. Instead, Giulianelli et al. [34], Conneau et al. [18] demonstrate that causal approaches are more informative than probing representations.

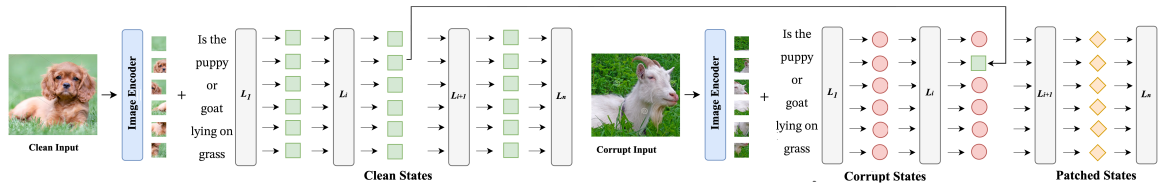


Figure 2.2: Activation Patching using SIP corruption. The image of the puppy is the clean image, I , and the goat is the corrupt image, I^* . Patching the correct answer token “puppy” at $M_I(I, T)$, from the clean image to the “puppy” token at $M(I^*, T)$ creates the patched states $M'(I^*, T)$ shown as orange diamonds.

Multimodal models typically integrate their inputs through one of two main approaches: (1) early fusion [81, 77], where self-attention processes combined image patches and text representations, or (2) cross-attention fusion [145, 84], where queries from one modality interact with representations from the other. [28] investigate cross-modal integration using input ablation techniques and conclude that self-attention models, such as VisualBERT, show less sensitivity to missing visual input compared to cross-attention models like ViLBERT,

which exhibit more significant reliance on cross-modal interactions. [56] finds that vision-text cross-attention learns visual grounding of noun phrases into objects, while text-to-text attention captures low-level syntactic word relations. [110] probes dual-stream VLMs through tasks like image-sentence alignment and counting, revealing strong alignment performance but limited grounding of visual symbols, resulting in poor counting capabilities. However, there has yet to be a comprehensive study applying mechanistic interpretability techniques to investigate how recent architectures like BLIP and LLaVA 1) use attention mechanisms to integrate multimodal information and 2) determine the specific functions that attention heads in these models learn to perform.

Causal Mediation Analysis and Mechanistic Interpretability. Causal mediation analysis (CMA) uncovers how interventions in one part of a system cause changes in another. Recent works have adapted CMA to study language models by using *activation patching* [91] to identify causal pathways, or *circuits*, responsible for certain behaviors in the model. Activation patching shows the indirect effect of a component on a model’s output by corrupting key input tokens and then restoring outputs by patching the internal modules from the clean run to the corrupted run. This approach has revealed that attention heads in transformers implement human-interpretable algorithms for “greater than” tasks [46], indirect object identification [156], and learning mathematical group operations [17]. Evidence suggests that models implement similar circuits across various tasks, indicating that studying a small set of computational units may provide insights into model behavior [43, 94]. Although works in mechanistic interpretability have provided critical insights into LLMs, works applying activation patching to multimodal models have been limited due to the absence of an effective semantic-based method for image corruption. Given that Zhang and Nanda [171] demonstrate Gaussian noise corruption can lead to “illusory” patching results, a semantic-based corruption scheme for images is needed to apply causal mediation analysis to VLMs. In this work, we introduce a semantic-based image corruption method (SIP) and token replacement (STR) for text, enabling meaningful causal mediation analysis to reveal distinct roles of self-attention and cross-attention in multimodal models like BLIP and LLaVA, advancing our understanding of vision-language integration.

2.3 Methods

NOTICE: Overview The first step in activation patching is developing a method to corrupt the input, preventing the model from predicting the correct answer without causing representations to be out-of-distribution. We propose the first semantic-based dual corruption scheme for VLMs to enhance understanding of each modality’s contribution to VQA tasks. We assess the impact of corrupting input text with symmetric token replacement [156] and images using our novel *Semantic Image Pairs* (SIP) framework.

Model Selection For our model selection, we use LLaVA-1.5-7B (Vicuna base) [81] and BLIP-VQA-Base [75], chosen for their relevance and alignment with recent interpretability research. BLIP-VQA was included to enable direct comparison with prior work applying activation patching to vision-language models [108], facilitating a meaningful evaluation of our corruption scheme (NOTICE) against Gaussian-noise patching. In addition, we also focus on LLaVA-1.5 to align with more recent interpretability studies, including [86], [101], and [62].

Text Corruption: Symmetric Token Replacement. Corrupting text inputs with symmetric token replacement (STR) swaps out the correct answer token of a prompt with another similar token that has the same length [156]. For indirect object identification, Wang et al. [156] swap the names of the indirect and direct objects with randomly sampled names. The goal of STR is to corrupt inputs so the language model cannot complete the task correctly while avoiding Gaussian noise corruption that pushes token embeddings off-distribution. For our VQA tasks, we swap both multiple-choice options in the input prompt with two randomly sampled incorrect multiple-choice options with the same tokenized length from a different sample in the dataset. Figure 2.1 illustrates how STR alters the input text for a prompt.

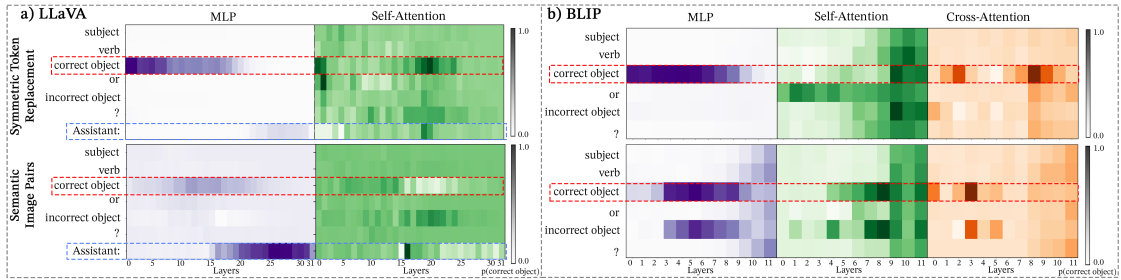


Figure 2.3: Module-wise activation patching results for BLIP and LLaVA on “objects” from SVO-Probes. We visualize the restoration probability after patching for MLP, self-attention, and cross-attention layers in the image-grounded text-encoder. The y-axis denotes which token we patch, and the x-axis denotes which layer we patch.

Image Corruption: Semantic Image Pairs. While STR is a widely used text corruption method, extending semantic corruption to the vision domain presents significant challenges as it requires a semantically aligned image pair. To address this, we propose SIP corruption, a visual corollary to STR, which modifies only one semantic concept in an image (see Figure 2.1).

We implement SIP in three diverse image-classification datasets: SVO-Probes [48] (an extension to FOIL it! [137]), MIT States [58], and Facial Expressions [42]. While natural variation in backgrounds can happen between semantic image pairs, we select datasets where nouns, subjects, verbs, and objects are known and can be controlled. The Facial Expressions dataset was curated to include only black-and-white photos of faces expressing different emotions. In MIT-States, object-state pairs are known and can be used to maintain objects (e.g., “building”) while varying their states (e.g., “modern”). The SVO-Probes dataset is designed to vary only one element (subject, verb, or object) at a time (e.g., only varying “puppy” and “goat”, while maintaining “grass” and “lying” constant, as seen in Figure 2.1 panel 1). Thus, for each dataset, we collect the known elements (emotions, objects/states, subjects/verbs/objects) and create image pairs that differ on that element. Our framework can be extended to any image dataset with clearly defined attributes, such as cars, shapes, animals, etc., using the class labels to find clean and corrupt pairs.

Previous works have shown vision transformers can correctly identify objects in various contexts regardless of the foreground or background [154]. Since image encoders in multimodal models, such as BLIP and LLaVA, are robust to minor variations to the context of an object, only the semantic contents of the image need to be controlled to obtain reliable patching results [154]. In contrast, masking pixels, adding Gaussian noise to image patches, or replacing pixels would create unnatural images, leading to unreliable results. In

Section 2.4.1, we demonstrate the effectiveness of SIP corruption over Gaussian-noise image corruption and experiment with using generative models to create SIP.

NOTICE VQA Task Design. We adapt the SVO-Probes, MIT States, and Facial Expressions image-classification datasets into VQA tasks. We prompt models to choose between a correct answer that matches the image and an incorrect, semantically different alternative, as seen in Figure 2.1. For example, using the image-class labels “happy” and “angry,” we create a VQA prompt: “Is this person feeling happy or angry?”. For LLaVA, we adapt the standard prompting format to ensure the model selects one of the multiple choice options: “USER: <image> [Prompt]. Answer with one word. Assistant:”. While in SVO-Probes, each subject, verb, or object comes with its own semantically different counterpart, we constructed such pairs for MIT-States and Facial Expression. In MIT-States, we separated states into color, shape, and texture and selected counterparts from the respective categories for each question. We grouped similar states (e.g., “huge”, “large”) together and ensured that the selected counterpart would not be similar in meaning to the correct answer (using cosine similarity and word2vec word embeddings [95]). Similarly, in the Facial Expressions dataset, we grouped the seven emotions into positive, negative, and neutral and selected counterparts from opposing emotions. See more details on the VQA design in Appendix A.5 and Appendix A.6 respectively

Activation Patching. Let M be a vision-language model and M_l its hidden states at layer $l \in \{0, 1, \dots, n\}$. The forward pass on an image (I) and text (T) pair at layer l is denoted as $M_l(I, T)$. We define I^* as a corrupted image and T^* as a corrupted text input. Note that we only corrupt one modality at a time. In a forward pass with I^* and T , activation patching replaces the hidden states of the corrupted run $M_l(I^*, T)$ with the clean run $M_l(I, T)$ to determine which components have the most significant indirect effect on the output. The hidden states at all subsequent layers, $M'_j(I^*, T)$ for $j \in \{l + 1, \dots, n\}$, are called *patched states*. In Figure 2.2 we provide an example illustrating patching using SIP image corruption. The image of the puppy is the clean image, I , and the goat is I^* . Patching the correct answer token “puppy” at $M_l(I, T)$, from the clean image to the “puppy” token at $M(I^*, T)$ creates the patched states $M'(I^*, T)$ shown as orange diamonds in Figure 2.2.

Measuring Patching Effects. The most common ways of measuring patching effect are *restoration probability* and *logit difference*. Let L^* and L' denote the logits from the corrupted and patched runs, and let τ, τ^{inc} denote the correct/incorrect multiple-choice answer. Logit difference measures the indirect impact patching has on the model’s logits. Let $L(\tau, \tau^{\text{inc}}) = \text{Logit}(\tau) - \text{Logit}(\tau^{\text{inc}})$. The logit difference is defined as $L'(\tau, \tau^{\text{inc}}) - L^*(\tau, \tau^{\text{inc}})$. Restoration probability measures the impact of patching clean states into a corrupt run on predicting the correct answer token: $P'(\tau) - P^*(\tau)$.

2.4 Results

We begin our investigation by conducting module-wise activation patching to understand the global patterns of self-attention, cross-attention, and MLP modules under both image and text corruption. We demonstrate that SIP image corruption produces results closely aligned with STR text corruption and effectively highlights the role of middle-layer MLPs, which have repeatedly proven crucial in a wide range of tasks [91]. In contrast, Gaussian Noise image corruption only identifies the final MLP layers as producing a significant

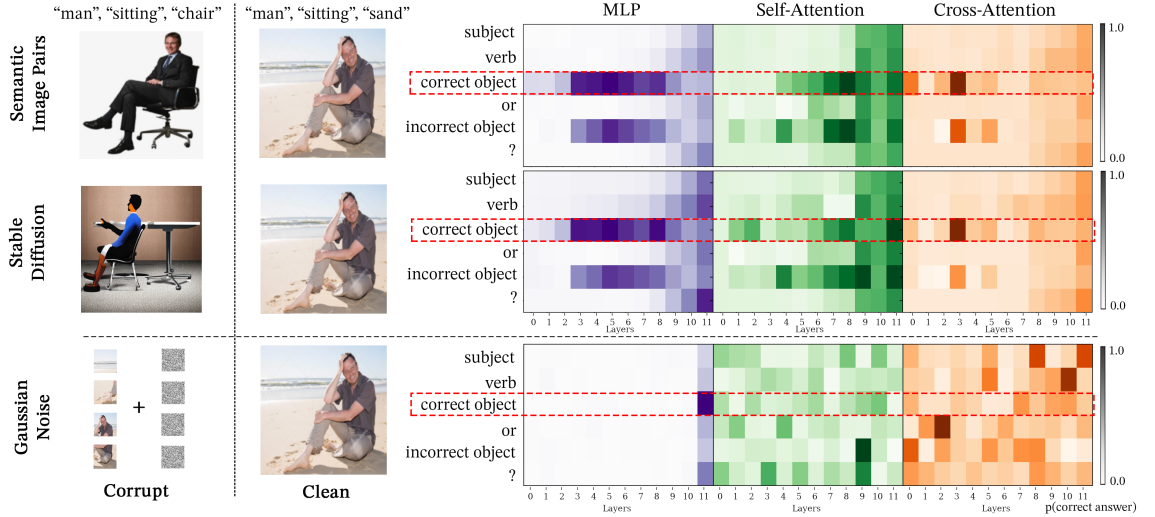


Figure 2.4: Module-wise activation patching results for SIP, and Gaussian-Noise corruption on SVO-Probes on BLIP. SIP corruption produces activation patterns that align with Stable Diffusion results and highlight the importance of middle layers, while Gaussian noise fails to reveal meaningful attention layers, emphasizing the effectiveness of SIP for probing vision-language models.

patching effect. Building on this, we use SIP to perform a fine-grained analysis of cross-attention heads in BLIP and self-attention heads in LLaVA, the critical point of multimodal integration for each model. As we identify the most significant cross-attention heads in BLIP and most significant self-attention heads in LLaVA, we categorize them into three distinct functional roles, offering deeper insights into their contributions and behaviors.

2.4.1 Module-wise Activation Patching

Module-wise activation patching reveals how different model components (i.e., attention blocks and MLPs) of VLMs respond to image and text corruption, highlighting which layers and modules (MLPs, self-attention, cross-attention) are most influential for successfully completing VQA tasks. First, we compare the effects of STR and SIP on BLIP and LLaVA, as shown in Figure 2.3. Across all tasks, we observe: (1) MLP patching reveals that image corruption emphasizes the importance of middle layers, while text corruption primarily affects earlier to middle layers; (2) Under text corruption, LLaVA’s self-attention demonstrates strong localization, on the correct object token in layer 0 and layer 21, similar to the cross-attention localization on the correct answer seen in BLIP at layer 3; (3) Under image corruption, LLaVA’s self-attention attends away from the correct answer token in layers 16-25; and (4) LLaVA’s earlier layers identify the correct answer token while in the later layers the emphasis shift to the instruction token “Assistant.”.

We evaluate the effectiveness of SIP corruption by comparing it with Gaussian-noise, the image corruption technique used in previous work [108]. In addition, we explore semantic image corruption using generative models, which shows the NOTICE framework is generalizable beyond curated datasets like SVO-Probes. Specifically, we use in-painting Stable Diffusion [124] to perform semantic image corruption by generating

“corrupted” versions of clean images that only vary on one key semantic property. For instance, generating an image of a man sitting on a chair from an image of a man sitting on sand. Figure 2.4 demonstrates the promise of using generative models to implement SIP corruption since stable-diffusion module-wise patching results are highly similar to SVO-Probes. Note that generative models may occasionally introduce undesirable image artifacts (e.g., adding extra fingers to a hand) that we cannot control. While using Stable Diffusion to create SIP shows promise for future CMA applications, Gaussian noise proves unreliable. In both LLaVA and BLIP, MLP modules highlight the importance of early and middle layers, while Gaussian noise only highlights the last layer (see Figure 2.4). Several studies highlight the importance of middle layers in transformers [25, 91], and show that MLPs store syntactic information and task-specific representations often peak in middle layer MLPs [25, 50, 35, 60]. Additionally, self-attention patching in SIP assigns the highest probability to the correct token, unlike Gaussian noise. SIP and STR also display consistent activation patterns—long horizontal regions in MLPs (aligning with [91]) and vertical regions in self-attention—whereas Gaussian Noise shows clear differences. These findings suggest that SIP is a more effective tool for probing VLMs and identifying key components and interactions.

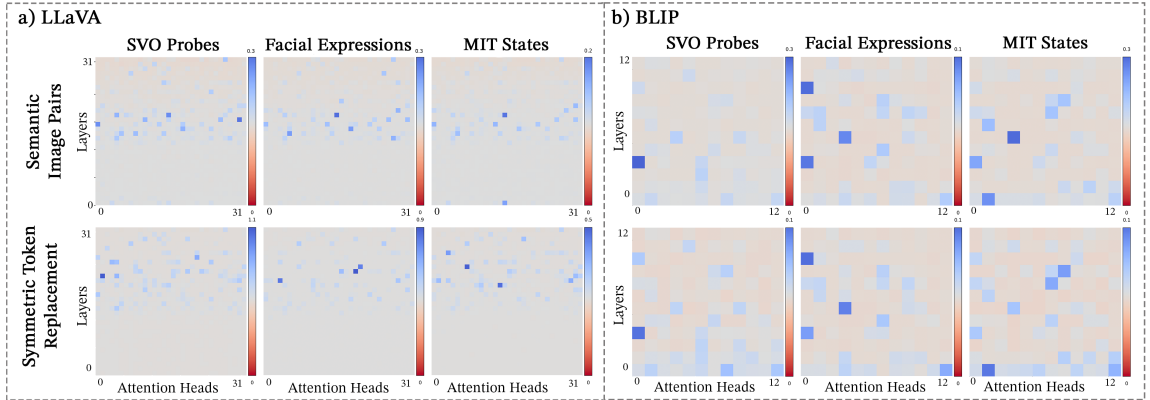


Figure 2.5: Logit difference demonstrating the impact of patching the correct answer for each LLaVA self-attention head and BLIP cross-attention head on the SVO-Probes, MIT States, and the Facial Expressions datasets. Many key attention heads overlap in importance across both modalities.

2.4.2 Universal Attention Heads Exist Across Corruption Schemes and Tasks

Following the module-wise analysis, we focus on the attention heads, investigating their role in facilitating vision-language integration. We calculate the logit difference for each cross-attention head in BLIP and self-attention head in LLaVA by analyzing the impact of patching the correct answer across the SVO-Probes, MIT States, and Facial Expressions datasets in BLIP and the “Assistant:” token in LLaVA. We perform activation patching on the “Assistant:” token since Figure 2.3 demonstrates that patching the “Assistant:” creates the largest module-wise patching effect. See Appendix A.10 for more details, as well as for the patching of the correct answer token in each self-attention head in LLaVA. Figure 2.5 demonstrates that, in BLIP, image corruption has a greater impact on logit difference compared to text corruption in SVO-Probes and MIT-States. Specifically, image corruption results in up to 20% logit difference, while text corruption

Model	Vision Heads	Multimodal Heads	Text Heads
LLaVA	L15.H5, L16.H18	L28.H7, L31.H27	L22.H16, L29.H19
	L14.H4, L15.H8	L16.H24, L19.H15	
	L18.H26, L17.H0	L18.H30, L18.10,	
	L17.H22, L14.H28	L21.H30	
	L19.H4, L15.H14		
	L17.H13, L20.H28		
BLIP	L5.H3	L3.H0	L0.H11

Table 2.1: Universal attention heads in LLaVA and BLIP, identified as having consistently high logit differences across tasks and modality corruptions. In LLaVA, most critical self-attention heads serve multimodal or vision-specific roles. In contrast, BLIP’s dual-stream design results in distinct vision, text, and multimodal heads, supporting more modular cross-modal integration.

causes a difference of 6%, suggesting BLIP relies more on visual input than text for predicting the correct answer. In LLaVA, Figure 2.5 shows the logit difference is consistent for both modalities. These results contrast with previous findings that VLMs are more influenced by textual priors [129]. Despite the difference in scale in BLIP, many critical attention heads identified under text corruption are also crucial when images are corrupted, with this overlap persisting across tasks. To quantify the overlap between the two modalities’ corruptions and across tasks, we identify three universal cross-attention heads in BLIP (L0.H11, L3.H0, and L5.H3) and seven universal multimodal self-attention heads in LLaVA (L28.H2, L31.H27, L18.H10, L16.H24, L21.H30, L19.H15, and L18.H30) where $Lx.Hy$ denotes layer x , head y .

Table 2.1 shows “universal attention heads”, which we defined as consistently having a logit difference two standard deviations above the average regardless of the task or the modality corruption. In LLaVA, we find only find two “text only” heads; a majority of important self-attention heads either overlap with multimodal heads that are important for both modalities or emerge under image corruption (“vision only” heads). This suggests visual inputs still influence the prompt, even under text corruption. We hypothesize this is due to LLaVA’s early fusion, where image tokens are embedded into the textual prompt. In contrast, BLIP has exactly one vision, one text, and one multimodal head, likely because its dual-stream cross-attention design allows the model to fuse visual and textual information more effectively without requiring multiple specialized heads for different modalities. These findings highlight key differences in cross-modal integration between the two models.

The presence of overlapping attention heads when patching different modalities and tasks is a novel discovery. Since cross-attention heads in BLIP and self-attention heads in LLaVA integrate vision and language, overlapping heads suggest that vision and language convey similar information to the VLM despite being distinct feature spaces. This challenges the longstanding notion that text is the dominant modality for VLMs. Additionally, the reuse of these components across tasks offers insights into model behavior and model editing, enabling targeted improvements that enhance performance, address bias, and better adapt the model to new tasks.

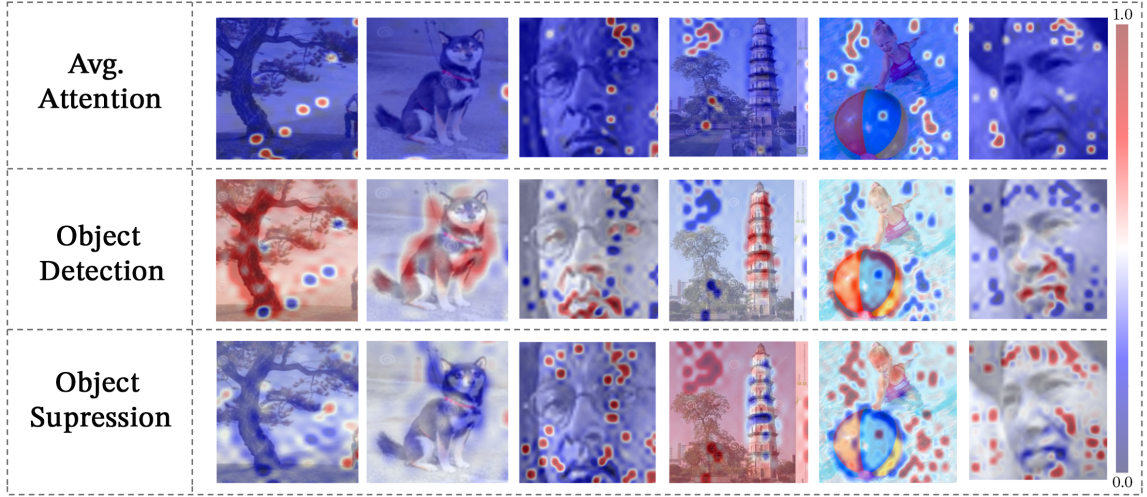


Figure 2.6: Universal cross-attention heads in BLIP implement object detection and object suppression. We visualize the average of all attention heads compared to the universal heads performing each function.

2.4.3 Universal Attention Heads Perform Distinct Functions

To better understand the role of universal attention heads in multimodal integration, we visualize the cross-attention patterns between the “correct answer” text token and image patches in BLIP. For LLaVA, we visualize self-attention patterns of universal multimodal self-attention heads from the “Assistant:” token to image patches. We find that universal cross-attention heads in BLIP fall into three function classes: *object detection*, *object suppression*, and *outlier suppression*, seen in Figure 2.6 whereas universal multimodal self-attention heads in LLaVA primarily implement outlier suppression seen in Figure 2.7

Outliers in Attention Patterns. Similar to Darcet et al. [20], we find that cross-attention patterns in BLIP and text-to-image self-attention patterns in LLaVA are dominated by “outlier features” with extremely high average attention scores. These outliers appear in Figure 2.6 as non-informative image patches with dark red scores. Unlike in ViTs, where outliers are thought to obscure interpretable features in vision attention, we find attention functions reduce attention to these outliers. In BLIP, this enables object detection by linking text tokens to the relevant image patches. Our finding that outliers play an important role in VLM representations is similar to recent works showing outliers in LLMs store task-specific information [125].

Object Detection. In BLIP, cross-attention layer 3, head 0 (L3.H0) implements object detection by linking a text token with relevant image patches. These heads associate physical objects in images with the nouns in SVO-Probes and MIT States. Notably, L3.H0 can segment images even when the word does not explicitly appear in the image. For example, in the Facial Expressions dataset, L3.H0 implicitly associates the correct emotion (e.g., “anger”) with the mouth region. L3.H0 is the only universal cross-attention head identified as “multimodal,” consistently demonstrating its importance across all tasks, regardless of modality corruption. Its ability to map text tokens to image patches highlights its critical role in fusing vision-language information and demonstrates its true multimodal capabilities.

Object Suppression. Cross-attention head L5.H3 in BLIP explicitly attends *away* from the image patches

associated with an input token and attends strongly to the outlier cross-attention patches. Object suppression cross-attention heads perform the inverse of object detectors, as they explicitly reduce attention to specific objects in an image. In SVO-Probes, this amounts to attending away from subjects or objects, and in MIT-States, L5.H3 attends away from relevant portions of the image that form the noun that is changing states. However, L5.H3 does not perform object suppression in Facial Expressions. Instead, L5.H3 performs the closely related function of outlier suppression. L5.H3 in BLIP provides a rare example of an *interpretable polysemantic* attention head that implements different interpretable functions depending on the input [43].

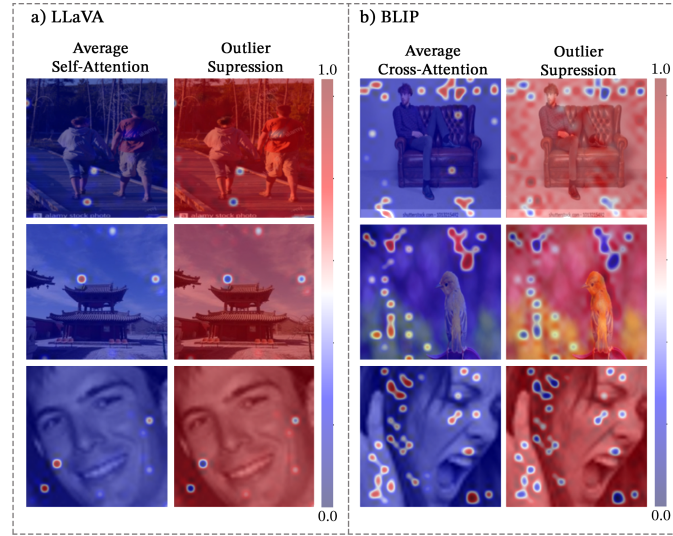


Figure 2.7: Universal heads in BLIP and LLaVA perform outlier suppression. We visualize the average of all attention heads compared to the universal heads performing outlier suppression.

Outlier Suppression. Outlier suppression heads reduce attention to high-probability image patches and distribute attention uniformly across remaining patches. All universal multimodal attention heads in LLaVA implement outlier suppression. In BLIP, L0.H11 implements outlier suppression across all tasks in this paper. For BLIP, L0.H11 is a text-only universal cross-attention head, meaning it does not produce a significant patching effect when images are corrupted. Due to its specialization, L0.H11 does not contribute to the semantic association between text tokens and image patches. Instead, it plays a crucial role in filtering out irrelevant visual information, enhancing the model’s robustness.

2.5 Conclusion

In this chapter, we study the roles of cross-attention and self-attention in multimodal integration. To accomplish this, we introduce NOTICE, a Gaussian-Noise-free Text-Image Corruption and Evaluation pipeline for mechanistic interpretability in VLMs. Using our SIP image corruption framework and STR text corruption, NOTICE enables causal mediation analysis across both modalities. Experiments across three tasks reveal

“universal attention heads” with consistent patching effects across BLIP and LLaVA. We show that cross-attention supports three functions: object detection, object suppression, and outlier suppression. However, self-attention primarily handles outlier suppression, indicating cross-attention’s unique role in image grounding. While attention presents a plausible, but not necessarily faithful rationale for model prediction [158], our findings show patterns of attention heads that consistently, and causally, influence model logits. With the support NOTICE, applications using VLMs can benefit from more transparent and adaptable models, facilitating deeper insights and enhanced performance.

Chapter 3

Controlling Factual Recall in Attention

In this Chapter, we present key results from our EMNLP 2025 paper, *Pixels Versus Priors: Controlling Knowledge Priors in Vision-Language Models through Visual Counterfactuals* [40]. While the preceding chapters explored how attention can be scaled and interpreted across modalities, in this chapter, we zoom in on its role in reconciling conflicting sources of information, particularly world knowledge versus visual evidence. To investigate this, we introduce Visual CounterFact, a new dataset of visually-realistic counterfactuals that put world knowledge priors (e.g, red strawberry) into direct conflict with visual input (e.g, blue strawberry). Using Visual CounterFact, we show that model predictions initially reflect memorized priors, but shift toward visual evidence in mid-to-late layers. This dynamic reveals a competition between the two modalities, with visual input ultimately overriding priors during evaluation. To control this behavior, we propose *Pixels Versus Priors* (PvP) steering vectors, a mechanism for controlling model outputs toward either world knowledge or visual input through activation-level interventions. On average, PvP successfully shifts 99.3% of color and 80.8% of size predictions from priors to counterfactuals. Together, these findings offer new tools for interpreting and controlling factual behavior in multimodal models.

3.1 Introduction

As multimodal large language models (MLLMs) demonstrate increasing success in real-world vision-language tasks [74, 157, 14], it is becoming increasingly important to understand their internal mechanisms in order to ensure the reliability and safety of these systems [37, 62, 86, 126]. Despite recent advances, interpretability in MLLMs remains underdeveloped compared to progress in natural language processing (NLP), where the encoding of world knowledge facts is well-researched and methods exist for systematically editing factual associations [90]. In NLP, *counterfactual datasets* consist of minimally altered input pairs that isolate specific factual changes, such as swapping one entity or relation while holding others constant. These datasets enable causal analysis of model behavior and have been central to understanding how factual associations are stored, retrieved, and manipulated [31, 32, 19, 169, 90]. Unlike language, where factual associations are well understood, there is no visual equivalent for locating or modifying stored associations in MLLMs. In particular,

there is no counterfactual dataset for testing how these models balance visual perception against memorized priors, nor any method for controlling their responses when the two sources of information conflict. To address this gap, we introduce Visual CounterFact, the first dataset designed to study world knowledge priors related to visual attributes in MLLMs, and use it to develop *Pixels Versus Priors* steering (PvP), a method for controlling whether the model relies on pixel-level information or on world knowledge.

Visual CounterFact modifies visual attributes, color and size, of everyday objects to create direct conflicts between memorized facts and input pixels. In our framework, *world knowledge priors* refer to linguistic associations between visual attributes and objects that the model has memorized during pretraining. In contrast, *visual perception* is defined by the in-context visual input, which we manipulate to create counterfactual images. These counterfactuals are designed to challenge the model’s world knowledge of visual attributes by presenting plausible but contradictory visual evidence. For example, we contrast the size-related knowledge prior “strawberries are bigger than flies” with the counterfactual “flies are bigger than strawberries,” violating expected size relations (see Figure 3.1).

Using Visual CounterFact, we find that MLLMs often ignore world knowledge when shown counterfactual images, favoring perceptual input even when prompted for general facts. We then trace where in the forward pass predictions shift from in-weight knowledge (e.g., strawberries are red) to in-context perception (*this* strawberry is blue), finding that this transition consistently emerges in mid-to-late layers. During this transition, models frequently flip between the two answers, revealing a competition between in-context pixel and in-weights prior information, with pixels often overriding priors in the model’s output. To control this behavior, we use our *Pixels Versus Priors* steering to control whether a vision-language model relies on knowledge priors or pixel information. PvP is a novel framework to construct steering vectors for vision-language models that control whether the model responds based on memorized knowledge or in-context visual input. Through this steering, we successfully shift an average of 99.3% of color predictions and 80.8% of size predictions from memorized priors to counterfactual answers.

Together, these contributions provide a new visual counterfactual benchmark and a mechanism for interpreting and controlling the behavior of vision-language models. We present the necessary foundation for a mechanistic understanding of how MLLMs integrate image input with prior knowledge of visual attributes, bridging the gap between interpretability research in language models and the emerging needs of multimodal models.

3.2 Related Works

Studies in mechanistic interpretability have shown that LLMs encode factual associations grounded in world knowledge within their weights, enabling precise manipulation through targeted interventions. In particular, feedforward layers often act as key-value memories, injecting factual knowledge into subject representations [31, 32], while clusters of “knowledge neurons” have been shown to store and control specific facts [19, 169]. These internal representations can be edited by introducing counterfactuals through weight-level interventions [90], or by tracing how attention mechanisms recover or suppress modified content during inference [63]. More recently, activation-level interventions have emerged as an alternative to weight-level

editing. In NLP, *steering vectors* are computed by subtracting internal representations from contrasting prompts to isolate meaningful activation directions [144, 150]. These directions can be added at inference to shift a model’s behavior without altering the model’s weights.

Although model editing and steering have been successful in shifting model outputs, Gekhman et al. [30] show that even when a model produces incorrect outputs, it may still internally represent the correct fact, highlighting a disconnect between stored knowledge and in-context generation. Investigating this disconnect further, there is a growing body of NLP literature that seeks to understand how language models flexibly deploy both in-context and memorized in-weight knowledge [13, 140, 3, 122, 70, 180, 112]. These studies suggest that models often switch between relying on memory and adapting to context, depending on training dynamics and task structure. This inherent conflict *within*-modality motivates our mechanistic analysis *across*-modalities.

In order to study the conflict of world knowledge priors across vision and language, a visual counterfactual dataset is needed. While many benchmarks test visual-textual alignment, none directly evaluate a model’s reliance on visual world knowledge. VL-Checklist [178] and VALSE [111] vary captions over fixed images, while FOIL-COCO [137], SVO-Probes [48], and Winoground [148] alter semantic content in real images. However, these methods often suffer from uncontrolled visual artifacts. COCO-Counterfactuals [71] uses generative models to edit images to replace a single object in the image with a new object. For example, for an image reflecting the caption “A large black ball sitting next to a glass of **milk**”, they generate a “counterfactual” image from the prompt “A large black ball sitting next to a glass of **water**”. While these images represent minimally altered pairs, they are not true counterfactuals designed to contradict visual world knowledge priors, such as the expectation that “strawberries are red” or that “strawberries are larger than flies”.

Similarly, model editing methods developed for LLMs have proven difficult to adapt to vision-language models. Early work shows that multimodal neurons can encode visual-textual concepts [109], and that vision and language encoders often share object-level semantics [130]. Yet, attempts to localize or edit factual knowledge in MLLMs, such as MMEdit [16] and VLKEB [53] face challenges with generalization and control. Model edits can introduce unintended side effects, such as altering predictions on unrelated inputs, and often fail to generalize across paraphrased prompts or unseen contexts. In addition to the side effects of model editing, in this work, we show that MLLMs tend to override memorized knowledge priors when presented with conflicting visual input. Given that pixels override priors, applying model editing to steer the model towards “strawberries are blue” would be overridden when presented with an image of a red strawberry. Instead, steering provides a reliable mechanism to modulate model responses between pixels and priors.

Steering has seen little adoption in the multimodal setting. Liu et al. [83] propose latent-space steering to reduce hallucinations by stabilizing vision features at inference time. Zhang et al. [175] show that embedded image prompts can act as hidden meta-instructions, influencing output style or sentiment. Luo et al. [86] demonstrate that vision-language models learn shared task vectors that generalize across modalities, suggesting that internal representations can be steered with compact task encodings. Similarly, Hojel et al. [51] identify visual task vectors in the activation space of prompting models, showing that these representations can be patched into attention heads to guide model behavior across tasks. However, these works do not examine how models handle conflicts between visual input and stored knowledge, nor do they provide mechanisms for explicitly controlling which source of information the model relies on. To our knowledge, no prior work

applies steering vectors to vision-language models for the purpose of modulating reliance on memorized visual priors versus image inputs.

3.3 Creating Visual CounterFact

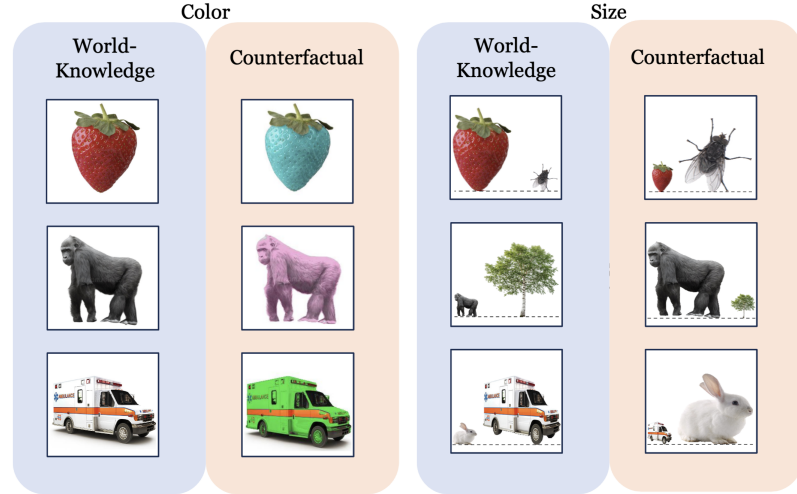


Figure 3.1: **Visual CounterFact**. A new benchmark to study how VLMs utilize world knowledge compared to visual inputs. (Left) images created using color relations, (right) images created using size relations.

First, we describe the creation of *Visual CounterFact*, a dataset designed to examine how MLLMs use visual input and world knowledge when presented with controlled counterfactual examples. *Visual CounterFact* contains images that deliberately introduce conflicts between visual input and world knowledge, spanning two tasks: color and size. Each image is created through a four-step pipeline designed to preserve realism and control for background noise while introducing counterfactual evidence. Additional details on each step are provided in Appendix Section [A.11](#)

(Step 1) Identifying objects with strong visual priors. We begin by selecting objects that have widely known visual attributes, such as canonical colors (e.g., “strawberries are red”) or typical size relationships (e.g., “strawberries are larger than flies”). These objects are sourced from human-annotated datasets (McRae norms [\[89\]](#)) and extended with GPT-4o estimates of typical attributes for CIFAR-100 [\[69\]](#) and ImageNet [\[22\]](#) categories.

(Step 2) Retrieving world knowledge images. For each object, we collect images using the Google Images API, specifying that the object should appear on a white background to reduce spurious visual cues. We aim to retrieve images that match the canonical visual prior (e.g., a red strawberry rather than a pink or green one). Each image is filtered and scored by GPT-4o for color accuracy, object correctness, and realism, and the highest-scoring image is selected.

(Step 3) Generating counterfactual relations. We construct counterfactuals that intentionally conflict with typical visual priors for each object. For the color task, we first prompt the LLaVA-Next model to generate likely colors for a given object (e.g., “What color is a strawberry?”), then sample from the five least likely

common colors (e.g., blue, orange, purple) to select a counterfactual color. To maintain visual clarity, we constrain these counterfactuals to be visually distinct from the original (e.g., avoiding red/orange or gray/black swaps). For the size task, we use GPT-4o to estimate the real-world dimensions of objects and compute their total size. We select object pairs that differ by at least a factor of 10 and generate two counterfactual relations per object by inverting the expected size ordering. For example, if object A is smaller than object B and object B is smaller than object C, we create counterfactuals such as “A is bigger than B” and “B is bigger than C.”

(4) Editing images to reflect counterfactual attributes. We use SAM2 [121] segmentation masks to apply controlled, localized transformations. In the color task, we modify hue values while preserving texture and shading; in the size task, we resize object masks and align them on a dashed line to reflect altered size relations without introducing depth ambiguity. The final dataset contains 575 color exemplars, 575 color counterfactuals, and 877 original and 877 counterfactual size images, for a total of 2,904 visually grounded examples.

3.4 Methods

We use Visual CounterFact to evaluate how MLLMs store world knowledge priors of visual characteristics using three models: LLaVA-Next-7B [74], Qwen2-VL-7B [157], and DeepSeek Janus Pro-7B [14]. These models were selected to cover a range of current state-of-the-art multimodal architectures, including established families like LLaVA, and emerging MLLMs like QwenVL and Janus Pro. We first apply early decoding to trace the evolution of the model’s prediction across layers and identify the point at which visual information overtakes linguistic priors or vice versa. We then develop PvP steering vectors that can actively shift model behavior toward either image-grounded or world knowledge responses. The results of these methods are presented in Sections 3.5.2 and 3.5.3, respectively.

3.4.1 Early Decoding

Early decoding is a technique for probing the intermediate computations of a model by decoding hidden states before the final output layer. Originally introduced by nostalgebraist [103] and extended in follow-up work [8, 107, 33, 153], this method applies the final layernorm, σ , to the intermediate hidden states h_l at layer l and then projects this representation onto the vocabulary space using the unembedding matrix W_U , yielding $W_U(\sigma(h_l))$. This produces a probability distribution over tokens, effectively allowing us to observe what the model “believes” at a given stage in its forward pass.

We use early decoding to identify when the model’s prediction shifts from being guided by knowledge stored in weights to being grounded in visual perception. By decoding the model’s predictions layer by layer, we observe how the probability distribution over possible output tokens evolves, allowing us to pinpoint where the model begins to favor a counterfactual (image-based) answer over the memorized world knowledge alternative.

3.4.2 Pixel Versus Prior Steering

Using Visual CounterFact, we introduce *Pixels Versus Priors* (PvP) steering vectors by calculating the difference in activations with contrasting prompts. Specifically, we present the model with a counterfactual image accompanied by one prompt that encourages the retrieval of world knowledge priors and another that directs it to analyze the image pixels. Consider the example in Figure 3.2. The prompt “What color *is this* strawberry?” encourages a visually grounded response, while “What color *are most* strawberries?” draws on memorized world knowledge priors about the color of a strawberry. When paired with a counterfactual image (e.g., a blue strawberry), the model should ideally answer “blue” in the first case and “red” in the second. When computing PvP steering vectors, the visual input is always the *counterfactual image*. For a given layer, l , we extract the hidden representations at the MLP block for both prompts at each layer and compute two steering vectors, S_{CF}^l and S_{WK}^l :

$$S_{CF}^l = \frac{1}{D} \sum_i^D ([h_n^l]_i^{\text{this}} - [h_n^l]_i^{\text{most}})$$

$$S_{WK}^l = \frac{1}{D} \sum_i^D ([h_n^l]_i^{\text{most}} - [h_n^l]_i^{\text{this}}).$$

Here, $i \in \{1, 2, \dots, D\}$ represent the text-image pairs in Visual CounterFact and h_n represents the hidden state of the last text token in the sequence of a sample, typically an instruction token, which has been shown to store more important information when compared to specific subject tokens [37]. After computing the world knowledge (S_{WK}^l) and counterfactual (S_{CF}^l) steering vectors, we control the model’s output by modifying the hidden state of the final token in the sequence at a given layer in the language decoder. Formally, let h_n^l denote the hidden state of the last token at layer l in the language decoder of an MLLM. To steer the representation toward pixel-level information from the image, we apply the following intervention:

$$h_n^l = h_n^l + S_{CF}^l$$

To instead steer the model toward world knowledge priors, we apply:

$$h_n^l = h_n^l + S_{WK}^l$$

These interventions are applied for all $l \in [l, l + w]$. Our method for calculating multimodal steering vectors captures the representational shift needed to modulate the model’s reliance on vision input versus world knowledge priors.

3.5 Results

3.5.1 MLLMs are Distracted By Counterfactual Images

We begin by analyzing how MLLMs behave when presented with counterfactual (CF) images that intentionally contradict common object priors, alongside baseline world knowledge (WK) images that reflect real-world

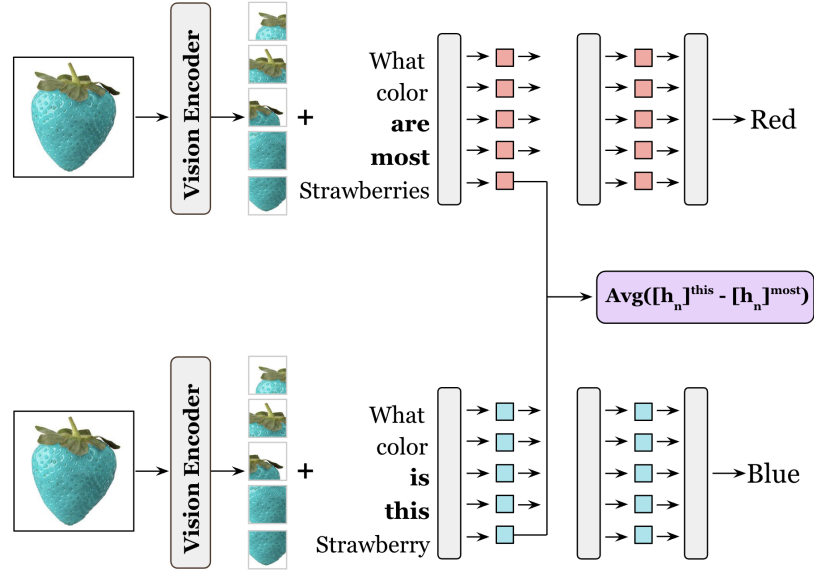


Figure 3.2: Pixel Versus Prior steering vectors are created by contrasting representations of prompts that emphasize pixel-level information (“this”) versus priors (“most”), using the last hidden state.

visual expectations (Figure 3.1). To test whether models rely more on memorized knowledge or on the current image, we use two types of prompts: “What color are **most** <objects>?” and “What color is **this** <object>?”

All models perform well on “this” prompts, achieving over 80% accuracy even when the input image presents a counterfactual. This indicates that MLLMs are highly effective at grounding their answers in the current visual input. Errors in this setting typically involve subtle hue disagreements such as gold versus orange or yellow, rather than confusion about the underlying object property.

In contrast, the “most” prompts reveal a critical weakness. When asked about what is generally true, models are expected to retrieve world knowledge rather than attend to the current image. This behavior holds when WK images are shown, but accuracy drops sharply when the same question is paired with CF images. In these cases, models often abandon their prior knowledge in favor of what is visually presented, even though the prompt clearly targets a generic concept. This suggests that MLLMs are easily distracted by the current image, even when instructed to generalize.

3.5.2 Localizing Visual Perception Shifts through Early Decoding

To understand how this visual override emerges during inference, we apply *early decoding* to track model predictions across layers. This reveals when the model transitions from relying on memorized priors to integrating counterfactual visual input. Figure 3.3 shows that in the color task when the model is prompted for the world knowledge answer but given a counterfactual image, the probability of the world-knowledge answer rises in mid to late layers, then flips to the counterfactual answer (orange) in the final layers. This “flipping behavior” is most common when the model is prompted to respond with the world knowledge answer and provided with a counterfactual image (Figure 3.3 Panel B). This delayed integration of visual input leads to

Model	Task	CF + “this”	WK + “this”	CF + “most”	WK + “most”
LLaVA-Next	Color	85.19	87.22	47.26	92.09
	Size	82.12	96.42	40.30	95.60
Qwen2-VL	Color	84.79	85.40	60.65	90.87
	Size	91.20	98.21	28.34	96.29
Janus-Pro	Color	86.00	88.03	59.23	90.47
	Size	85.14	96.84	18.02	96.01

Table 3.1: Accuracy (%) for color and size tasks under “this” (e.g., “What color is this strawberry?”) and “most” (e.g., “What color are most strawberries?”) questions with counterfactual (CF) and world knowledge (WK) images. Models perform well when grounded in the current image, but accuracy drops sharply in the “most + CF” setting, indicating that MLLMs are overly influenced by misleading visual input.

	Size			Color		
	LLaVA-Next	Qwen2-VL	Janus-Pro	LLaVA-Next	Qwen2-VL	Janus-Pro
% Samples w/o Flip	45%	69%	70%	35%	71%	88%
% Samples w/ Flip	55%	31%	30%	65%	29%	12%
Avg. # CF → WK	1.02	0.47	0.76	0.84	0.18	0.56
Avg. # WK → CF	1.12	0.70	0.86	1.31	0.90	0.56

Table 3.2: Flip statistics for size and color attributes with counterfactual images and prompts designed to elicit world knowledge responses. A “flip” occurs when the initially less probable response later exceeds the alternative by at least 5%.

errors when the image contradicts memorized associations, matching the results in Table 3.1.

In contrast, when using “this” prompt (e.g., “what color is **this** strawberry?”) for identifying the counterfactual attribute, models are confident in the counterfactual answer by the middle layers and rarely flip to the world-knowledge alternative (Figure 3.3 panel A and Table 3.2). This confidence is supported by the high inference accuracies seen in Table 3.1. Despite their confidence in the counterfactual answer, there is still a slight spike in world knowledge answer probability in mid-to-later layers. This slight spike shows that memorized knowledge does not fully disappear from the model, even when presented with contradicting inputs.

Table 3.2 shows how often models alternate between world knowledge and counterfactual answers on the color and size tasks when provided with a counterfactual image and a “most” prompt. On average, LLaVA-Next flips from world knowledge to counterfact 1.24 times on samples where a flip occurs, compared to 0.79 in the reverse direction. This indicates that MLLMs are prone to overriding prior knowledge when presented with a counterfactual image. These results suggest a consistent pattern: models initially rely on linguistic priors rooted in world knowledge, and only later override these with visual evidence as processing progresses through the layers.

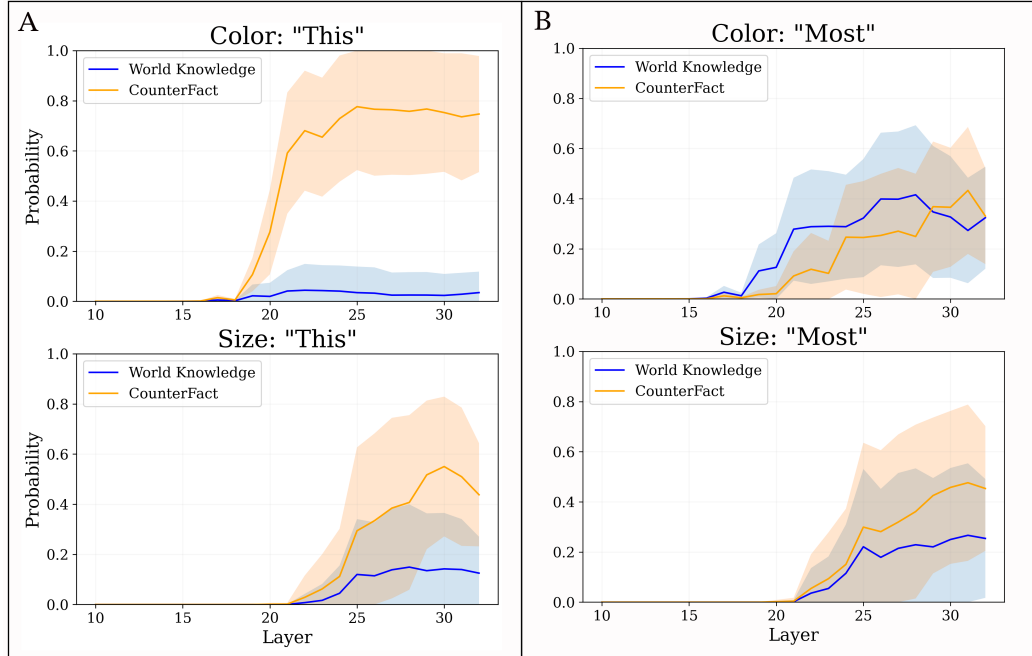


Figure 3.3: Early decoding results on LLaVA-Next show a conflict between answering “world knowledge” using priors or answering “counterfact”.

Table 3.3: Flip performance (%) under Color and Size tasks for two directions (WK $\rightarrow$ CF, CF $\rightarrow$ WK). Layer ranges indicate where flips occur.

Task	Dir.	LLaVA		QwenVL		Janus-Pro	
		Flips	Layers	Flips	Layers	Flips	Layers
Color	WK $\rightarrow$ CF	99.5	[14–16]	99.7	[17–19]	98.6	[14–16]
Color	CF $\rightarrow$ WK	86.4	[10–17]	78.8	[12–17]	78.2	[15–18]
Size	WK $\rightarrow$ CF	71.3	[8–16]	89.9	[16–22]	81.2	[16–19]
Size	CF $\rightarrow$ WK	33.5	[12–21]	61.8	[13–23]	70.37	[16–20]

3.5.3 Controlling World Knowledge Associations with PvP Steering Vectors

In Section 3.5.2, we show that MLLMs tend to rely on world knowledge in early layers and shift to visual information later, often flipping between the two. This delayed integration of visual input often results in unstable predictions when images conflict with prior knowledge (seen in Table 3.1). To stabilize predictions and control whether the model attends to the image or draws from prior knowledge, we use Pixel Versus Prior Steering (see Section 3.4.2). Practically, PvP steering offers an interpretable method to causally intervene in model processing, revealing the layers and temporal windows where the balance between vision and world knowledge can be effectively manipulated.

Table 3.3 shows the effectiveness of our steering approach across tasks and models, highlighting both the percentage of successful steering of the model and the key layers at which intervention has the highest impact. We apply PvP steering vectors to the set of inputs that the model originally gets incorrect, meaning

without PvP-steering, the model performance on this subset of data is 0%. Remarkably, we achieve at least a 98% success rate in flipping model predictions from world knowledge to counterfactual answers in the color task for all models. This demonstrates that MLLMs are not only steerable but highly responsive to targeted interventions, particularly when guided away from strongly encoded world knowledge priors. The size task is more difficult by nature: it requires detecting two objects and reasoning about their size relationship, making it more dependent on deeper, distributed visual processing. This is reflected in lower flip rates and broader intervention windows.

Across models, we observe that the most effective interventions tend to occur within specific mid-to-late layer ranges, typically requiring sustained influence over multiple layers (Table 3.3). In general, steering the model from world knowledge to a counterfactual (WK $\rightarrow$ CF) demands less intervention than reversing that shift (CF $\rightarrow$ WK), suggesting that overriding memorized priors is easier than restoring them once suppressed by a counterfactual input.

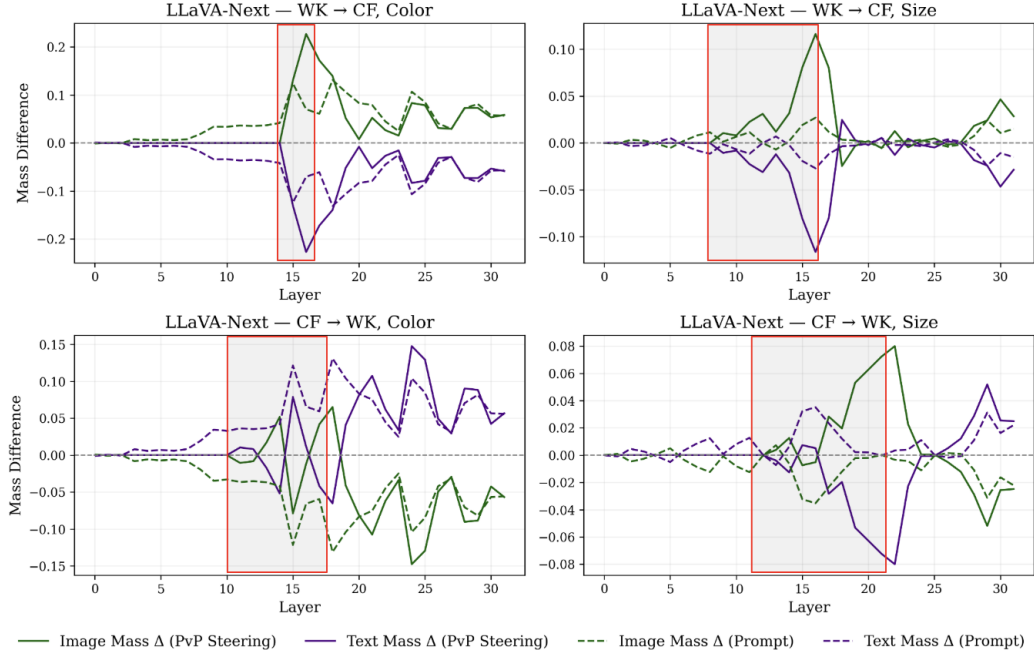


Figure 3.4: Effect of prompt changes and interventions on attention mass across layers for LLaVA-Next in the color and size tasks. Solid lines show changes when applying the steering vector; dashed lines show the effect of modifying the prompt. Green and purple lines represent attention shifts toward image and text tokens, respectively. The red shaded region highlights the layers where the intervention was applied (corresponding to Table 3.3). We see that intervention has a much stronger effect than changing the prompt.

3.5.4 Impact of Prompting and Steering Vectors on Attention Mass

While our results demonstrate that PvP steering vectors can reliably shift model outputs, we do not know how this shift is implemented internally. We hypothesize that this shift is implemented in the attention layers, as these components gather information from either the image or text tokens. To study the impact of steering

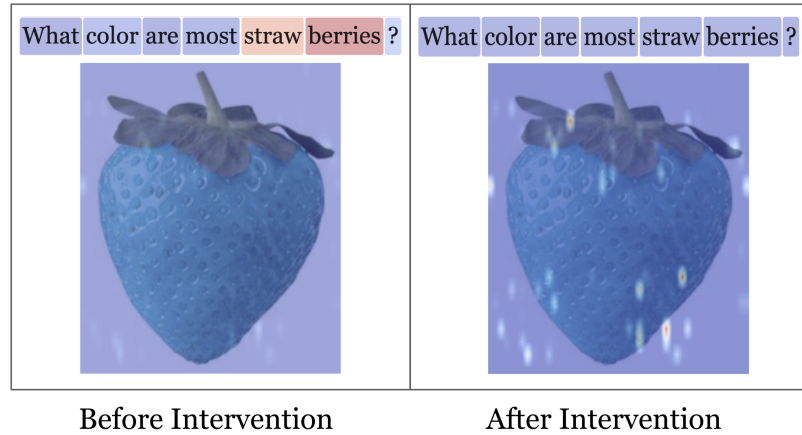


Figure 3.5: Impact of intervention on attention distribution. We observe more attention on text before intervention, and see a shift in attention towards the image and away from the text after intervention.

vectors on model predictions, we analyze their impact on the model’s attention patterns. We compare two settings: (1) changing the prompt from “most” to “this” (dashed lines), and (2) applying our PvP steering vector (solid lines) that steers the model toward or away from a counterfactual response. The first setting explores how changing the prompt shifts the model’s attention. Asking about the color of “most” strawberries should encourage the model to focus on prior knowledge (red), while asking about “this” strawberry directs attention to the specific pixels (blue). The second setting shows how injecting the PvP steering vector guides the model’s internal attention beyond the effect of the prompt.

As shown in Figure 3.4, changing the prompt from “most” to “this” yields a modest shift in attention toward image tokens. For example, on the color task, LLaVA-Next increases the attention mass to image tokens by 13%. In contrast, the PvP intervention vector causes a much stronger shift, increasing attention mass to image tokens to 40% (see Table A6 in Appendix A.13 for all models).

Among all models, LLaVA-Next shows the strongest and most consistent shifts in attention when interventions are applied, followed by Qwen2-VL and then Janus-Pro. For the size task, steering vectors must act earlier in the network and across more layers to be effective, reflecting the fact that size requires integrating more visual features than color. In contrast, the color task is more localized and easier to influence with a smaller intervention window. Figure 3.4 illustrates how attention moves across layers in response to both prompt changes and steering task interventions in LLaVA-Next. To illustrate the same effect but with a concrete example, Figure 3.5 shows the attention before intervention being heavily focused on text, with “strawberries” highlighted in red (highest attention). After intervention, attention shifts to the image, with most red regions being inside the image, rather than over the text.

These findings show that PvP steering vectors reshape internal attention mechanisms more effectively than prompt changes alone. They offer precise control over how models allocate attention to visual inputs, especially in tasks like size comparison that require broader spatial reasoning. By intervening directly in the model’s representation space, PvP steering enables deeper interpretability and control over MLLM behavior.

3.6 Conclusion

In this chapter, we investigate how multimodal large language models (MLLMs) reconcile memorized world knowledge and visual input. Understanding this balance is essential for building reliable models that can correctly choose between conflicting sources of information. To study this, we introduce *Visual CounterFact*, a dataset that constructs realistic visual counterfactuals targeting familiar attributes like object color and size. These examples violate learned priors while preserving visual plausibility, enabling precise comparisons between perception and memory. Using this dataset, we find that MLLMs often default to perception, even when prompted to retrieve general knowledge. In these cases, performance on knowledge-based prompts drops significantly, suggesting that models are overly influenced by visual inputs, even when the question targets memorized facts. Through studying the forward-pass, we observe that model predictions initially reflect stored priors, then transition to visually grounded answers in mid-to-late layers. This transition is often unstable, with models flipping between the two sources of information across layers. To control this behavior, we introduce *Pixels Versus Priors* steering vectors, which allow us to edit model behavior toward preferring either world knowledge priors or visual input. These activation-level interventions produce significant attention shifts towards or away from the image, depending on our steering vector direction. Our findings offer a new framework for interpreting and controlling MLLMs, advancing our ability to understand and control the interaction between memory and perception in multimodal models.

Chapter 4

Reducing Hallucinations through Attention Heads

In the last chapter, we present results from our work, *Mechanisms of Prompt-Induced Hallucination in Vision–Language Models* [127]. Continuing the work of the previous chapter on controlling attention, we turn to the problem of hallucinations and solve it by modifying attention heads. We study the hallucination failure mode in a controlled object-counting setting, where the prompt overstates the number of objects in the image (e.g., asking a model to describe *four* waterlilies when only *three* are present). At low object counts, models often correct the overestimation, but as the number of objects increases, they increasingly conform to the prompt regardless of the discrepancy. Through mechanistic analysis of three VLMs, we identify a small set of attention heads whose ablation substantially reduces prompt-induced hallucinations (PIH) by at least 40% without additional training. Across models, PIH-heads mediate prompt copying in model-specific ways. We characterize these differences and show that PIH ablation increases correction toward visual evidence. Our findings offer insights into the internal mechanisms driving prompt-induced hallucinations, revealing model-specific differences in how these behaviors are implemented.

4.1 Introduction

Vision–language models (VLMs) often follow textual prompts even when they contradict visual evidence [21, 176, 52]. These hallucinations are a growing concern as VLMs are deployed in real-world settings with noisy or inaccurate user inputs. To better understand this behavior, we study *prompt-induced hallucinations* (PIH) in a controlled object-counting setting, where prompts systematically overestimate the number of objects in an image.

Figure 4.1 illustrates the phenomenon where, although the model correctly counts three waterlilies when asked directly, it hallucinates additional, nonexistent flowers when asked to describe a mismatched count (e.g., “Describe the four waterlilies”). This undesirable behavior is unlikely to be driven by limitations in numerical reasoning [172, 120, 126], since the model correctly solves the underlying counting task. Instead, this is

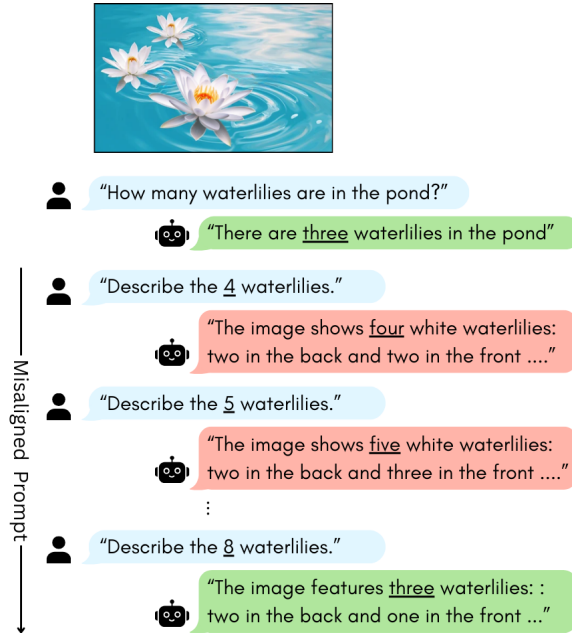


Figure 4.1: **Example of prompt-induced hallucination (PIH) in numerical reasoning.** The model correctly answers the baseline counting question. However, it conforms to the prompt and hallucinates additional waterlilies when asked to describe a mismatched number of objects. As the mismatch between the prompt and visual evidence grows, the model increasingly favors the image and recovers the correct count.

consistent with a known tendency of VLMs to rely on textual prompts over conflicting visual evidence [21].

To characterize PIH, we study **when** such errors occur and **how** they can be reduced via targeted analysis of attention-heads. Across models, PIH outputs follow a consistent and structured pattern. When the ground-truth number of objects exceeds four, VLMs are more likely to conform to the prompt and hallucinate missing objects. In contrast, at lower object counts, models are more likely to copy the prompt under small discrepancies but transition to visually grounded responses as the mismatch increases (Figure 4.1).

Through mechanistic analysis, we identify attention heads associated with PIH. Ablating these heads reduces hallucinations and restores visually grounded counts by up to 54%. Importantly, this intervention does not require additional data or training and alters only the object count, leaving the described objects unchanged. To test generalization beyond counting, we apply the same PIH head ablations to a color identification task [40] in which the prompt specifies an incorrect color. We find that these heads similarly mediate PIH in the color domain, mitigating prompt-color copying by up to 94.25%.

We further investigate the functional role of PIH attention heads and find that their ablation consistently reduces prompt-based copying, increases correction toward ground-truth responses, and shifts attention toward the image. While these effects are shared across all tested models, the mechanisms by which copying is suppressed differ. We make this distinction by examining how models format their responses, how attention shifts between text and image inputs, and how output probabilities change after ablation. Across models, these analyses show that PIH ablation consistently reduces direct copying from the prompt. While the way this reduction appears varies by model, it is always accompanied by increased reliance on visual information and

improved correctness. LLaVA-OneVision shows the strongest reduction in prompt-following after PIH-heads ablation, with clear suppression of prompt-consistent formatting, together with the largest increase in image attention and improved general counting performance.

Our contributions are threefold:

1. We introduce and analyze PIH, a failure mode in which VLMs prioritize prompt information over conflicting visual evidence. (section 4.3).
2. We identify a small subset of attention heads whose ablation substantially reduces PIH across models and tasks, without additional training or performance degradation on aligned prompts (section 4.4 & section 4.5).
3. We characterize the function of PIH-heads and show their ablation suppresses prompt copying while increasing correction toward visual evidence. (section 4.6).

4.2 Related Work

4.2.1 Prompt-Induced Hallucinations in LLMs

A substantial body of work has investigated hallucinations in large language models (LLMs) [61, 149, 54]. Prior work has shown that hallucinations can be reliably induced through out-of-distribution prompting [167], long context windows [76, 162], or the use of explicit personas [67, 139]. However, these settings often reflect adversarial or non-realistic usage patterns, rather than the models’ behavior under typical user interactions. Previous studies have demonstrated that LLMs may produce incorrect outputs when prompts contradict internal knowledge [160, 163] or contain errors subsequently repeated or amplified by the model [173, 139]. This behavior has been linked to cognitive-like biases, including anchoring bias, where models rely disproportionately on information introduced in the prompt [66, 26, 87], and sycophancy, where models conform to user assumptions despite conflicting evidence [135].

4.2.2 Modality Conflicts in VLMs

Vision–language models (VLMs) integrate textual and visual information, introducing additional complexity. Prior work shows that VLMs achieve higher accuracy on textual tasks compared to visual equivalents [28, 29, 151]. This gap is attributed to distinct internal mechanisms for visual and textual processing [102, 41] and to modality-specific parametric knowledge [179]. VLMs also struggle to disentangle different signals, both across their input modalities [52] and between their input and parametric knowledge [146, 40, 105].

Recent work investigated how VLMs perform under inconsistent or conflicting textual and visual inputs [164, 177]. Given conflicting information, current VLMs tend to disproportionately prefer textual information [128, 21], which can manifest in sycophantic behavior [114]. Further exploring modality conflicts, Zhang et al. investigated modality preference through the lens of unimodal uncertainty. In contrast, our work studies modality conflicts stemming from minor, realistic inconsistencies and examines a more complex task that more closely reflects real-world usage.

4.3 Prompt-Induced Hallucinations in Object Counting

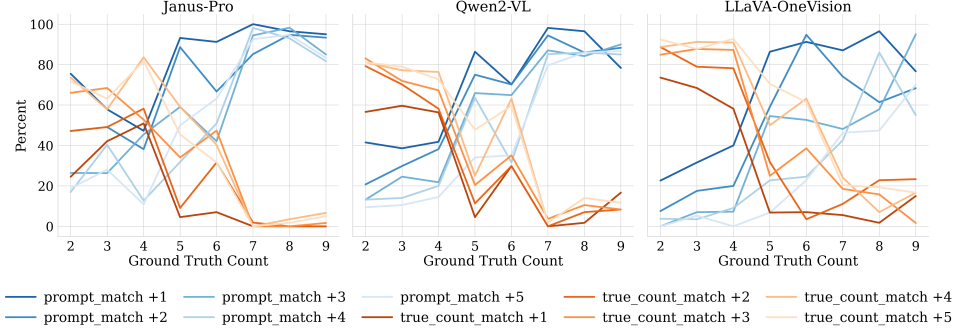


Figure 4.2: **PIH rates across different numbers of ground-truth objects.** Blue lines show the percentage of outputs matching the prompted number for different offsets, k . Orange lines show matches to the true count. As the object count increases, models stop correcting the prompt.

In this section, we present Prompt-Induced Hallucinations (PIH) and study when they manifest in the context of object counting. Object counting provides a clean setting for isolating prompt-image conflicts: ground-truth answers are verifiable, numerical deviations can be precisely controlled, and model errors can be explicitly measured. While a language model could legitimately answer the question "Describe 4 waterlilies" as part of an instruction-following task without hallucinating, for VLMs these responses constitute hallucinations: (1) we explicitly ground prompts with the phrase "in the image," and (2) the models describe additional objects not present in the image.

4.3.1 Experimental Setup

We investigate three VLMs covering a range of state-of-the-art model families: LLaVA-OneVision-7B [72], Qwen2-VL-7B [157], and Janus-Pro-7B [14]. We study counting using *CountBench* [106], a benchmark designed to measure object-count understanding in VLMs. For each image with ground-truth object count N , we first present a *baseline prompt*, P_B = "How many [objects] are there in the image?", retaining only samples where the model correctly predicts N . We then use *misaligned prompts* asking the model to "Describe the $N + k$ [objects] in the image," where $k \in \{1, 2, \dots, 5\}$ denotes the *discrepancy distance* between the prompt and the visual input. A PIH occurs when the model proceeds to describe $N + k$ objects without correcting the count, despite contradictory visual evidence.

4.3.2 PIH at Different Object Counts

Figure 4.2 shows the proportion of outputs that either match the prompted count ($N + k$, blue curves) or the true count (N , orange curves) across a range of object quantities and prompt offsets (k). The x -axis denotes the true number of objects (N), and the y -axis shows the fraction of responses aligned with either the prompt or ground truth. We observe a consistent pattern across all three models. For images containing small numbers of objects (2-4 objects), models typically resist misleading prompts and report the true count,

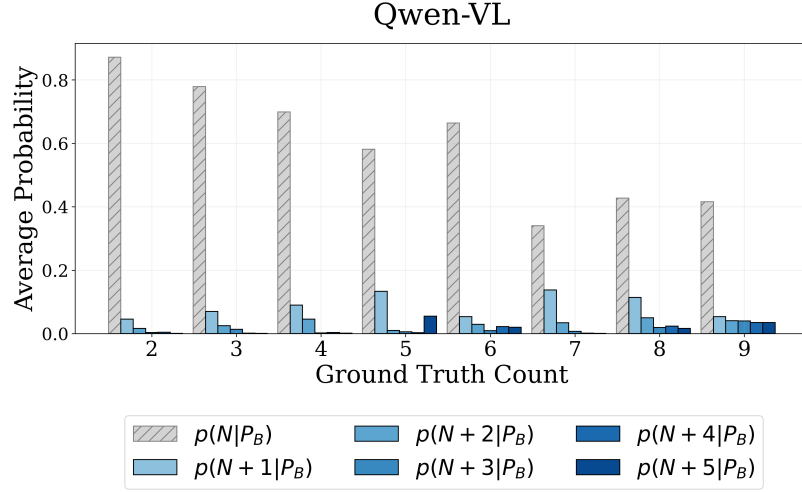


Figure 4.3: **Base probabilities on the object counting task.** We plot the probability assigned to the true count N and neighboring counts $N + k$ ($k \in \{1, \dots, 5\}$) given the prompt “How many [objects] are in the image?”. As object counts increase, confidence in N decreases and probability mass spreads to neighboring counts.

demonstrated by the orange curves being mostly higher than the blue curves. However, models still produce PIH in approximately 20–40% of cases, which primarily occur when the discrepancy is small ($N + 1$).

Beyond four objects, this behavior changes. Models increasingly conform to the prompt regardless of the discrepancy size, with prompt-matching responses approaching 80–90% even when the prompt substantially overestimates the true count. Accordingly, ground truth responses drop to near zero. In these cases, the discrepancy distance ceases to play a critical role, as shown by the convergence of the blue prompt-aligned lines. This effect persists for extreme offsets (with $k \in \{10, 20, 50\}$), with models readily describing, for example, fifty-nine cats in an image containing only nine (more details in Appendix A.16).

4.3.3 PIH Increase as Confidence Decreases

We hypothesize that models are more confident when counting fewer objects and therefore more resistant to misaligned prompts, with failures concentrated at minimal discrepancies ($k = 1$). We test this by collecting the probabilities for the true count N and neighboring counts $N + k$ under the baseline prompt (P_B). Figure 4.3 shows results for Qwen-VL, with similar trends for Janus-Pro and LLaVA-OneVision shown in Appendix A.17. We filter out all samples for which P_B is answered incorrectly, ensuring that the reported values reflect base confidence rather than raw accuracy. Indeed, for $N \leq 4$, the probability $p(N | P_B)$ is high while $p(N + k | P_B)$ is low, indicating strong confidence in the predicted count. At $N \geq 5$, base confidence drops and $p(N + k | P_B)$ increases, matching our PIH findings. Computing the Pearson correlation, we observe a moderately high ρ between the base probability and confirmation to the prompt (computed per-sample), with $\rho = 0.37, 0.46$ for Qwen-VL, and Janus-Pro, respectively, as well as a small positive correlation of 0.1 for LLaVA-OneVision. These results demonstrate that base probabilities are correlated with model behavior under prompt-image mismatches, suggesting that PIH is more prevalent when visual confidence is low.

	LLaVA-OneVision	Qwen-VL	Janus-Pro
Baseline Prompt: Exact Match ($\uparrow$)			
Before	76.89	78.49	80.32
After	81.24	79.29	79.41
Misaligned Prompt: Prompt Match ($\downarrow$)			
Before	42.58	56.51	64.10
After	1.42	3.22	10.19
Misaligned Prompt: True Count Match ($\uparrow$)			
Before	45.68	37.70	30.54
After	77.80	70.66	70.90

Table 4.1: **Response accuracies on baseline and misaligned prompts before and after PIH ablations.** The baseline prompt is “How many [objects] are there in the image?”. Misaligned prompts take the form “Describe the $N + k$ [objects] in the image” ($k \in \{1, 2, \dots, 5\}$) with results averaged across all k . Lower prompt match indicates reduced PIH, while a higher true-count match reflects stronger correction toward visual evidence. Prompt and true-count matches do not sum to 100%, as some responses contain no numerical value.

4.4 Locating PIH Mechanisms

To better understand and reduce PIH, we aim to locate model components responsible for PIH. Prior work has shown that attention heads often serve specific functional roles, such as information copying [104, 27] and context tracking [118], whereas MLP layers are more closely associated with storing parametric knowledge [91, 4]. Therefore, we focus our analysis on identifying attention heads associated with PIH.

4.4.1 Method: Attention Head Knockouts

To analyze the causal role of individual attention heads in PIH, we perform attention head knockouts using mean ablation (Wang et al. [156], Mueller et al. [100], see Appendix A.18 for technical details).

Intuitively, this intervention removes token-specific information carried by the head while preserving its overall activation magnitude. We apply this procedure head-by-head and quantify knockout success as the fraction of samples that switch from the prompted count to the true count under the misaligned prompt. We rank heads by success and select the top- m heads per model. We then perform a second-stage grouped knockout in which all top- m heads are ablated simultaneously for $m \in \{1, 3, 5, 10\}$, selecting $m = 3$ for Qwen-VL and $m = 10$ for LLaVA-OneVision and Janus-Pro as a per-model hyperparameter. We refer to the final set of per-model attention heads as **PIH attention heads**. See Appendix A.19 for a full list of top-10 heads per model and their success scores.

4.4.2 Attention Knockouts for PIH Mitigation

Table 4.1 shows models’ outputs given the baseline and misaligned prompts, before and after knockouts on the model-specific sets of PIH-heads. First, our results show that ablating PIH-heads does not disrupt standard counting behavior, as demonstrated by the exact-match accuracies on the baseline prompt. Notably, LLaVA-OneVision exhibits an improvement after ablation (+4.35 points), whereas Qwen-VL and Janus-Pro

show minor changes of approximately 1 point. Additionally, we verify that the object mentioned in the response remains unchanged and correct. This confirms that ablating PIH-heads does not impact general counting capabilities.

Under the misaligned prompts, ablating PIH-heads substantially reduces prompt-following behavior. Prompt match rates (previously corresponding to the blue curves in Figure 4.2) drop from 42-64% before intervention to below 11% for all models after ablation. Models also show a complementary increase in true count match rates, reaching 70-78% after intervention. This shift indicates that the ablated heads are responsible for propagating the incorrect numerical information specified by the prompt. Qualitative analysis of model generations before and after PIH-head ablation shows not only a reduction in incorrect numerical repetition but also the disappearance of spurious objects that were previously hallucinated and described in detail (examples are shown in Appendix A.20).

Our findings show that PIH is causally mediated by a small set of attention heads that propagate tokens from the prompt. Removing these heads shifts behavior from prompt reliance toward image reliance without harming general counting abilities, indicating that observed hallucinations are prompt-induced rather than due to counting failures.

4.4.3 Early-Layer LM Heads Mediate PIH

We observe that PIH-heads exhibit a consistent pattern in their layer-wise localization. Across all three models, top-10 PIH-heads (detailed in Appendix A.19) emerge in early and middle layers, suggesting that prompt-induced numerical information is introduced early and subsequently propagated forward. In both Qwen-VL and LLaVA-OneVision, the top-ranked PIH-heads are L0H3 and L0H6, while in Janus-Pro the most influential head is likewise located in the first layer (L0H20). This pattern extends beyond the top ranks: among the top-10 PIH-heads, 5/10 occur in the 0th layer for Qwen-VL, 7/10 for LLaVA-OneVision, and 3/10 fall within the first two layers (layers 0–1) for Janus-Pro. The early-layer concentration of PIH-heads suggests that numerical information from the prompt is injected before extensive multimodal integration or higher-level reasoning takes place, enabling it to shape downstream computation even when it conflicts with visual evidence.

This early-layer localization raises a related question about the source of these mechanisms: namely, whether PIH arises from vision–language interaction or from the language model itself. One way to interpret this question is through the overlap of PIH-heads between Qwen-VL and LLaVA-OneVision, which share the same underlying Qwen2 language model but use different vision backbones and image–language fusion architectures.

The top-ranked PIH-heads exhibit substantial overlap across the two models: the top-1 and top-2 heads overlap perfectly, and at $m = 10$, half of the identified heads are shared. This level of overlap strongly suggests that PIH is primarily mediated by language-model attention heads, rather than arising from vision-specific components or cross-modal alignment layers. This is consistent with prior work showing that fine-tuning tends to amplify or repurpose existing mechanisms rather than introduce entirely new mechanisms [e.g., 117]. Further, this supports the view that PIH reflects language-internal information routing that persists across multimodal models.

Response	LLaVA-OneVision		Janus-Pro		Qwen-VL	
	Before	After	Before	After	Before	After
NC (correct)	0.96	46.01	5.17	20.94	8.93	30.24
FC (correct)	0.00	33.23	9.61	24.14	11.34	39.86
NC (no color)	0.00	15.97	0.00	10.34	0.00	9.62
No PIH	0.96	95.21	14.78	55.42	20.27	79.72
EC (incorrect)	93.61	3.51	82.76	41.13	28.87	14.09
SC (incorrect)	5.43	1.28	2.46	3.45	50.86	6.19
PIH	99.04	4.79	85.22	44.58	79.73	20.28

Table 4.2: **Distribution of PIH and no PIH cases before and after PIH attention head ablation for the color task and semantic response types (%)**. NC, FC, SC, and EC denote no-, format-, soft-, and exact-copying, respectively. Correct denotes the ground-truth image color, while incorrect denotes the color in the misaligned prompt. Summed rows highlight the shift from PIH to image-grounded behavior after ablation.

4.5 Generalization Beyond Counting

Given that PIH appears to be mediated by shared language-model attention heads that persist across multimodal architectures, we next ask whether these same heads generalize beyond counting to mitigate prompt-induced hallucinations in other tasks. Specifically, we use a color prediction task based on the Visual CounterFact dataset [40], as color naturally allows gradual discrepancies. We use “What color is the [object] in the image?” as the baseline prompt, and “Describe the $C + k$ [object]” as the misaligned prompt, where C denotes the ground-truth color of the object and $k \in \{1, 2, 3\}$ denotes the perceptual “difference” from the C on the color wheel. Additional details on the color task are available in Appendix A.21. To test generalization, we use the same set of PIH-heads per model (Table 4.1) on the color task. Table 4.2 show that ablating the same PIH heads yields large improvements under misaligned color prompts across all models, reducing prompt-induced hallucinations by between 40% and 95%. This demonstrates that PIH heads identified through counting generalize beyond numerical reasoning and are not task-specific.

4.6 Analyzing the Function of PIH-heads

In this section, we analyze PIH-heads to better understand their functionality. We hypothesize that ablating PIH-heads induces one of two mechanisms: PIH-heads could reduce hallucinations by inhibiting copying behavior or by increasing attention flow to the image. Namely, PIH-heads could enable the model to propagate incorrect information from the prompt without referencing the image. Alternatively, PIH-heads may influence hallucinations by modulating the model’s reliance on the image, so that removing these heads increases reliance on visual evidence.

To determine the functionality of PIH heads, we first investigate their attention patterns over the input prompts and find that they do not reveal a consistent pattern connecting the generated answer to either copying the prompt or relying on the image (Examples shown in Appendix A.19.1). This is in line with several studies showing that influential attention heads may not exhibit human-interpretable patterns [59, 133, 11, 44].

Prompt: “Describe the 3 cats.” (N=2)	
Response	Copying Form
“There are 3 cats.”	Exact copying
“There are three cats.”	Soft copying
“There are 2 cats.”	Format copying
“There are two cats.”	No copying
Prompt: “Describe the purple banana.” (N=yellow)	
Response	Copying Form
“The purple banana . . .”	Exact copying
“The banana is purple.”	Soft copying
“The yellow banana . . .”	Format copying
“The banana is yellow.”	No copying

Table 4.3: Definition of copying forms under misaligned prompts for count and color tasks.

Therefore, we study the function of PIH-heads by examining the generated texts before and after knockouts, shifts in attention mass after PIH-ablation, and the impact of PIH-ablation on model confidence. Through these analyses, we show that PIH ablation reduces hallucinations by both **inhibiting copying** of information from the prompt and **increasing attention** to the image.

4.6.1 Ablation Impacts Copying Form

To characterize how models incorporate misaligned prompt information, we categorize several forms of copying behavior, summarized in Table 4.3. *Soft-copying* occurs when the response contains the misaligned prompt information but deviates in structure from the prompt’s structure. *Format copying* occurs when the response contains the correct (visually grounded) information, while maintaining the same prompt format. *Exact copying* occurs when both the prompt information and the format are present. Using these definitions, we analyze copying behavior first in the counting task and then in the color recognition task.

Table 4.4 shows copying patterns in the counting task, where the input prompts always include the number as a digit (i.e., “Describe the N_{digit} [objects]”). Prior to ablation, responses that include the number as a digit are more likely to be incorrect. For example, Janus-Pro outputs a digit in 43.95% of the cases, consisting of 43.45% of the incorrect cases, and 0.05% correct cases. This indicates that incorrect responses stem from exact copying of misaligned prompt information. After ablating PIH-heads, digit responses are almost entirely eliminated across all models. Notably, the reduction in digit responses occurs for both incorrect and correct outputs, showing that the suppression of format copying is not merely a byproduct of improved numerical accuracy.

The different copying forms in Table 4.2 show that PIH in the color task is likewise associated with copying

Model	Accuracy	Digit (%)		Word (%)	
		Before	After	Before	After
Janus-Pro	Incorrect	43.45	2.67	25.40	23.30
	Correct	0.50	0.14	30.65	73.89
	Total	43.95	2.81	56.05	97.19
LLaVA-OneVision	Incorrect	7.57	0.25	45.62	15.29
	Correct	0.30	0.25	46.51	84.21
	Total	7.87	0.50	92.13	99.50
Qwen-VL	Incorrect	28.44	4.53	33.11	16.94
	Correct	0.05	0.00	38.40	78.53
	Total	28.49	4.53	71.51	95.47

Table 4.4: **Response format and correctness before and after PIH attention head ablation.** Digit responses indicate copying format used in the prompt, while word responses correspond to spelled-out number expressions.

the format of the prompt. Before ablation, models frequently reproduce the prompt’s structure through format copying responses, accounting for over 90% of outputs for both LLaVA-OneVision and Janus-Pro. After PIH head ablation, this behavior is substantially reduced and accompanied by a corresponding increase in free-form descriptions.

These results suggest that when PIH-heads are active, models tend to reproduce both the content and the structure of misaligned prompts; when they are ablated, responses shift toward freer, less prompt-constrained descriptions that better reflect visual evidence. This parallel between counting and color tasks indicates that PIH is not limited to numerical reasoning, but reflects a broader tendency to copy information from the prompt, even when it is inconsistent with the image.

4.6.2 PIH-Ablation Shifts Attention Mass to Image Tokens

One notable outcome of PIH head ablation is observed in LLaVA-OneVision, where performance improves even for non-conflicting prompt–image cases. As shown in Table 4.1, baseline counting accuracy increases by 4.35%, despite the absence of misleading prompt information. Similarly, in the color task, PIH ablation yields a strong reduction in prompt-copying, with PIH reduced by 94%. These gains suggest that PIH ablation affects more than prompt-following, potentially altering how models balance textual and visual inputs.

To investigate this effect, we analyze changes in *attention mass* from assistant tokens to image tokens. For each layer, we compute the fraction of attention allocated to image tokens, with the remainder allocated to text tokens (summing to 1). Figure 4.4 shows the layer with the largest shift in attention from text to image, after PIH ablation: layer 2 in LLaVA-OneVision, 22 in Qwen-VL, and 27 in Janus-Pro; see Appendix A.22 for all layers. PIH head ablation increases image attention across all models, most strongly in LLaVA-OneVision ($\Delta = 0.121$) which shows an increase in an early model layer. This is significant, as previous work has shown the importance of visual information processing in early layer for multimodal integration [102]. These indicate that, particularly in LLaVA-OneVision, PIH head ablation induces a reallocation of attention toward visual

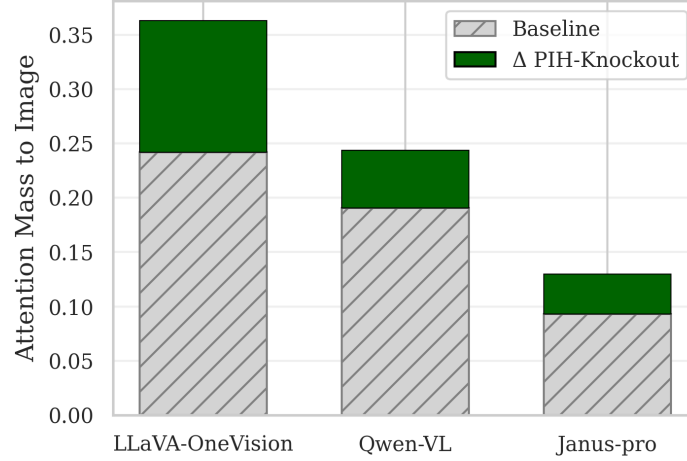


Figure 4.4: Image attention mass at the layer with the largest intervention-induced change for each model. Namely, LLaVA-OneVision (layer 2, $\Delta=0.121$), Qwen-VL (layer 27, $\Delta=0.053$), and Janus-pro (layer 22, $\Delta=0.037$). Gray bars indicate the baseline attention mass for image tokens, while green bars show an increase after PIH-knockout.

inputs, providing a plausible explanation for the observed improvements in both baseline counting accuracy and robustness to PIH.

4.6.3 PIH Heads Mediate Model-Specific Copying Mechanisms

Beyond analyzing changes in model outputs, we seek to understand how PIH head ablation affects ground-truth probabilities when presented with prompts that do not conflict with the image. We examine the conditional probability of the ground-truth count N when expressed as either a digit (N_{digit}) or a word (N_{word}) under prompts of the form “Describe the N_{digit} [objects]”, shown in Figure 4.5. We use a digit-word analysis because it offers a controlled way to distinguish format-copying from soft-copying by analyzing a single token.

Janus-Pro inhibits format copying when correcting misaligned prompts. Figure 4.5 demonstrates that $P(N_{\text{digit}} | N_{\text{digit}})$ significantly decreases when ablating PIH-Heads, which indicates a reduction in format-copying in the counting task. This is supported by the finding that PIH ablation in Janus-Pro reduces format copying by nearly 30%, a shift that is consistently accompanied by an increase in correct, free-form responses. While $P(N_{\text{digit}} | N_{\text{digit}})$ decreases, PIH ablation increases $P(N_{\text{word}} | N_{\text{digit}})$, indicating that when correcting misaligned prompts, Janus-Pro more frequently produces the ground-truth response in a word-based, free-form format rather than copying the digit form used in the prompt. Additionally, our findings show that PIH ablation produces only a modest shift in attention mass toward the image, consistent with almost no change in baseline counting performance. Together, these results suggest that PIH ablation in Janus-Pro does not increase visual reliance but primarily inhibits prompt-based format copying when correcting misaligned prompts.

Qwen-VL encourages format copying when correcting misaligned prompts. In contrast to Janus-Pro, we find that PIH-ablation causes Qwen-VL to more heavily rely on the prompt *format*. Figure 4.5 shows

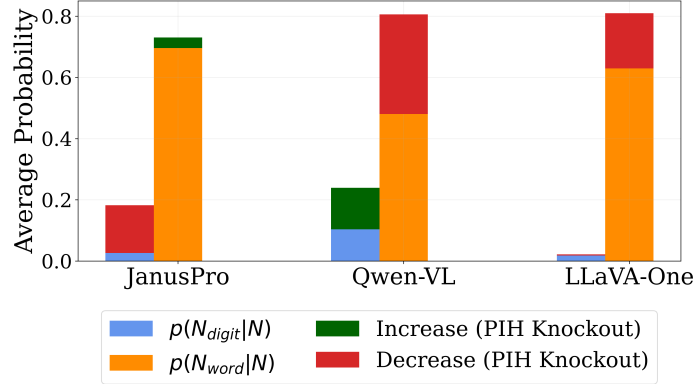


Figure 4.5: **Impact of knockout on the probability of the correct count in digit (blue) and word form (orange).** We measure the probabilities given the prompt “Describe the N_{digit} [objects]”. The knockout effects are shown in green (increase) and red (decrease).

a 32.6% decrease in probability of $P(N_{\text{word}} | N_{\text{digit}})$ with a corresponding *increase* in corrected responses that follow the digit format of the prompt. Namely, $P(N_{\text{digit}} | N_{\text{digit}})$ increases by 13.6%. This is consistent with Table 4.2 which shows that Qwen-VL is the only model for which PIH head ablation increases format copying, with copying rates rising from 40.21% to 53.95%. While PIH ablation increases image attention and improves correction performance in Qwen-VL, Figure 4.5 reveals that these gains arise from a shift away from soft copying toward increased format copying, rather than a strict reduction in copying.

LLaVA-OneVision inhibits all forms of copying and shows a greater reliance on the image when correcting misaligned prompts. Figure 4.5 and Table 4.4 show that LLaVA-OneVision rarely produces digit-form responses even prior to PIH ablation, indicating format copying suppression. Next, Figure 4.5 demonstrates that $P(N_{\text{word}} | N_{\text{digit}})$ heavily decreases (18%), indicating LLaVA-OneVision also suppresses soft copying. In terms of image reliance, LLaVA-OneVision is the only model that, after PIH ablations, showed a substantial increase in performance on the baseline counting prompt, indicating an improvement in general counting abilities. In parallel, we find that PIH-ablation leads to the largest shift in attention mass toward the image (+12%), indicating a greater reliance on the image.

Although implemented differently across models, PIH head ablation consistently reduces prompt copying and increases visual reliance, highlighting the promise of targeted interventions.

4.7 Conclusion

In this chapter, we study prompt-induced hallucinations in VLMs through a controlled object-counting task, isolating text-image conflicts. Through mechanistic analysis of three models, we show that such hallucinations stem from systematic prompt-based copying rather than failures of visual perception, and can be substantially reduced by ablating a small set of attention heads. We find that PIH is localized primarily to early layers of the language model, where prompt-based numerical information is likely processed and propagated before multimodal integration. Additionally, we find that Qwen-VL and LLaVA-OneVision share many of these

influential heads, indicating they largely originate from the shared language-model backbone rather than vision-specific components.

Ablating PIH heads consistently shifts behavior toward visually grounded corrections, generalizes beyond counting to color prediction, and does not degrade general performance. Finally, we find that PIH heads mediate different forms of prompt copying with model-specific implementations. LLaVA-OneVision exhibits the strongest effects, including broad suppression of all forms of prompt-copying and substantially increased reliance on visual information after PIH head ablations. Our results highlight PIH as a mechanistically localized and model-dependent failure mode in VLMs, suggesting targeted interventions can improve visual grounding under prompt-image mismatches.

Chapter 5

Conclusion and Outlook

This thesis investigated the role of attention mechanisms in multimodal deep learning models, with a focus on addressing three core challenges: scalability, interpretability, and controllability. As multimodal systems become increasingly central to real-world applications such as healthcare and decision support, understanding how attention integrates, prioritizes, and propagates information across modalities is essential for building models that are both effective and trustworthy.

In Chapter 1, I addressed the scalability limitations of existing multimodal attention paradigms. We introduced *One-Versus-Others* (OvO) attention, a novel integration mechanism that scales linearly with the number of modalities, in contrast to the quadratic complexity of self-attention and cross-attention. Through experiments across diverse clinical datasets with up to 6 modalities, we demonstrated that OvO achieves substantial reductions in computational cost, often exceeding 90% in FLOPs, while matching or improving predictive performance. These results show that rich cross-modal integration does not require expensive pairwise interactions, and that scalable attention mechanisms can unlock the use of multimodal models in high-modality domains such as healthcare.

In Chapter 2, I turned to the problem of interpretability, introducing *NOTICE*, a causal mediation framework for probing attention in vision-language models. By combining semantic corruption with activation patching, NOTICE enabled the identification of attention heads that perform consistent, human-interpretable functions across tasks and corruption schemes. Our analysis revealed that self-attention and cross-attention specialize into distinct roles, including object detection, object suppression, and outlier suppression, challenging the view of attention as a homogeneous or purely relational mechanism. These findings provide mechanistic evidence that attention heads can implement stable computational primitives, offering a foundation for causal interpretability in multimodal transformers.

Chapter 3 examined how attention mediates competition between memorized priors and new perceptual input. Using counterfactual visual evidence and early decoding, we showed that multimodal models often overweight misleading perceptual cues, even when they conflict with world knowledge. We introduced steering vectors that reallocate attention between visual tokens and linguistic priors, demonstrating that attention can be directly controlled to favor either perceptual grounding or factual recall. This work shows that targeted interventions can meaningfully shift model behavior without retraining.

In Chapter 4, I applied these insights to a concrete failure mode: prompt-induced hallucinations in vision-language models. We identified specific groups of early-layer attention heads that mediate prompt copying behavior, leading models to hallucinate objects or attributes implied by the prompt but absent from the image. Through targeted attention-head ablation, we reduced hallucinations across numerical and semantic tasks and observed a consistent shift in attention mass from text tokens to image tokens. Importantly, while the precise mechanisms of copying varied across architectures, the resulting behavioral shift toward visual grounding was shared, suggesting a common functional role for these heads across models.

Together, the contributions of this thesis provide a unified perspective on attention in multimodal systems. I show that attention is not merely a flexible alignment mechanism, but a structure that governs scalability, implements interpretable functions, and mediates critical trade-offs between competing information sources. By introducing OvO for efficient integration, NOTICE for causal interpretability, PvP steering vectors for control, and attention head ablations for hallucination mitigation, this work bridges theoretical understanding with practical intervention.

Several promising directions emerge from this work, many of which are already the focus of my ongoing research. First, the OvO attention formulation has been extended beyond supervised prediction to contrastive learning paradigms, enabling the integration of a broader set of medical modalities [39]. This extension positions OvO as a general mechanism for scalable multimodal representation learning. Moreover, because OvO assigns a single attention weight per modality, it naturally supports modality-level interpretability, allowing direct analysis of modality contributions in complex, high-modality settings.

Second, the NOTICE framework enables deeper causal investigations of multimodal attention. Ongoing work extends activation and path patching with Visual Counterfact [40] to study information flow between modalities. Another active direction, building on Visual Counterfact, focuses on color perception in multimodal models. By incrementally adding colored pixels to otherwise black-and-white images of objects, this work studies when both humans and models judge an object to be sufficiently colored, enabling direct comparison of perceptual thresholds and attention dynamics.

Finally, the interpretability tools developed in this thesis can be applied to safety-critical settings. Ongoing work uses attention-based causal analysis to study multimodal jailbreaking and adversarial prompting, and to extend these techniques to generative multimodal architectures such as diffusion models, where attention plays a central but poorly understood role in generation dynamics.

More broadly, this thesis suggests that progress in multimodal AI requires a deeper mechanistic understanding of the attention mechanisms that bind modalities together. In an era where multimodal systems are increasingly deployed in high-stakes settings, developing attention mechanisms that are scalable, interpretable, and controllable is critical. This work contributes both conceptual insights and concrete tools toward that goal, laying the groundwork for multimodal models that are not only powerful, but also transparent and reliable.

Acknowledgments

First and foremost, I would like to thank my advisors, Ritambhara Singh and Carsten Eickhoff, for their guidance, mentorship, and support throughout my PhD. The path they chose for me (multimodal learning) shaped my PhD and my future career. There is no doubt in my mind that joining their labs was the best decision I could have made. Their patience and commitment to my development shaped not only this thesis but also the way I think about mentorship and academia. I am deeply grateful for their ability to answer a call when I need them the most, their trust in me, their encouragement to grow independently, and the time and care they invested in me over the years. I would also like to acknowledge Chen Sun, for being on my committee since my first qualifying exam. The input and perspective he has offered have helped shape my research.

I am forever thankful to my partner, Noah Fischer, who has been by my side since before my PhD journey. I am still in awe that at 22, he chose to move across the country with me and fully support my decision to pursue a PhD. He listened to me practice every important presentation, stood by me through the inevitable ups and downs of academia, and helped me through tough decisions. During tight deadlines, he has made my life easier and taken care of me. He is the first person I share every accomplishment with. Noah has been both my biggest cheerleader and my voice of reason, and his unbelievable support made this journey possible.

I am grateful to my family for providing the foundation that made this achievement possible. Especially my parents, Olga and Boris Golovanevsky, who instilled in me a strong respect for education and learning, and they gave me both the resources and the confidence to pursue ambitious goals. There is no doubt that they are the reason I am passionate about what I do, and their pride propelled me forward. Thank you for every encouraging conversation and for always being there for me.

Finally, I would like to thank my friends and lab mates for making the PhD experience richer. Thank you to William Rudman, who has immensely shaped my PhD trajectory. We have authored four papers together in roughly one year, and I have no doubt that he will be a lifelong collaborator. Thank you to my friends and lab mates, especially Amina Abdullahi, Catherine Chen, Ruochen Zhang, and Nate Gillman, for being not only my close friends and inspiring researchers but also for always being a shoulder to lean on. Thank you to everyone at Brown for the conversations, collaboration, and celebrations along the way. Your support, humor, and camaraderie made even the most challenging moments more manageable, and I am grateful to have shared this journey with you.

This thesis is dedicated to my grandmothers, Vera Rovinsky and Laura Golovanevsky, who saw me begin this journey but did not get to see its completion. Their love, strength, and belief in me continue to guide and inspire me.

Bibliography

- [1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.
- [2] Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. Perceptionclip: Visual classification by inferring and conditioning on contexts, 2024. URL <https://arxiv.org/abs/2308.01313>.
- [3] Suraj Anand, Michael A Lepori, Jack Merullo, and Ellie Pavlick. Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. *International Conference on Learning Representations*, 2025.
- [4] Dana Arad, Hadas Orgad, and Yonatan Belinkov. Refact: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2558, 2024.
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.
- [6] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.
- [7] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL <https://aclanthology.org/2022.c1-1.7>.
- [8] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023.

- [9] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2), January 2024. ISSN 1091-6490. doi: 10.1073/pnas.2304406120. URL <http://dx.doi.org/10.1073/pnas.2304406120>.
- [10] Nathaniel Braman, Jacob WH Gordon, Emery T Goossens, Caleb Willis, Martin C Stumpe, and Jagadish Venkataraman. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V* 24, pages 667–677. Springer, 2021.
- [11] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.
- [12] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 565–580, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58538-9. doi: 10.1007/978-3-030-58539-6_34. URL https://doi.org/10.1007/978-3-030-58539-6_34.
- [13] Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891, 2022.
- [14] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL <https://arxiv.org/abs/2501.17811>.
- [15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- [16] Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*, 2023.
- [17] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations, 2023.
- [18] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL <https://aclanthology.org/P18-1198>.

- [19] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.
- [20] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024.
- [21] Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876, 2025.
- [22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- [23] Beddiar Djamil-Romaissa, Oussalah Mourad, and Seppänen Tapio. Automatic captioning for medical imaging (mic): a rapid review of literature. 2022.
- [24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [25] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*, 2020.
- [26] Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 12640–12653, 2024.
- [27] Sheridan Feucht, Eric Todd, Byron Wallace, and David Bau. The dual-route model of induction. *arXiv preprint arXiv:2504.03022*, 2025.
- [28] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.
- [29] Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*, 2024.
- [30] Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*, 2025.
- [31] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

- [32] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.
- [33] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models, 2024.
- [34] Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information, 2021.
- [35] Yoav Goldberg. Assessing bert’s syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.
- [36] Michal Golovanevsky, Carsten Eickhoff, and Ritambhara Singh. Multimodal attention-based deep learning for alzheimer’s disease diagnosis. *Journal of the American Medical Informatics Association*, 29(12):2014–2022, 2022.
- [37] Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. *arXiv preprint arXiv:2406.16320*, 2024.
- [38] Michal Golovanevsky, Eva Schiller, Akira Nair, Eric Han, Ritambhara Singh, and Carsten Eickhoff. One-versus-others attention: Scalable multimodal integration for biomedical data. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 580–593. World Scientific, 2024.
- [39] Michal Golovanevsky, Pranav Mahableshwarkar, Carsten Eickhoff, and Ritambhara Singh. Picme: Pipeline for contrastive modality evaluation and encoding in the mimic dataset. *arXiv preprint arXiv:2507.03165*, 2025.
- [40] Michal Golovanevsky, William Rudman, Michael Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfactuals. *arXiv preprint arXiv:2505.17127*, 2025.
- [41] Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation, 2025. URL <https://arxiv.org/abs/2406.16320>.
- [42] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.
- [43] Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild, 2023.

- [44] Christopher Grimsley, Elijah Mayfield, and Julia RS Bursten. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1780–1790, 2020.
- [45] Songhao Han, Le Zhuo, Yue Liao, and Si Liu. Llms as visual explainers: Advancing image classification with evolving visual descriptions. *arXiv preprint arXiv:2311.11904*, 2023.
- [46] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model, 2023.
- [47] Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR, 2022.
- [48] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.
- [49] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it’s harder to tell than you might think! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62. URL <https://aclanthology.org/2020.emnlp-main.62>.
- [50] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.
- [51] Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. In *European Conference on Computer Vision*, pages 257–273. Springer, 2024.
- [52] Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*, 2025.
- [53] Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Vlkeb: A large vision-language model knowledge editing benchmark. *arXiv preprint arXiv:2403.07350*, 2024.
- [54] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- [55] Nikolai Ilinykh and Simon Dobnik. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL*

- 2022, pages 4062–4073, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.320. URL <https://aclanthology.org/2022.findings-acl.320>
- [56] Nikolai Ilinykh and Simon Dobnik. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, 2022.
 - [57] Sergey E Ilyin, Stanley M Belkowski, and Carlos R Plata-Salamán. Biomarker discovery and validation: technologies and integrative approaches. *Trends in biotechnology*, 22(8):411–416, 2004.
 - [58] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015.
 - [59] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.
 - [60] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.
 - [61] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.
 - [62] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*, 2024.
 - [63] Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*, 2024.
 - [64] Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 2019.
 - [65] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
 - [66] Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.
 - [67] Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359, 2024.

- [68] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021.
- [69] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.
- [70] Andrew Kyle Lampinen, Stephanie CY Chan, Aaditya K Singh, and Murray Shanahan. The broader spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*, 2024.
- [71] Tiej Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Advances in Neural Information Processing Systems*, 36: 71195–71221, 2023.
- [72] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL <https://arxiv.org/abs/2408.03326>.
- [73] BOCHONG LI, Toshiya Nakaguchi, Yuichiro Yoshimura, and Ping Xuan. Robust multi-modal prostate cancer classification via feature disentanglement and dual attention. *Transactions of Japanese Society for Medical and Biological Engineering*, (Abstract):308–308, 2021.
- [74] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. URL <https://arxiv.org/abs/2407.07895>.
- [75] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [76] Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in) stability in language model dialogs. *The Conference on Language Modeling COLM 2024*, 2024.
- [77] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [78] Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.
- [79] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models, 2023.

- [80] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL <https://arxiv.org/abs/2310.03744>.
- [81] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [82] Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics*, 145:104466, 2023.
- [83] Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering, 2024. URL <https://arxiv.org/abs/2410.15778>.
- [84] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [85] Yujie Lu, Xiujun Li, William Yang Wang, and Yejin Choi. Vim: Probing multimodal large language models for visual embedded instruction following, 2023.
- [86] Grace Luo, Trevor Darrell, and Amir Bar. Task vectors are cross-modal. *arXiv preprint arXiv:2410.22330*, 2024.
- [87] Simon Malberg, Roman Poletukhin, Carolin Schuster, and Georg Groh Groh. A comprehensive evaluation of cognitive biases in llms. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 578–613, 2025.
- [88] Răzvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Polina Golland, Stefan Klein, et al. Tadpole challenge: Accurate alzheimer’s disease prediction through crowdsourced forecasting of future data. In *Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2*, pages 1–10. Springer, 2019.
- [89] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.
- [90] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.
- [91] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.
- [92] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

- [93] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 17–38. Springer, 2020.
- [94] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models, 2024.
- [95] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL <https://arxiv.org/abs/1301.3781>
- [96] Yue Ming, Xiyang Dong, Jihuai Zhao, Zefu Chen, Hao Wang, and Nan Wu. Deep learning-based multimodal image analysis for cervical cancer detection. *Methods*, 205:46–52, 2022.
- [97] Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. Spuriousity rankings: Sorting data to measure and mitigate biases, 2023. URL <https://arxiv.org/abs/2212.02648>.
- [98] Mohanad Mohammed, Henry Mwambi, Innocent B Mboya, Murtada K Elbashir, and Bernard Omolo. A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific reports*, 11(1):1–22, 2021.
- [99] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.
- [100] Aaron Mueller, Atticus Geiger, Sarah Wiegrefe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.
- [101] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.
- [102] Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*, 2025.
- [103] nostalgebraist. interpreting gpt: the logit lens. *LessWrong*, 2020. URL <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- [104] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- [105] Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. *arXiv preprint arXiv:2507.13868*, 2025.

- [106] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023.
- [107] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.conll-1.37. URL <http://dx.doi.org/10.18653/v1/2023.conll-1.37>.
- [108] Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2856–2861, October 2023.
- [109] Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. Finding and editing multi-modal neurons in pre-trained transformers. *arXiv preprint arXiv:2311.07470*, 2023.
- [110] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*, 2020.
- [111] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.
- [112] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.
- [113] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer’s disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010.
- [114] Renjie Pi, Kehao Miao, Li Peihang, Runtao Liu, Jiahui Gao, Jipeng Zhang, and Xiaofang Zhou. Pointing to a llama and call it a camel: On the sycophancy of multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20177–20191, 2025.
- [115] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.
- [116] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

- [117] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*, 2024.
- [118] Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. *arXiv preprint arXiv:2505.14685*, 2025.
- [119] Tanzila Rahman, Mengyu Yang, and Leonid Sigal. Tribert: Full-body human-centric audio-visual representation learning for visual sound separation. *arXiv preprint arXiv:2110.13412*, 2021.
- [120] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024.
- [121] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL <https://arxiv.org/abs/2408.00714>.
- [122] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*.
- [123] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [124] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [125] William Rudman, Catherine Chen, and Carsten Eickhoff. Outlier dimensions encode task-specific knowledge. *arXiv preprint arXiv:2310.17715*, 2023.
- [126] William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. Forgotten polygons: Multimodal large language models are shape-blind. *arXiv preprint arXiv:2502.15969*, 2025.
- [127] William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Ritambhara Singh, Carsten Eickhoff, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision-language models. *arXiv preprint arXiv:2601.05201*, 2026.
- [128] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022.

- [129] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11248–11257, 06 2022. doi: 10.1609/aaai.v36i10.21375.
- [130] Fawaz Sammani and Nikos Deligiannis. Interpreting and analyzing clip’s zero-shot image classification via mutual knowledge. *arXiv preprint arXiv:2410.13016*, 2024.
- [131] Stefan Schulz, Ann-Christin Woerl, Florian Jungmann, Christina Glasner, Philipp Stenzel, Stephanie Strobl, Aurélie Fernandez, Daniel-Christoph Wagner, Axel Haferkamp, Peter Mildemberger, et al. Multimodal deep learning for prognosis prediction in renal cancer. *Frontiers in oncology*, 11:788740, 2021.
- [132] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022.
- [133] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.
- [134] Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826, 2021.
- [135] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.
- [136] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.
- [137] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.
- [138] Luís A Vale Silva and Karl Rohr. Pan-cancer prognosis prediction using multimodal deep learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 568–571. IEEE, 2020.
- [139] Adi Simhi, Jonathan Herzig, Itay Itzhak, Dana Arad, Zorik Gekhman, Roi Reichart, Fazl Barez, Gabriel Stanovsky, Idan Szpektor, and Yonatan Belinkov. Hack: Hallucinations along certainty and knowledge axes. *arXiv preprint arXiv:2510.24222*, 2025.
- [140] Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36:27801–27819, 2023.

- [141] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL <https://arxiv.org/abs/2010.02502>
- [142] George Sterpu, Christian Saam, and Naomi Harte. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 111–115, 2018.
- [143] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- [144] Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*, 2022.
- [145] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [146] Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14333–14348. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.886. URL <https://doi.org/10.18653/v1/2023.emnlp-main.886>.
- [147] Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, 2022.
- [148] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.
- [149] SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6, 2024.
- [150] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.
- [151] Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. Same content, different answers: Cross-modal inconsistency in mllms. *arXiv preprint arXiv:2512.08923*, 2025.

- [152] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [153] Martina G Vilas, Timothy Schaumlöffel, and Gemma Roig. Analyzing vision transformers for image classification in class embedding space. *Advances in neural information processing systems*, 36: 40030–40041, 2023.
- [154] Martina G Vilas, Timothy Schaumlöffel, and Gemma Roig. Analyzing vision transformers for image classification in class embedding space. *Advances in neural information processing systems*, 36, 2024.
- [155] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.
- [156] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022.
- [157] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024. URL <https://arxiv.org/abs/2409.12191>
- [158] Sarah Wiegrefe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.
- [159] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=gl3D-xY7wLq>.
- [160] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2023.
- [161] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. doi: 10.1109/TPAMI.2023.3275156.
- [162] Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating llms’ belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, 2024.
- [163] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

- [164] Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*, 2025.
- [165] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023.
- [166] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org, 2023.
- [167] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.
- [168] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.
- [169] Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. *arXiv preprint arXiv:2310.15910*, 2023.
- [170] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.
- [171] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024.
- [172] Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. Quantity matters: Towards assessing and mitigating number hallucination in large vision-language models. *arXiv preprint arXiv:2403.01373*, 2023.
- [173] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.
- [174] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.
- [175] Tingwei Zhang, Collin Zhang, John X Morris, Eugene Bagdasarian, and Vitaly Shmatikov. Soft prompts go hard: Steering visual language models with hidden meta-instructions. *arXiv preprint arXiv:2407.08970*, 2024.

- [176] Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. *arXiv preprint arXiv:2511.02243*, 2025.
- [177] Jinkun Zhao, Lei Huang, and Wenjun Wu. What color is it? a text-interference multimodal hallucination benchmark. *arXiv preprint arXiv:2511.13400*, 2025.
- [178] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VI-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.
- [179] Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024.
- [180] Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.

Appendix

A.1 Computational Complexity Analysis for Multimodal Integration Schemes

In this section, we present the step-by-step details of the computational complexity analysis presented in Section 3.3. The analysis is done with respect to the size of the input modalities associated with the three paradigms used in our experimental setting: early fusion followed by self-attention, cross-modal attention, and One-Versus-Others (OvO) Attention.

A.1.1 Early Fusion

The early fusion approach involves first combining the modalities and then processing the concatenated sequence with the self-attention mechanism.

Step 1: Concatenation of Modalities.

Let k be the number of modalities and n be the feature-length of each modality.

$$\text{Total length after concatenation} = k \times n$$

The complexity for this operation is linear:

$$\mathcal{O}(k \cdot n)$$

Step 2: Compute Queries, Keys, and Values.

The self-attention mechanism derives queries (Q), keys (K), and values (V) for the concatenated sequence (length $k \cdot n$) using linear transformations with representation dimension, d . The complexity of each transformation operation is:

$$\mathcal{O}(k \cdot n \cdot d)$$

Step 3: Compute Attention Scores.

Attention scores are computed by taking the dot product of queries and keys. The self-attention mechanism has quadratic complexity with respect to the sequence length and linear complexity with respect to the representation dimension d [152]. Thus, given the concatenated sequence's length of $k \cdot n$ and the dimension of the keys and queries d , the complexity of this step is:

$$\mathcal{O}((k \cdot n)^2 \cdot d) = \mathcal{O}(k^2 \cdot n^2 \cdot d)$$

Step 4: Calculate the Weighted Sum for Outputs.

For each of the $k \cdot n$ positions in the concatenated sequence, we compute the softmax of the attention scores to produce the attention weights. These weights are then multiplied with their corresponding d -dimensional values to compute the weighted sum, which becomes the output. The computational complexity of these operations is:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

When combining all steps, the dominating terms in the computational complexity stem from the attention scores' computation and the weighted sum, culminating in an overall complexity of:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

A.1.2 Cross-modal Attention

For cross-modal attention, each modality attends to every other modality.

Step 1: Compute Queries, Keys, and Values for Inter-Modal Attention.

From a given modality, compute a query (Q), and from the remaining $k - 1$ modalities, compute keys (K) and values (V). Keys, queries, and values are obtained using linear transformations with representation dimension d . The complexity of each transformation operation is:

$$\mathcal{O}(n \cdot d) \text{ for each query, key, value set}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k - 1) \cdot n \cdot d)$$

The term $k \cdot (k - 1)$ comes from the number of pairwise permutations of k , given by ${}_kP_2 = \frac{k!}{(k-2)!} = k(k - 1)$.

Step 2: Calculate Attention Scores for Inter-Modal Attention.

The queries and keys from different modalities are used to compute attention scores, which represent how much one modality should attend to another.

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities [152]}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k - 1) \cdot n^2 \cdot d)$$

Step 3: Calculate the Weighted Sum for Outputs.

For every modality interaction, calculate the softmax of the attention scores to obtain the attention weights. These weights are then used in conjunction with the values vector to derive the weighted sum for the output:

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k - 1) \cdot n^2 \cdot d)$$

When evaluating all steps together, the dominating factors in computational complexity arise from the computation of attention scores and the weighted sum. Thus, the collective complexity for cross-modal attention, where each modality attends to every other, equates to:

$$\mathcal{O}(k \cdot (k - 1) \cdot n^2 \cdot d) = \mathcal{O}((k^2 - k) \cdot n^2 \cdot d)$$

For the complexity of cross-modal attention, the dominant term is k^2 . The $k - 1$ term effectively becomes a constant factor in relation to k^2 . As k tends toward larger values, the difference between k^2 and $k^2 - k$ diminishes. This is a consequence of the principles of big $\mathcal{O}$ notation, which focuses on the fastest-growing term in the equation while dismissing constant factors and lower-order terms. As a result, for asymptotic analysis, the complexity

$$\mathcal{O}(k^2 - k) \cdot n^2 \cdot d$$

can be simplified to:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

A.1.3 One-Versus-Others (OvO) Attention Complexity

Step 1: Averaging of "Other" Modalities.

Let k be the number of modalities and n be the feature-length of each modality. For each modality m_i , averaging over the other $k - 1$ modalities results in a complexity of:

$$\mathcal{O}(n)$$

Given that this needs to be computed for all k modalities:

$$\mathcal{O}(k \cdot n)$$

Step 2: Calculate Attention Scores with Shared Weight Matrix W .

The modality vector m_i and the average of "other" modalities, $\frac{\sum_{j \neq i}^n m_j}{n-1}$, are used to compute attention scores, which represent how much one modality should attend to the others. Multiplication with the weight matrix W (with representation dimension d) and the dot product with the summed modalities lead to:

$$\mathcal{O}(n^2 \cdot d)$$

Considering this operation for all k modalities:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

Step 3: Calculate the Weighted Sum for Outputs.

For every modality interaction, calculate the softmax of the attention scores to obtain the attention weights. These weights are then used in conjunction with the m_i vector (analogous the values (V) vector) to derive the weighted sum for the output:

$\mathcal{O}(n^2 \cdot d)$ for each pair of modalities

Considering all modalities:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

When evaluating all steps together, the dominating factors in computational complexity arise from the computation of attention scores. Thus, the collective complexity for cross-modal attention, where each modality attends to every other, equates to:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

In summary, One-Versus-Others (OvO) Attention exhibits a computational complexity that grows linearly with respect to the number of modalities ($\mathcal{O}(k \cdot n^2 \cdot d)$). In contrast, both early fusion through self-attention and cross-attention approaches demonstrate quadratic growth with respect to the number of modalities ($\mathcal{O}(k^2 \cdot n^2 \cdot d)$). This makes OvO a more scalable option for multimodal integration.

A.2 OvO Simulation Dataset Details

We consider two classes: (1) 20 random feature values that sum up to 1.0, and, (2) 20 random feature values that are each less than 0.15. The threshold was chosen at 0.15 because if 0.10 was the threshold, the mean of the 20 values would be 0.05, and thus, the sum would also be very close to 1, on average. This would render the task too difficult, and there would not be a significant difference between the samples across the two labels. Setting the threshold to 0.2 would render the task too easy, as on average, the numbers are consistently greater in the second class and the classes could be differentiated using only one modality. Thus, we chose 0.15 as the threshold.

A.3 Compute Resources for OvO Attention

For each experiment, we use one NVIDIA GeForce RTX 3090 GPU. For the MIMIC task, single-modality models ran for roughly 40 minutes, and multi-modal models ran for roughly 55 minutes on average. For the eICU, the single modality pre-trained models ran for roughly 50 minutes, the single modality neural network ran for a minute, and the multi-modal models ran for approximately an hour on average. For the TADPOLE task, single-modality models ran for 5 minutes, while multi-modal models ran for roughly 15 minutes on average. In the simulation dataset, the maximum modalities was 20 which took our model, OvO, roughly 2 minutes to run, while the cross-modal attention baseline took about 20 minutes to run on average.

A.4 Significance Testing for OvO Attention

We use a t-test to determine if there is a significant difference the performance metrics (AUROC, AUPRC, MAUC, BCA) means between OvO attention and the next best-performing multimodal model. Our sample

Table A1: Average runtimes for different tasks and model types using one NVIDIA GeForce RTX 3090 GPU.

Task	Model Type	Runtime (minutes)
MIMIC	Unimodal	40
MIMIC	Multimodal	55
eICU	Unimodal (pre-trained)	50
eICU	Unimodal neural net	1
eICU	Multimodal	60
TADPOLE	Unimodal	5
TADPOLE	Multimodal	15
Simulation	OvO (20 modalities)	2
Simulation	Cross and self-attention	20

size is 10 from each group, as we initialized the models with 10 random seeds. For the MIMIC IV and CXR dataset, we compare against self-attention as it performed the second best after OvO. Using an $\alpha = 0.01$, we have evidence to reject the null hypothesis and conclude that there is a statistically significant difference in means between single-attention and OvO attention. The p-value for the AUROC scores is 0.00363 and the p-value for AUPRC is 0.000948. For the TADPOLE challenge, we compare against cross-attention as it performed the second best after OvO. We get a p-value for MAUC scores of $2.09e^{-7}$ and a p-value of $8.19e^{-12}$ for BCA. Thus, we demonstrate a statistically significant difference in MAUC and BCA means between self-attention and OvO attention. Lastly, for the eICU dataset, we compare against cross-attention as it performed the second best after OvO. We get a p-value for AUROC scores of $2.00e^{-15}$ and a p-value of $8.24e^{-14}$ for AUPRC. Thus, we demonstrate a statistically significant difference in AUROC and AUPRC means between self-attention and OvO attention.

Table A2: Results of t-tests comparing the performance metrics between OvO attention and the next best-performing multimodal models across different datasets.

Dataset	Comparison Model	Metric	p-value
MIMIC	Self-attention	AUROC	0.00363
MIMIC	Self-attention	AUPRC	0.000948
TADPOLE	Cross-attention	MAUC	2×10^{-7}
TADPOLE	Cross-attention	BCA	8×10^{-12}
eICU	Cross-attention	AUROC	2×10^{-15}
eICU	Cross-attention	AUPRC	8×10^{-14}

A.5 NOTICE VQA-Prompts

For each task, we create a VQA prompt that asks the model to choose between a correct answer, which corresponds to the image, and an incorrect answer that corresponds to a semantically minimal pair. While in SVO-Probes, each subject, verb, or object, came with it’s own semantically different counterpart, in MIT-States dataset and Facial Expression dataset, we constructed such counterparts manually. In MIT-States, we separated

states into color, shape, and texture, and selected counterparts for each question from the respective categories. We grouped similar states (e.g., “huge”, “large”) together, and ensure that the selected counterpart would not be similar in meaning to the correct answer. Similarly, in the Facial Expressions dataset, we grouped the seven emotions into positive, negative, and neutral, and selected counterparts from opposing emotions, to ensure the model can correctly chose one answer. Recall that the SVO-Probes dataset consists of subject, verb, and object triplets. Each triplet is presented in singular form and non-conjugated. We use GPT-3.5 Turbo to perform grammatical error correction on all input prompts in SVO-Probes while minimally perturbing the template structure. We evaluate BLIP using exact match accuracy, which requires the generated answer to match the correct answer exactly, including capitalization and conjugation. BLIP achieves 66.1% accuracy on the facial expressions dataset, 44.6% on MIT States, and 16.1% on SVO-Probes (see Table A3 for category-wise accuracy). In our experiments, we use samples that BLIP accurately predicted. For each task, we select 500 samples across all categories: subjects, verbs, and objects for SVO-probes; colors, shapes, and textures for MIT-States; and one of the five emotions BLIP correctly identified for the facial expression identification task. To mitigate model bias toward the position of the correct answer in multiple-choice prompts, we equally distribute 250 samples where the correct option precedes and 250 where it follows the “or” token.

A.6 NOTICE Datasets Details

SVO-Probes. The Subject-Verb-Object (SVO) Probes dataset [48] consists of over 48,000 image-language pairs that vary on exactly one of the subject, verb, and object tuples that describe an image. For example, Figure 2.1 depicts image-text pairs with the captions “A child crossing the street” (child, crossing, street) and “A lady crossing the street” (lady, crossing, street). In this instance, the two image pairs only vary on the subject of the image, but the verb (crossing) and the object (street) are identical. SVO-Probes tests if multimodal models can capture fine-grained linguistic information about the content of an image.

MIT-States. The MIT States dataset [58] is a collection of 245 object classes, 115 attribute classes, and over 53,000 images. The dataset is used to study how models can generalize state transformations across different object classes. For instance, learning what “melted” looks like on butter can help recognize “melted” in chocolate.

Facial Expressions. The Facial Expression Recognition 2013 (FER-2013) [42] dataset includes 35,887 grayscale images of faces resized to 48x48 pixels. The images are categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The FER-2013 dataset is valuable for comparing feature learning methods with hand-engineered features in emotion recognition tasks.

A.7 BLIP and LLaVA Performance on NOTICE Datasets

In evaluating the BLIP and LLaVA, we focused on exact match accuracy, which necessitates that the model’s generated answer must exactly match the correct answer in terms of both content and form, including specifics like capitalization and conjugation. This stringent criterion ensures that the model’s understanding and response generation align closely with the intended semantics of the test prompts. Table A3 shows the performance of

Table A3: Comparison of BLIP and LLaVA accuracy (%) by dataset.

Dataset	BLIP	LLaVA
SVO Probes	16.10	40.89
MIT States	44.60	67.81
Facial Expressions	66.10	79.31

BLIP and LLaVA (zero-shot) on each task, and each category within the task.

A.8 SIP with Stable Diffusion

With the primary goal of establishing a more controlled environment conducive to the generation of semantically similar corrupt (negative) images from clean(positive) images, we also explore the use of stable diffusion to produce images that are similar to clean images from the SVO-Probes dataset, with the sole variation being in only one among the subject, verb, or object of the image.

Our experiment design utilizes the *stable-diffusion-2-inpainting* [124] to generate a corrupt image using the clean image as a base template. In this context, we define a prompt structure for the model as “a picture of $\text{sub}_{neg} + \text{verb}_{neg} + \text{obj}_{neg}$ ” where $\langle \text{sub}_{neg}, \text{verb}_{neg}, \text{obj}_{neg} \rangle$ denote the svo-triplet for the negative image.

For the parametric constraints on the diffusion model pipeline, we utilize a DDIM Scheduler [141], setting the DDIM value = 120 and the prompt guidance scale value to 8 for all the image generations. For generating images on the template of the base positive image, we use a contour mask, which allows for inpainting of the negative subject, verb, or object into the positive image. In such cases, a variation in quality is observable in the images introduced due to the probabilistic nature of the model.

Holistically, the generated images from *stable-diffusion-2-inpainting* are largely closer to the original corrupt images from SVO-Probes dataset where the image annotations utilize simpler terminologies to depict any one of the three $\text{sub}_{neg}, \text{verb}_{neg}, \text{obj}_{neg}$ such as “grass” instead of “meadow” (refer to Fig. A1 and A2).

A.9 NOTICE Result Details

A.9.1 SIP vs. Gaussian Noise Image Corruption.

Figure A3 illustrates the comparative effects of Semantic Image Pairs (SIP) and Gaussian Noise (GN) image corruption across all tasks, highlighting SIP’s superior ability to emphasize the significance of middle layers in the MLP, in contrast to GN’s focus on later layers. These observations affirm SIP’s enhanced reliability and effectiveness in revealing the nuanced functioning of vision-language models, supporting its use for in-depth interpretability studies.

A.9.2 Module-wise Activation Patching

Module-wise activation patching offers a detailed examination of how different components of the BLIP model respond to image and text corruption, using Semantic Image Pairs (SIP) and Gaussian noise for images

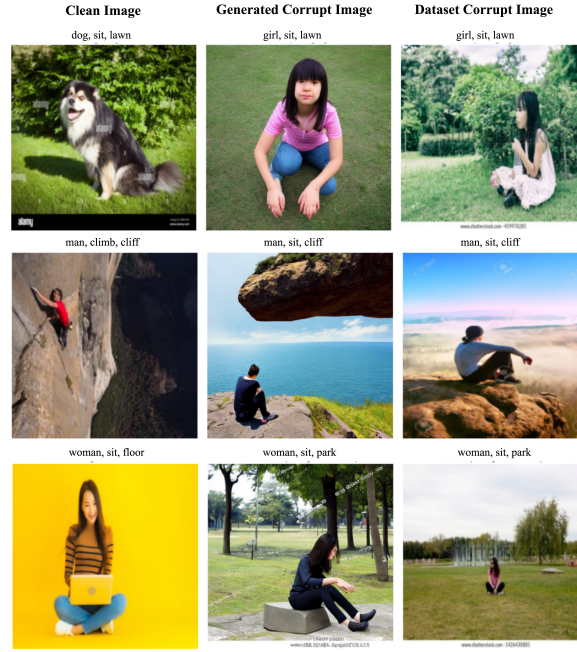


Figure A1: Successful examples of images generating using *stable-diffusion-2-inpainting* that showcase corrupt images close enough to the clean image yet different in (1) Subject (2) Verb (3) Object

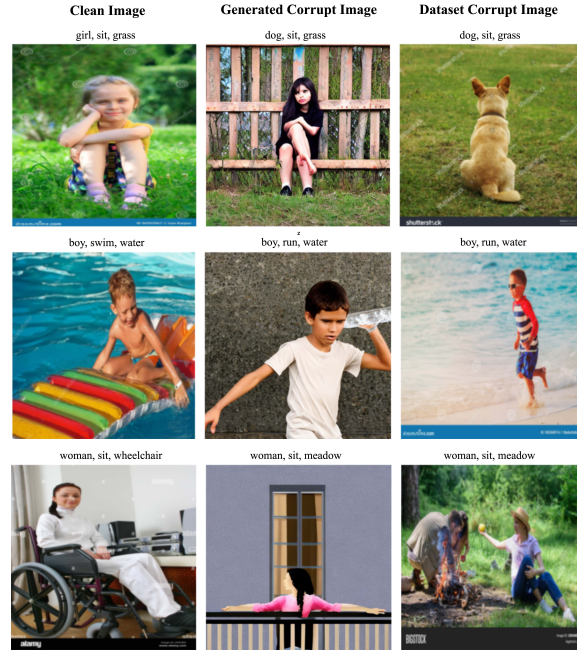


Figure A2: Poor examples of images generating using *stable-diffusion-2-inpainting* that showcase corrupt images which deviate from the clean image in (1) Subject (2) Verb (3) Object

and Symmetric Token Replacement (STR) for texts. This approach is crucial for uncovering the roles of self-attention, cross-attention, and MLP modules within the architecture, particularly focusing on their reaction

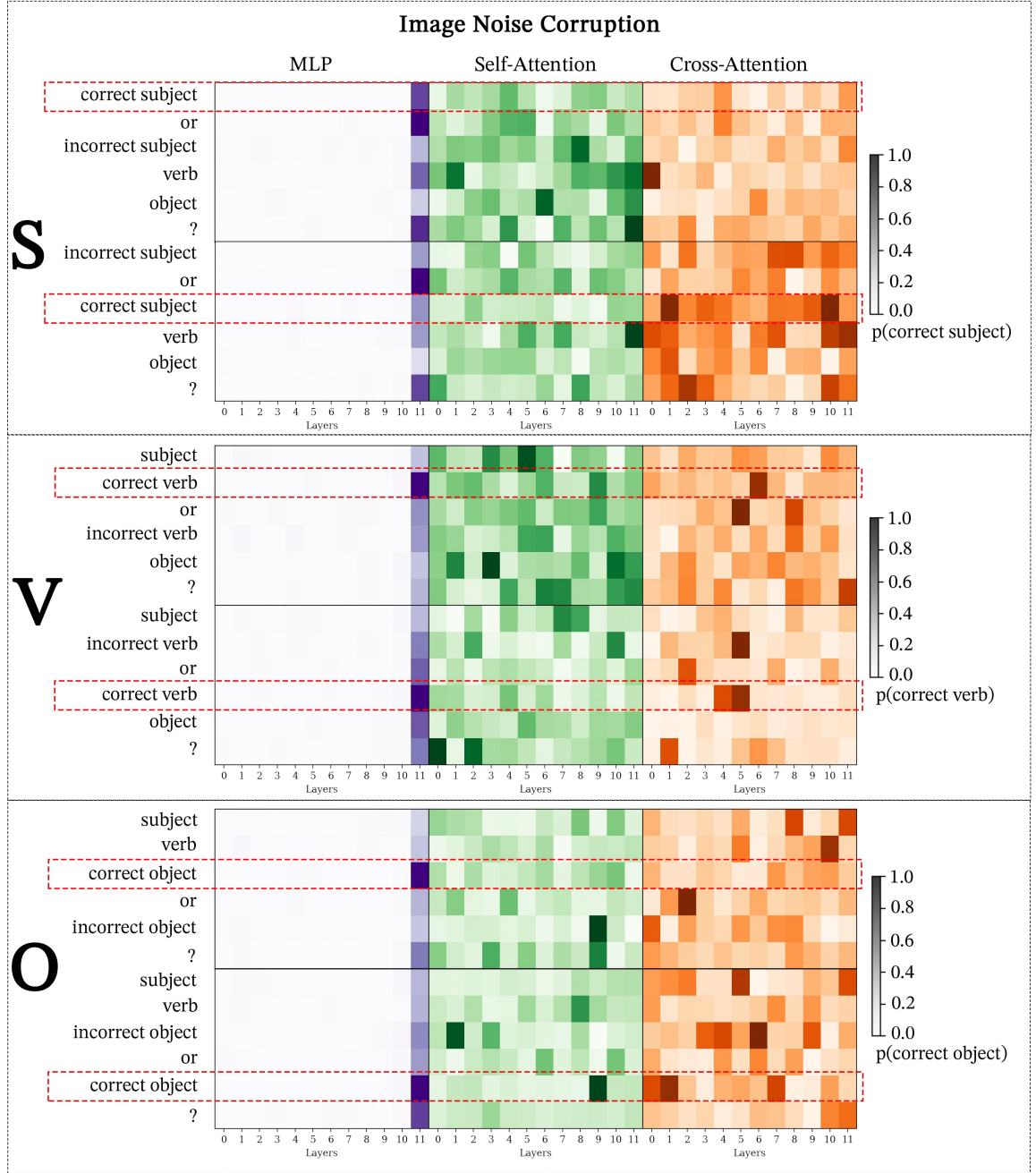


Figure A3: Comprehensive view of Gaussian noise effects across all tasks in the SVO-Probes dataset, providing a contrast to SIP effects and highlighting the differential impact of noise on model accuracy and layer activation.

to varied corruption methods. Through these experiments, we aim to pinpoint the layers most critical for the model's decision-making processes and explore how different corruption strategies affect the transparency and reliability of the model's internal workings. This analysis not only tests the robustness of the SIP technique against the traditional Gaussian noise approach but also highlights the strategic layers that are pivotal in

handling multimodal data integration.

First, Figure A4 and Figure A5 show module-wise activation patching for SVO Probes, broken down into subjects, verbs, and objects, VQA prompts.

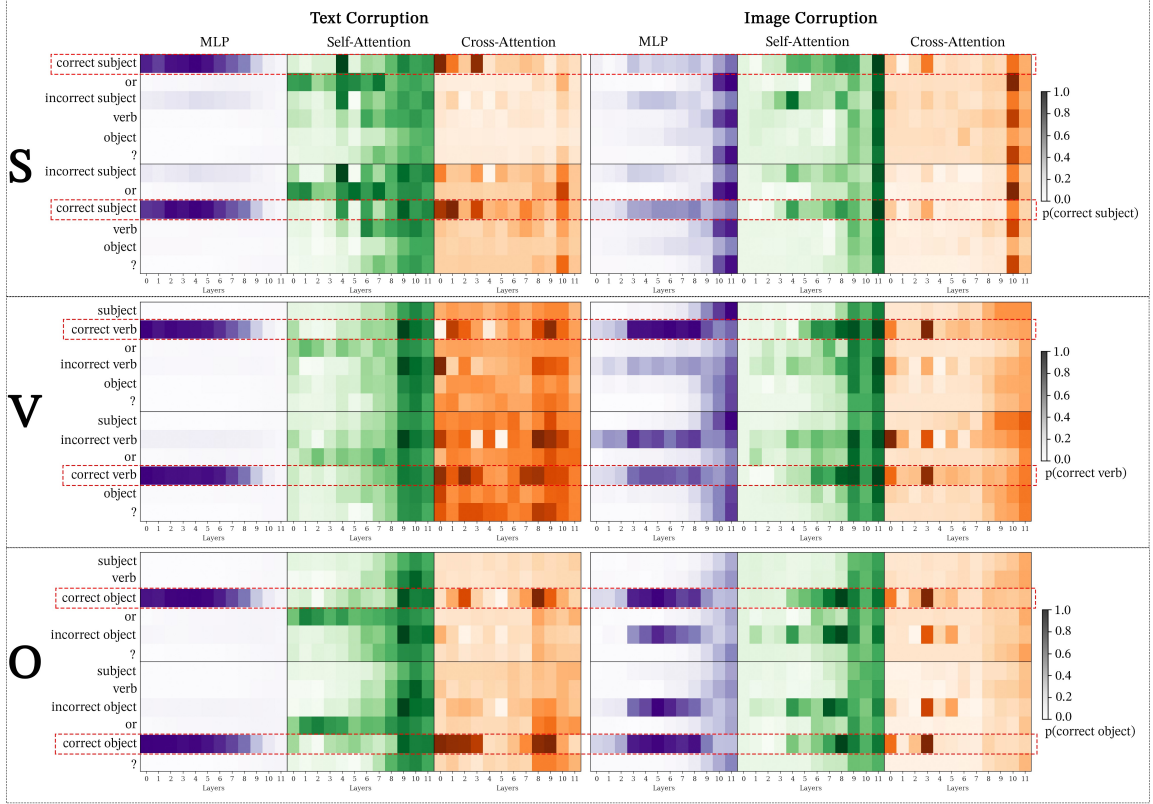


Figure A4: Module-wise activation patching results on BLIP for the SVO-Probes dataset.

Next, we examine module-wise activation patching for the Facial Expressions dataset in Figure A6 for BLIP and Figure A7. We note that image and text corruption results in similar patterns. This provides further support to SIP corruption aligning with STR corruption.

Lastly, Figure A8 and Figure A9 exhibit outcomes similar to the other tasks but with a notable distinction: there's an increased likelihood of accurately predicting the correct token within the cross-attention heatmaps during text corruption. This suggests that the multiple-choice VQA prompts used in the MIT States dataset may be more definitive compared to the other datasets. While distinguishing between emotions such as “angry” and “sad,” or identifying subtle differences like “woman” versus “girl” can be challenging in the Facial Expressions and SVO Probes datasets, respectively, the task of differentiating between clearly contrasting attributes like “modern” versus “ancient” buildings in MIT States appears to be more straightforward.

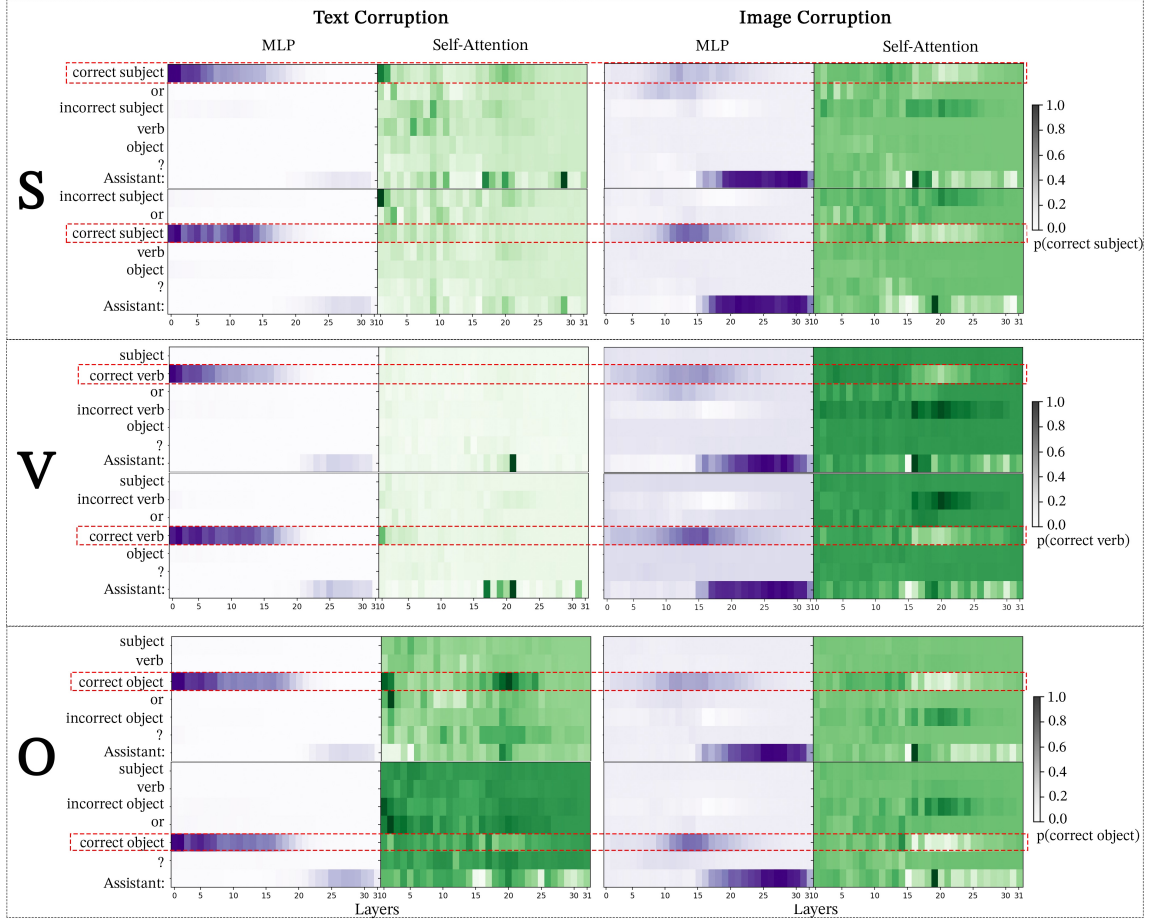


Figure A5: Module-wise activation patching results on LLaVA for the SVO-Probes dataset.

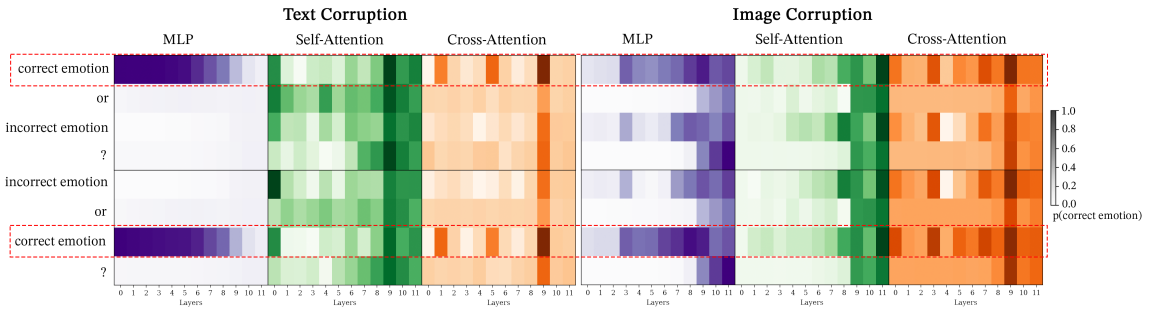


Figure A6: Module-wise activation patching results on BLIP for the Facial Expressions dataset.

A.10 LLaVA Self-Attention Patching Details

We consider two methods for performing activation on LLaVA’s self-attention heads. First, we patch only on the correct answer token. Although Figure 2.5 demonstrates a strong patching effect in BLIP, patching the correct answer token in LLaVA results in an extremely small logit difference, with the largest average

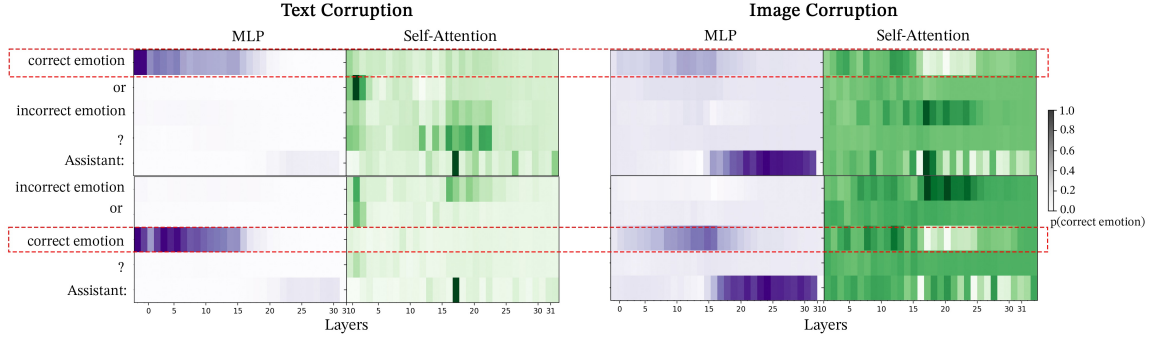


Figure A7: Module-wise activation patching results on LLaVA for the Facial Expressions dataset.

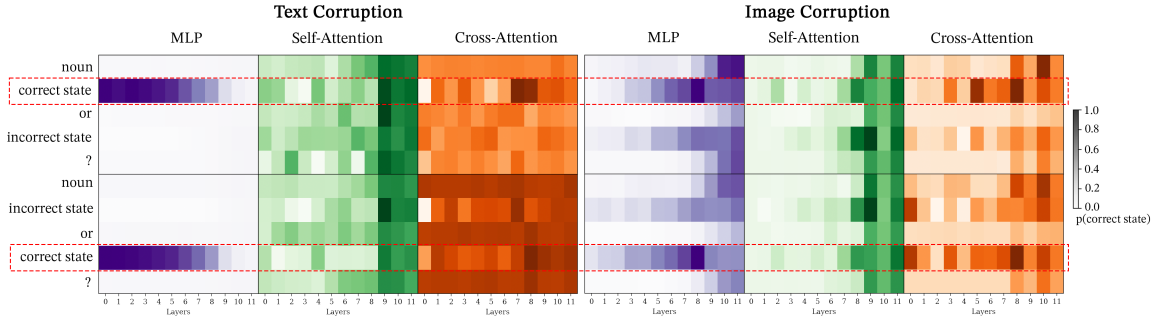


Figure A8: Module-wise activation patching results on BLIP for the MIT States dataset.

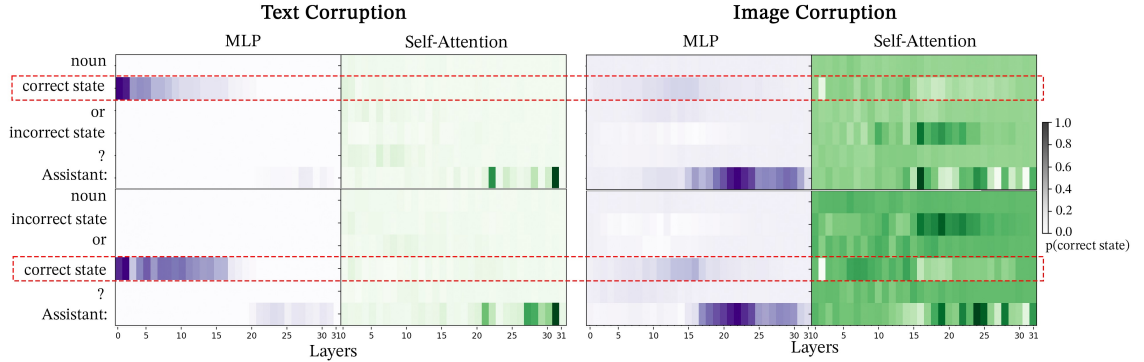


Figure A9: Module-wise activation patching results on LLaVA for the MIT States dataset.

logit difference being 0.016 as shown in Figure [A.10](#), regardless of modality corruption or dataset. Patching the correct answer token does not result in any “universal attention heads” indicating that self-attention does not play a pivotal role in associating the relevant images patches with the correct answer token in LLaVA. Module-wise patching using image corruption for both MLPs and Self-Attention blocks shows that patching the ‘Assistant:’ token consistently produces the strongest patching effect for LLaVA. Accordingly, we experiment with patching the ‘Assistant:’ token for LLaVA’s self-attention heads. Patching the “Assistant:” token creates a far stronger logit difference with some heads producing an average logit difference as high as 0.4. Patching

Table A4: Universal cross-attention heads identified via head patching.

Function	SVO	MIT	Facial Expr.
Object detection	L3.H0	L3.H0	L3.H0
Object suppression	L5.H3	L5.H3	—
Outlier suppression	L0.H11	L0.H11	L0.H11, L5.H3

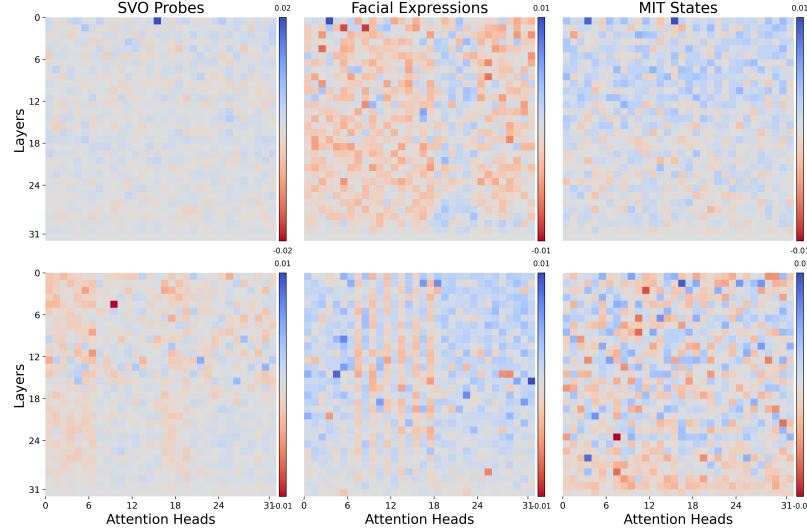


Figure A10: Logit difference when patching self-attention heads in LLaVA on the correct multiple choice option in SVO-Probes, MIT States and Facial Expressions. Top row shows SIP image corruption, the bottom row shows STR text corruption.

'Assistant:' produces consistent results with 6 universal self-attention heads compared to 0 when patching the correct answer token. Both the consistency and strength of patching the 'Assistant:' token highlight the importance of the 'Assistant:' token in storing information relevant to the image. Future studies performing activation patching on instruction-tuned VLMs that utilize causal self-attention to fuse modalities should focus on the information in the specific instruction tokens.

A.10.1 Comparing Cross-Attention and Self-Attention Head Functions

Our result that the cross-attention mechanism VLMs consistently associates the nouns with their corresponding image patches and that causal self-attention does not associate nouns to image regions is supported by previous works investigating cross-attention and self-attention patterns in toy VLMs [55]. Although self-attention heads in LLaVA associate neither the correct answer token nor the 'Assistant:' token with the relevant image patches, universal self-attention heads in LLaVA have the primary function of mitigating outlier attention patterns in images.

Important cross-attention heads in BLIP are scattered throughout the network ranging from layer 0 to layer 9. In contrast, import self-attention heads in LLaVA are concentrated in the middle layers of the network with x / y self-attention heads between 14-18.

A.10.2 Cross-Attention Head Function Classes

Understanding the specific functions of cross-attention heads within vision-language models (VLMs) is crucial for mechanistic interpretability. In this section, we delve into the distinct roles played by universal cross-attention heads identified by our NOTICE scheme. These heads can be categorized into three primary function classes: *implicit object detection*, *object suppression*, and *outlier suppression*. Below, we provide detailed examples of these functions along with corresponding visualizations from the SVO-Probes, MIT-States, and Facial Expressions datasets.

L0.H11 - outlier suppression. Universal cross-attention head L0.H11 implements outlier suppression in all tasks. It actively attends away from high-probability image patches and uniformly attends to the remaining patches. This function significantly reduces the mean of the average cross-attention outputs, thereby enhancing the model’s robustness by filtering out irrelevant visual information. L0.H11 does not produce a significant patching effect when images are corrupted, highlighting its role as a text-only universal cross-attention head.

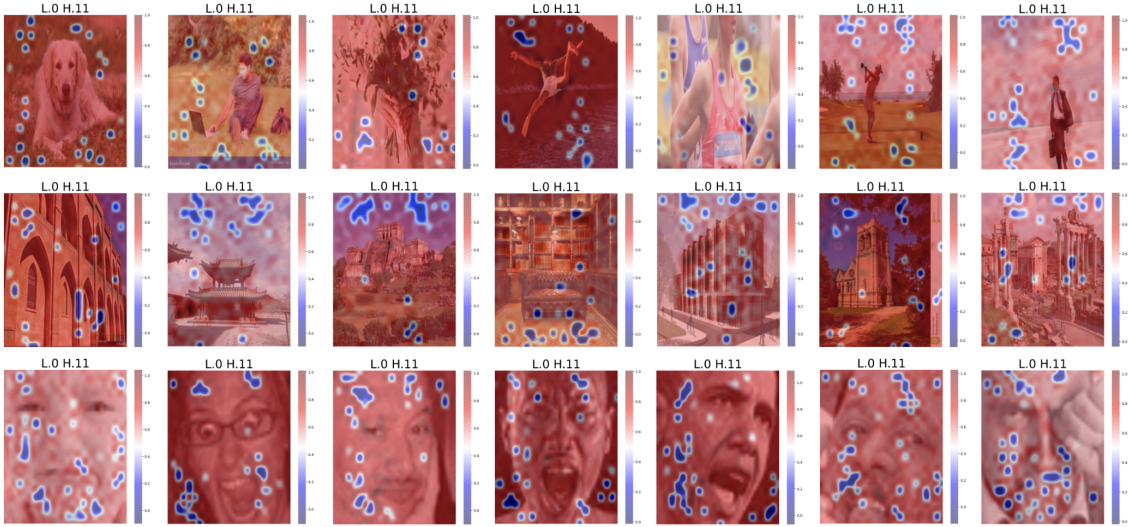


Figure A11: Cross-Attention patterns for universal cross-attention head L0.H11 on a sample of images from each dataset.

L3.H0 - Object detection. Cross-attention layer 3, head 0 (L3.H0) performs implicit object detection by associating text tokens with relevant image patches. It consistently demonstrates its importance across all tasks, reliably associating the correct emotion text token with the mouth region in the Facial Expressions dataset, and associating physical objects in images with their corresponding nouns in SVO-Probes and MIT-States. This head’s ability to map input text tokens to associated image patches underscores its critical role in fusing vision-language information.

L5.H3 - object suppression and outlier suppression. Cross-attention head L5.H3 performs object suppression by attending away from objects associated with an input token and attending strongly to outlier patches. In SVO-Probes, it attends away from subjects or objects, while in MIT-States, it attends away from relevant portions forming the noun that is changing states. In the Facial Expressions dataset, L5.H3 performs

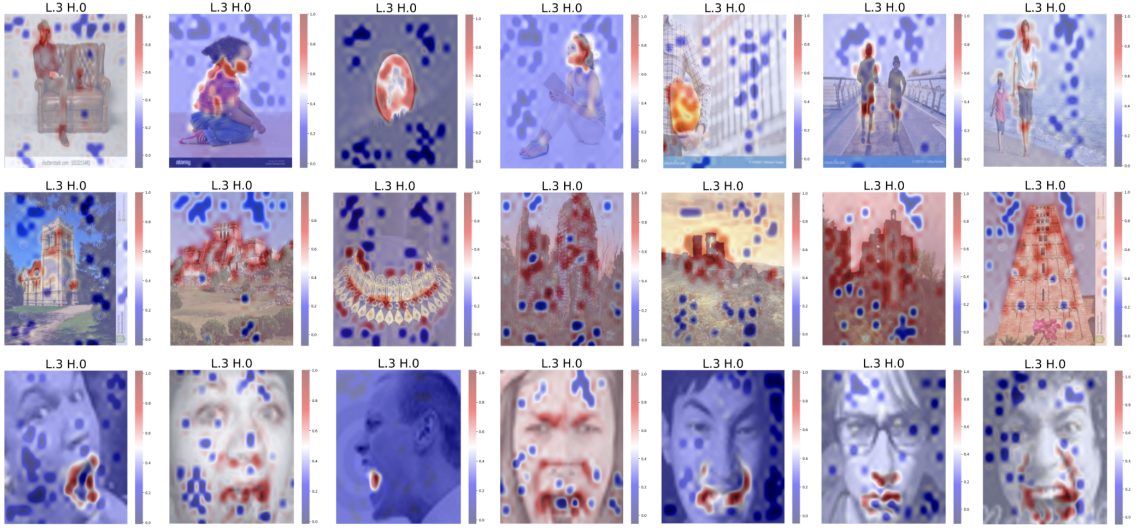


Figure A12: Cross-Attention patterns for universal cross-attention head L3.H0 on a sample of images from each dataset.

outlier suppression instead of object suppression.

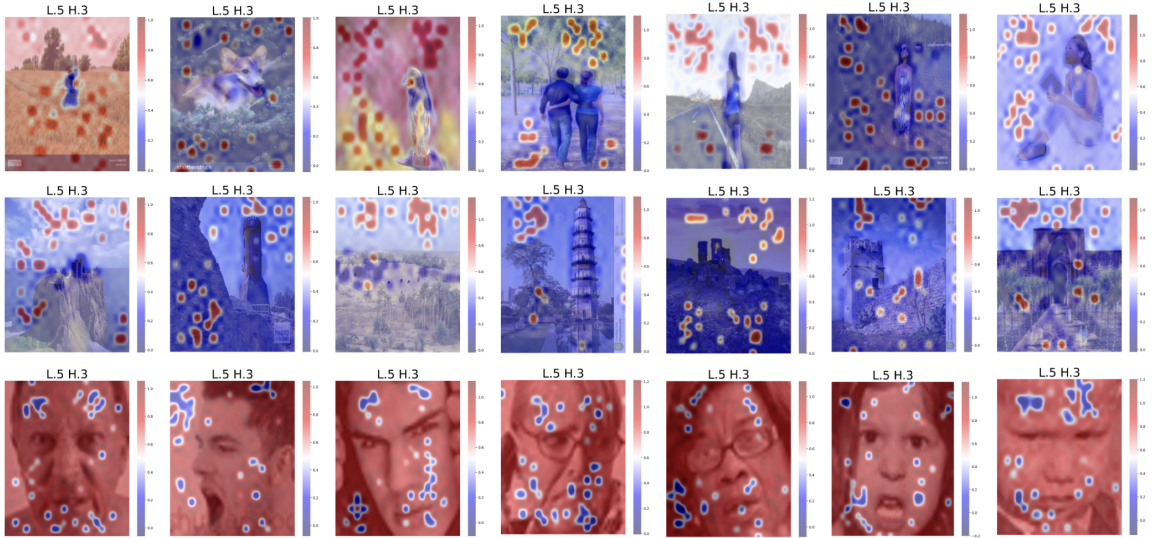


Figure A13: Cross-Attention patterns for universal cross-attention head L5.H3 on a sample of images from each dataset.

These examples underscore the diverse functions of cross-attention heads within BLIP, each playing a pivotal role in processing and integrating multimodal information. The identification and categorization of these heads provide a deeper understanding of the internal mechanisms driving VLMs, highlighting the significance of targeted model improvements and adaptations.

A.11 Creating Visual CounterFact

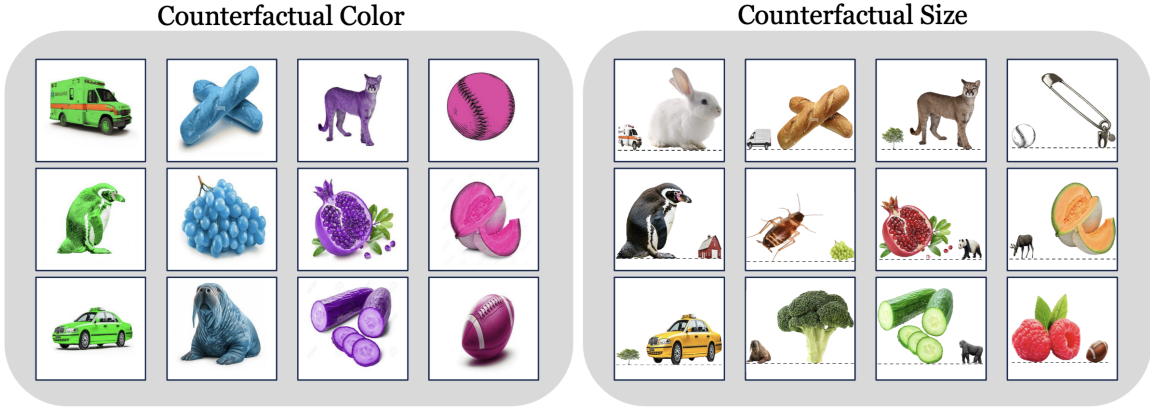


Figure A14: Examples from our dataset. (Left) images created using color relations, (right) images created using size relations.

Step 1: Sourcing Objects. We begin by identifying a set of real-world *objects* (e.g., *strawberry*, *squirrel*, *cherry*) from the McRae feature norms dataset [89], ImageNet [22], and CIFAR-100 [69]. The McRae dataset tasks 100 human participants with listing common attributes for an object. If at least 30% of participants respond with a specific color as a key attribute of that object, we include it in our dataset. After filtering with this 30% threshold, we obtain 190 objects. To increase the number of objects, we use the GPT-4o to elicit typical colors for ImageNet 1000 and CIFAR-100 classes and discard any categories lacking strong color ground-truth priors (e.g., clothing items). After both procedures, we obtain **622** unique objects.

Step 2: Sampling Real-World Images. We use the Google Images API to retrieve three candidate images of each object, specifying that they must be on a white background (a `<world knowledge color>` `<subject>` on a white background). Since prior work has shown that background information can heavily bias object detection models, we extract objects against a white background to ensure that model decisions are not influenced by spurious background features [159, 97, 166, 2]. Sampled images are then evaluated using GPT-4o, which scores them based on object correctness, color accuracy, presence of a white background, and overall realism. Specifically, we prompt GPT-4o with the following questions:

1. Is this an image of a `<color>` `<object>`? (yes/no)
2. Is this image on a white background? (yes/no)
3. Is this image an illustration or a realistic image? (illustration/realistic)
4. On a scale from 1 to 10, how realistic is this `<object>`? (numerical score)

We retain the highest-scoring image of the three to ensure visual fidelity with our inclusion criteria. For each yes/no question, the image receives a score of 0 for “no” or a score of 10 for “yes”. For images scoring 0 overall, we repeat the querying process but remove the world knowledge color (query: a `<subject>` on a white background), as most of the images resulting in a score of 0 are multi-colored (e.g., zebra, bee). After the second round of querying, we drop any remaining images with score 0, resulting in **575** unique objects.

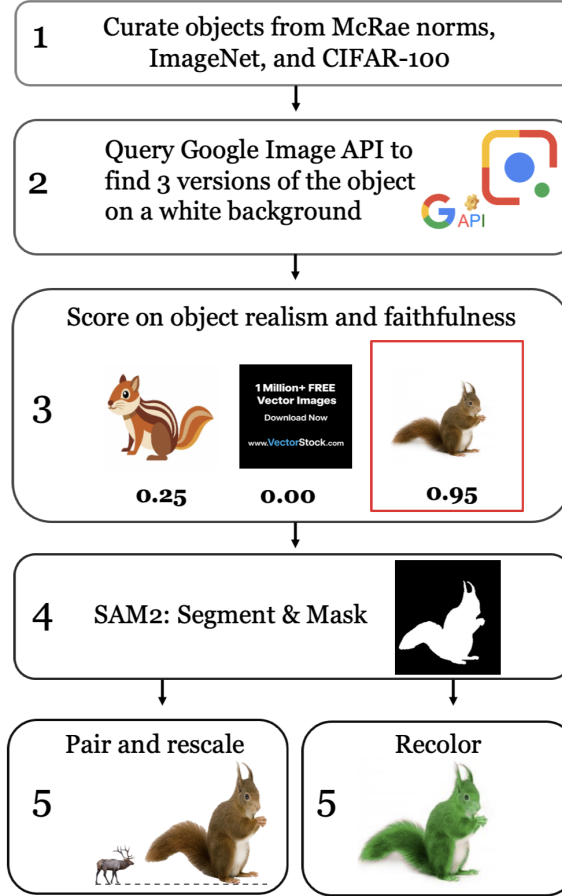


Figure A15: Construction pipeline for the Visual Counterfactual Dataset. We identify typical traits using semantic knowledge sources, retrieve realistic visual exemplars, and apply transformations to create perceptually plausible counterfactuals that conflict with language priors.

Step 3: Constructing Object-Size and Object-Color Counterfactuals.

(1) Color: To generate color counterfactuals, we prompt the LLaVA-Next model (Mistral-7B backbone) with “What color is a <object>?” and randomly sample from the five least likely color predictions (using common colors such as red, blue, green, pink, orange, etc.). This ensures that counterfactuals challenge the model’s linguistic priors, encouraging reliance on visual input rather than memorized associations. We constrain counterfactual colors to be perceptually distinct from the original color (e.g., avoiding red/orange or gray/black swaps). **(2) Size:** For the size task, we use the same set of objects from the color task and estimate their typical real-world dimensions using GPT-4o. The model provides height and width in inches, which we multiply to compute a total size metric. We then identify object pairs that differ in size by at least a factor of 10. For each object, we create two counterfactual images. Given three objects that satisfy $\text{object}_1 < \text{object}_2 < \text{object}_3$, where “<” denotes increasing real-world size, we generate two counterfactual images containing object_2 . Namely, $(\text{object}_1, \text{bigger_than}, \text{object}_2)$ and $(\text{object}_2, \text{bigger_than}, \text{object}_3)$. For example, if a squirrel is typically bigger than a cherry and smaller than an alligator, we create the counterfactual

images (cherry, bigger_than, squirrel) and (squirrel, bigger_than, alligator). This creates twice the number of samples since we construct two size relations for one object. After manual filtering of sizes that GPT-4o reported incorrectly, we are left with 877 unique size-object relations.

Step 4: Creating Counterfactual Images. Given a retrieved object on a white background, the first step in creating counterfactual images is to use SAM2 [121] to obtain a segmentation mask. After we obtain segmentation masks, we use two separate pipelines to create color and size counterfactuals, respectively.

(1) Color: Given a color-object relation with the relation “[object] is [color]” and its counterfactual relation “[object] is [counterfactual color]”, we apply a segmentation mask to isolate the object and modify only hue values in the HSV color space in order to change the color of the object to the counterfactual color while preserving the original saturation and brightness. This produces realistic and semantically surprising transformations (e.g., turning a red strawberry blue) while maintaining texture and shading. For objects with minimal hue (e.g., gray or black), we apply a set of hand-written remapping rules.

(2) Size: Given two objects with the world knowledge relation “[object 1] is larger than [object 2]” we create the counterfactual image reflecting “[object 2] is larger than [object 1]” by combining the segmentation masks of object₁ and object₂. For the world knowledge image, we resize the masks so that object₁ appears significantly larger than object₂, and in the counterfactual image, we resize the masks so that object₂ appears significantly larger than object₁. Specifically, 250×250 pixels versus 80×80 pixels. This size difference visually reflects the intended relation.

To make the size comparison clear and avoid depth ambiguity, we place both objects on the same horizontal baseline and add a black dashed reference line that both objects touch. This helps ensure that differences in perceived size are interpreted as scale changes rather than perspective shifts.

Visual CounterFact consists of 575 original (world knowledge) object images, 575 color counterfactual images, and 877 size original and 877 counterfactual images, totaling 2,904 unique images. Figure 3.1 shows examples from each split of the dataset. These transformations yield a dataset that explicitly conflicts with world knowledge priors of an object’s color and size while preserving perceptual plausibility, enabling targeted evaluation of visual reasoning models under counterfactual conditions. In Appendix Section A.12 we provide additional examples of images as well as dataset statistics on the kinds of objects we include in Visual Counterfact.

A.12 Visual Counterfact Dataset Statistics and Examples

Figure 3.1 illustrates representative examples from the Visual CounterFact dataset, including counterfactual edits based on object color (left) and relative size (right). Each example maintains visual realism while introducing semantically meaningful contradictions to typical object properties.

Table A5 summarizes the distribution of object categories in the dataset. The majority of counterfactuals involve animals, followed by a diverse set of objects spanning furniture, produce, vehicles, tools, and more. This broad coverage ensures the dataset tests model reliance on both visual input and memorized associations across varied semantic domains.

Category	Count
Animals	218
Household Items and Furniture	59
Fruits and Vegetables	58
Vehicles and Transportation	42
Electronics and Appliances	29
Tools and Hardware	22
Food and Drink (non-produce)	21
Buildings and Structures	21
Plants and Trees	19
Musical Instruments	15
Clothing and Accessories	15
Weapons and Military Items	13
Medical and Hygiene Items	11
Toys and Recreational Items	11
Natural Objects (non-living)	10
Office Supplies	6
Miscellaneous	5

Table A5: Distribution of object categories in the dataset.

A.13 PvP Attention Mass Details

To better understand how steering vectors and prompt changes affect the internal attention mechanisms of MLLMs, we visualize the change in attention mass over layers for each model and task. Figures A16 and A17 show these effects across color and size tasks, respectively. Each subplot compares attention mass difference toward image tokens (green) and text tokens (purple), with solid lines indicating PvP steering interventions and dashed lines indicating prompt-only changes.

We find that across all models and both tasks, interventions consistently produce stronger shifts in attention mass compared to prompt changes alone. In the color task (Figure A16), steering from WK to CF reliably increases image attention, while the reverse direction decreases it (as expected). The effect is particularly pronounced in LLaVA-Next, with peaks around the intervention window. The size task (Figure A17) shows a similar but more muted pattern, consistent with the task’s higher visual complexity. These trends reinforce that PvP steering vectors exert precise, causal control over how attention is allocated between vision and language streams.

Table A6 reports the maximum change in attention mass directed toward image tokens across models and tasks, comparing the effects of prompt changes (“most” to “this”) and PvP steering interventions. As described in Section 3.4.2, we measure the peak increase or decrease in image attention during inference.

Across all models, PvP steering consistently produces larger attention shifts than prompt modifications. This effect is most pronounced in the color task, where interventions increase image attention by up to 40% in LLaVA-Next, compared to 13% from prompting. As visualized in Figure 3.4, prompting leads to moderate reallocation of attention, while steering vectors induce strong and targeted redistribution.

We also observe an asymmetry between WK → CF and CF → WK directions: steering toward perception

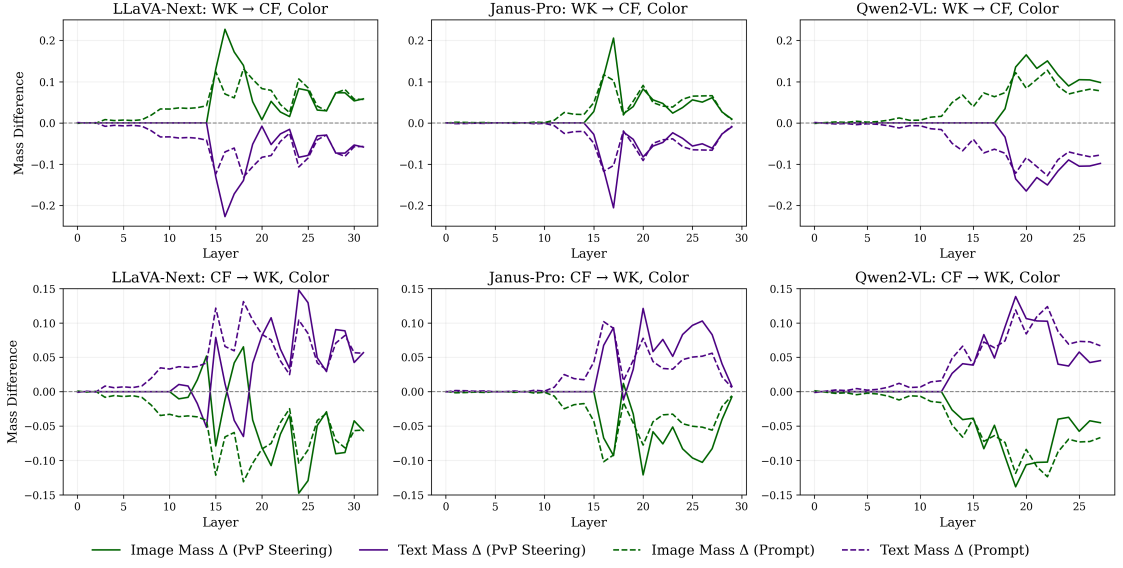


Figure A16: Attention mass difference across layers for all models on the color task. Solid lines show changes from PvP steering vectors; dashed lines show prompt-only changes. Green represents attention to image tokens, purple to text tokens. Each subplot shows one model and intervention direction (WK $\rightarrow$ CF or CF $\rightarrow$ WK).

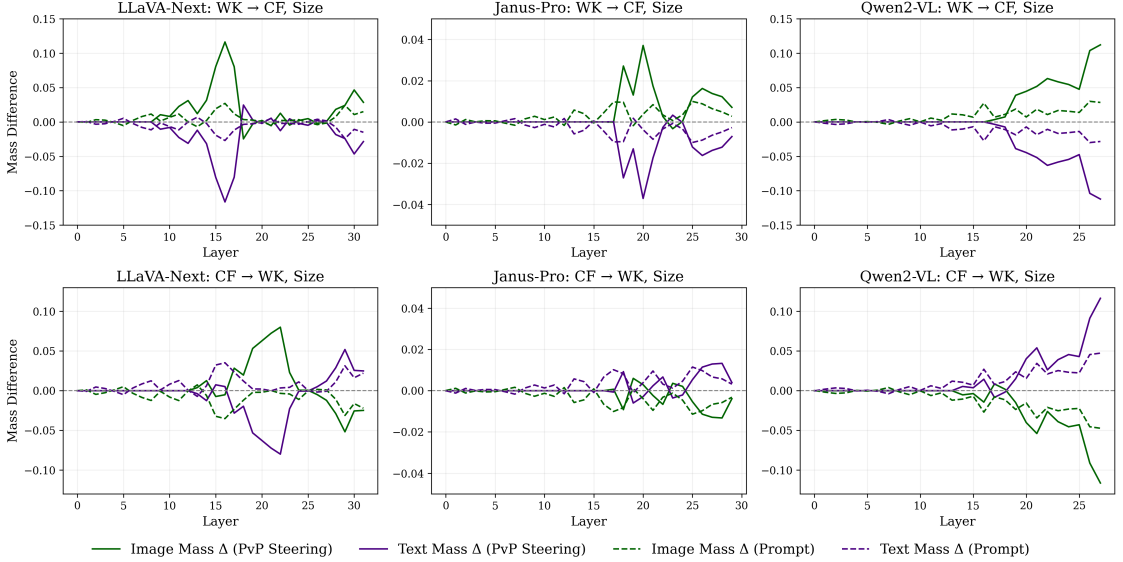


Figure A17: Attention mass difference across layers for all models on the size task. Solid lines show changes from PvP steering vectors; dashed lines show prompt-only changes. Each subplot shows one model and intervention direction (WK $\rightarrow$ CF or CF $\rightarrow$ WK).

(WK $\rightarrow$ CF) is generally more effective than restoring attention to prior-based information (CF $\rightarrow$ WK). This aligns with accuracy results in Section 3.5.3, where interventions that shift models away from priors are more

LLaVA-Next		Intervention	Prompt
Color	WK $\rightarrow$ CF	40.0%	13.1%
Color	CF $\rightarrow$ WK	-15.6%	-13.1%
Size	WK $\rightarrow$ CF	10.9%	3.2%
Size	CF $\rightarrow$ WK	-7.6%	-3.2%
Qwen2-VL			
Color	WK $\rightarrow$ CF	21.8%	12.8%
Color	CF $\rightarrow$ WK	-25.2%	-12.4%
Size	WK $\rightarrow$ CF	14.1%	4.7%
Size	CF $\rightarrow$ WK	-10.8%	-4.7%
Janus-Pro			
Color	WK $\rightarrow$ CF	19.5%	11.7%
Color	CF $\rightarrow$ WK	-11.0%	-10.2%
Size	WK $\rightarrow$ CF	2.4%	1.1%
Size	CF $\rightarrow$ WK	-1.1%	-1.1%

Table A6: Max change in image attention mass (Δ) under intervention and prompt changes for each model and task.

successful than those that attempt to recover them.

These findings provide further evidence that PvP steering offers fine-grained control over internal attention dynamics in MLLMs, outperforming prompt-based techniques in both strength and specificity.

A.14 PvP Early Decoding Details

Figure A18 shows early decoding traces for Qwen2.5-VL and Janus-Pro, extending our main-layer analysis from Figure 3.3. Consistent with the behavior observed in LLaVA-Next, both models initially assign high probability to the world knowledge answer when prompted with a “most” question and shown a counterfactual image. However, as the forward pass progresses, the probability of the counterfactual answer rises and ultimately dominates by the final layer.

For both models, this flipping behavior illustrates the delayed integration of visual information, often leading the model to override its prior with perceptual evidence late in the forward pass. When prompted with “this” questions, both models quickly favor the counterfactual answer and rarely flip to world knowledge. These early decoding results across all three models reinforce our central finding: MLLMs are highly sensitive to visual input and tend to prioritize perception over memorized priors when the two conflict, particularly in later layers of processing.

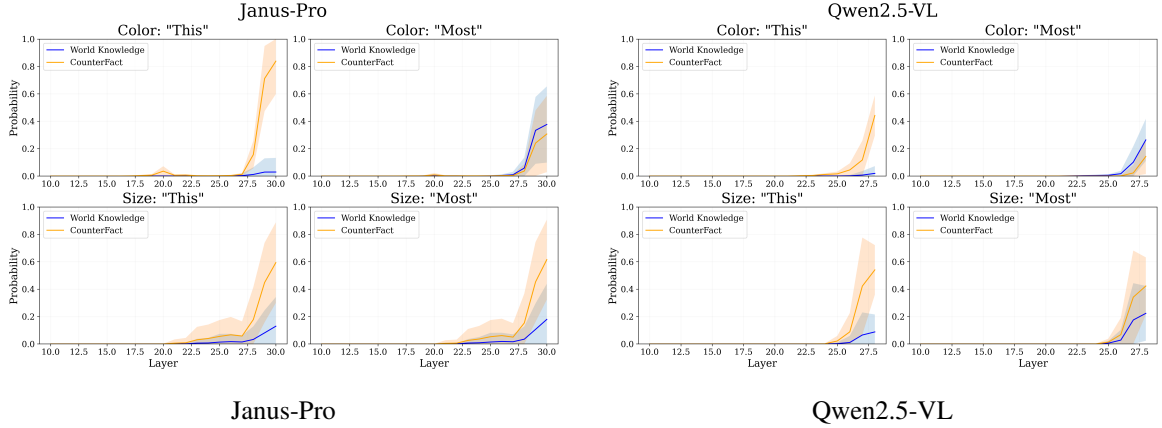


Figure A18: Early decoding results on Qwen2.5-VL and Janus-Pro show a conflict between answering “world knowledge” using in-weight memorization or answering “counterfact” using visual perception.

Model Name	Underlying LM	Vision Encoder	Size	HuggingFace Path
LLaVA-OneVision	Qwen2	SigLIP	7B	llava-hf/llava-onevision-qwen2-7b-ov-hf
Qwen-VL	Qwen2	DFN-ViT w/ RoPE-2D	7B	Qwen/Qwen2-VL-7B-Instruct
Janus Pro	DeepSeek-LLM	SigLIP-Large-Patch16-384	7B	deepseek-ai/Janus-Pro-7B

Table A7: Details of evaluated models, including underlying language models, vision encoders, sizes, and Hugging Face model paths.

A.15 PIH Experimental Details

A.15.1 Accuracy on CountBench

We first evaluate the baseline accuracy of LLaVA-OneVision, Qwen-VL and Janus-Pro on CountBench [106]. Table A8 breaks performance down by the number of objects. All models perform reliably at very small counts (2–4 objects), but accuracy declines as counts increase. Notably, there are sharp dips at **5 objects** (e.g., LLaVA-1.6: 87.27% to 65.91%, QwenVL: 90.91% to 75.00%) and again at **7 objects** (Janus-Pro: 87.72% to 55.56%, LLaVA-OneVision: 89.47% to 48.15%). These discontinuities suggest the presence of training biases that systematically disadvantage certain numbers, a phenomenon supported by prior work linking VLM counting performance to biases in object frequency distributions seen during pretraining [126, 155].

Num	LLaVA-OV	Janus-Pro	QwenVL
2	90.57	84.91	90.57
3	91.23	87.72	85.96
4	87.27	85.45	90.91
5	65.91	70.45	75.00
6	89.47	87.72	85.96
7	48.15	55.56	51.85
8	71.93	80.70	75.44
9	68.33	86.67	71.67

Table A8: Performance across different numbers of objects on CountBench.

A.16 PIH with Large Offsets

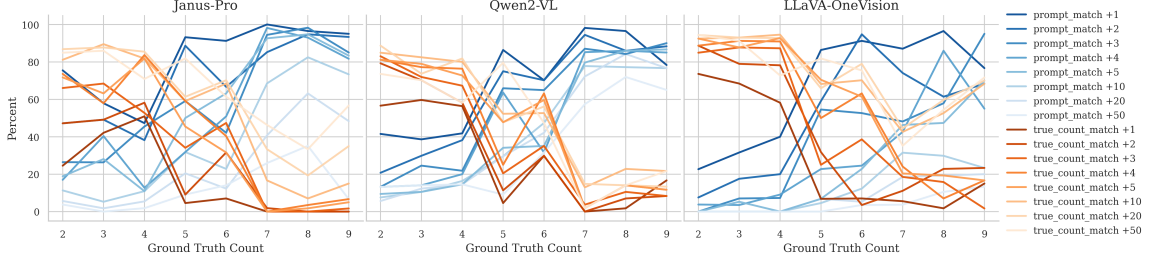


Figure A19: **PIH rates across different number of ground-truth objects.** Blue lines show the percentage of outputs matching the prompted number for different offsets k , while orange lines show matches to the true count. Bias grows with object count, with four objects marking the approximate threshold at which models stop correcting the prompt. Prompt match and true count match do not sum to 100% because some model responses do not contain a number.

Figure A19 extends the analysis in Figure 4.2 to large discrepancy offsets ($k \in \{10, 20, 50\}$), where the prompted count substantially exceeds the number of objects present in the image. Across all three models, we observe that once the number of ground-truth objects exceeds approximately four, prompt-induced hallucinations persist even under extreme misalignment: models frequently produce responses matching the prompted count, while corrections to the true count become rare.

However, LLaVA-OneVision exhibits a notable deviation from this pattern. While Qwen-VL and Janus-Pro show near complete convergence toward prompt matching at higher object counts, regardless of the offset magnitude, LLaVA-OneVision maintains a wider separation between prompt-aligned and ground-truth-aligned responses. In particular, for offsets of +10, +20, and +50, LLaVA-OneVision continues to produce a non-trivial fraction of corrections to the true count, even when the discrepancy is large.

This behavior indicates that LLaVA-OneVision is more resilient to extreme prompt–image mismatches, retaining some capacity to reject implausible prompt specifications at higher object counts. The persistence of ground-truth corrections under large offsets is consistent with our broader findings that LLaVA-OneVision exhibits stronger visual grounding and reduced reliance on prompt-based copying compared to Qwen-VL and Janus-Pro. Nevertheless, even for LLaVA-OneVision, prompt-induced hallucinations dominate once object counts are sufficiently high, underscoring the strength of prompt-driven biases in this regime.

A.17 Base PIH Probabilities for All Models

Figure A20 shows the probabilities for the true count N and neighboring counts $N + k$ under the baseline prompt (P_B) for all models.

Table A9: Top-10 attention heads per model ranked by individual knockout success. Success denotes the percentage of samples switching from PIH to the correct image-grounded count when the head is ablated.

LLaVA-OneVision			Qwen-VL			Janus-Pro		
Rank	Head	Succ.	Rank	Head	Succ.	Rank	Head	Succ.
1	L0H3	48.99	1	L0H3	38.90	1	L0H20	31.18
2	L0H6	37.92	2	L0H6	26.89	2	L1H7	27.96
3	L0H26	30.20	3	L11H15	16.71	3	L14H27	17.74
4	L11H15	27.85	4	L0H11	8.09	4	L12H9	17.47
5	L0H23	27.52	5	L15H8	7.57	5	L11H18	15.59
6	L0H24	21.48	6	L14H13	7.05	6	L13H2	13.98
7	L0H15	20.81	7	L0H10	6.79	7	L0H24	13.44
8	L14H9	17.79	8	L7H3	6.53	8	L14H28	13.17
9	L0H11	16.78	9	L0H24	6.01	9	L8H11	12.90
10	L17H22	15.10	10	L8H13	5.48	10	L8H3	12.63

A.18 PIH Attention Head Knockout Details

To analyze the causal role of individual attention heads in PIH, we perform attention head knockouts using mean ablation [156, 100]. Let $H^{(l,h)} \in \mathbb{R}^{T \times d}$ denote the output of head h at layer l across T tokens. We compute the head-wise mean

$$\mu^{(l,h)} = \frac{1}{T} \sum_{t=1}^T H_t^{(l,h)},$$

and replace the head output at every token position with this mean, i.e.,

$$\tilde{H}_t^{(l,h)} = \mu^{(l,h)} \quad \forall t \in \{1, \dots, T\}.$$

This intervention removes token-specific information carried by the head while preserving its overall activation magnitude.

We apply this procedure head-by-head and quantify knockout success as the fraction of samples that switch from the prompted (hallucinated) count to the true image-grounded count under the misaligned prompt. We rank heads by their corrective effect and select the top- m heads per model. We then perform a second-stage grouped knockout in which all top- m heads are ablated simultaneously for $m \in \{1, 3, 5, 10\}$, selecting m as a hyperparameter per-model.

A.19 Full List of PIH-heads

Table A9 reports the top-ranked attention heads for each model, ranked by their individual knockout success. Knockout success is defined as the fraction of misaligned-prompt samples that switch from matching the prompted (hallucinated) count to matching the true image-grounded count when a single attention head is ablated via mean ablation. Higher values therefore indicate a stronger causal role in propagating prompt-induced numerical errors.

For both Janus-Pro and LLaVA-OneVision, performance improves monotonically with larger head sets, with the best results achieved when ablating the top 10 PIH heads. In contrast, Qwen-VL exhibits its highest

Model	Best-1	Best-3	Best-5	Best-10
Janus-Pro	31.18	65.06	73.19	77.91
LLaVA-OV	48.99	72.35	73.68	83.08
Qwen-VL	38.90	67.14	66.86	3.43

Table A10: Global mean success rates (%) under different PIH head selection strategies. Best- m denotes ablating the top- m PIH heads ranked by individual knockout success.

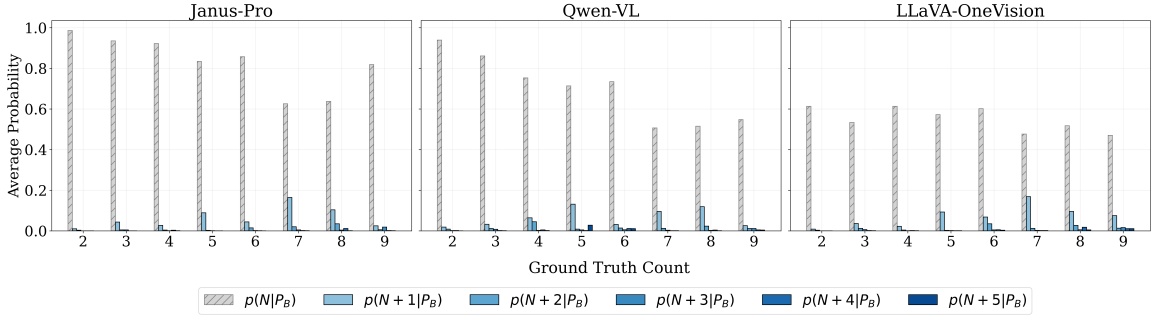


Figure A20: **Base probabilities on the object counting task.** We plot the probability assigned to the true count N and neighboring counts $N + k$ ($k \in \{1, \dots, 5\}$) given the prompt “How many [objects] are in the image?”. As object counts increase, confidence in N decreases and probability mass spreads to neighboring counts, matching the increased susceptibility to PIH.

success rate when ablating only the top 3 heads, with performance degrading substantially when larger groups are removed. This divergence suggests that while PIH in Janus-Pro and LLaVA-OneVision is distributed across a broader set of attention heads, Qwen-VL relies on a smaller subset of highly specialized heads, and removing additional heads begins to interfere with general generation behavior. Accordingly, we select the top-10 heads as the final PIH head sets for Janus-Pro and LLaVA-OneVision, and the top-3 heads for Qwen-VL in all subsequent experiments.

A.19.1 Attention Patterns of PIH-heads

To gain qualitative insight into the behavior of PIH attention heads across models, we visualize the attention patterns of the highest-ranked PIH head in each model, Qwen-VL (L0H3), LLaVA-OneVision (L0H3), and Janus-Pro (L0H20), over the full input and generation sequences, averaged across all samples. Representative examples are shown in Figure A21.

In all visualizations, query tokens are shown on the y-axis and key tokens on the x-axis, such that each row of attention scores sums to 1. Since images are represented as long sequences of image pad tokens (often several hundred), we aggregate the majority of these tokens into a single row and column, denoted as *IMAGE_PADS*, by summing their attention scores. We retain the first three and last three image pads as individual entries to preserve boundary structure and improve interpretability.

The PIH head in Qwen-VL (L0H3) exhibits a mixed attention pattern. A substantial portion of attention lies along the diagonal over the text sequence, indicating behavior similar to local self-attention. At the same

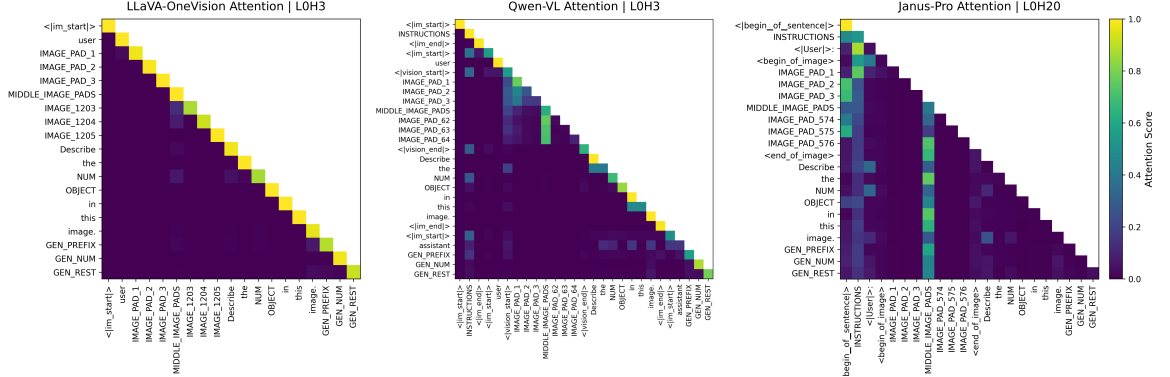


Figure A21: Attention patterns for the top-ranked PIH attention head in three multimodal models, averaged over all samples and computed over the full input–generation sequence. **Left:** LLaVA-OneVision (L0H3), **Center:** Qwen-VL (L0H3), **Right:** Janus-Pro (L0H20). Query tokens are shown on the y-axis and key tokens on the x-axis, with each row summing to 1. Image pad tokens are largely aggregated into a single row and column (*IMAGE_PADS*), with boundary tokens retained.

time, we observe consistent and structured attention within the image-token region, including attention to the aggregated *IMAGE_PADS* and nearby vision boundary tokens. This combination places Qwen-VL L0H3 between purely textual self-attention and fully image-focused attention.

In contrast, the PIH head identified in Janus-Pro (L0H20) shows attention patterns that are predominantly concentrated on image tokens. The diagonal structure characteristic of local textual self-attention is largely absent, and attention is instead focused almost entirely within the image-token region across many query positions. This behavior suggests that the PIH head in Janus-Pro functions primarily as an image-focused attention head.

The PIH head in LLaVA-OneVision (L0H3) behaves almost entirely as a local self-attention head over text. Attention is concentrated along the diagonal, with little to no direct attention to image tokens in the unablated model. Ablation of the PIH head in LLaVA-OneVision leads to the strongest performance improvements among the three models. One possible explanation is that the targeted PIH head in LLaVA-OneVision is not primarily focused on image tokens, but instead behaves as a largely textual self-attention head. As a result, ablating this head may predominantly disrupt copying or local textual interactions, without directly impairing image processing. More generally, ablation may induce second-order effects across all models, whereby other heads adjust their attention patterns and potentially increase attention to image tokens. In the case of LLaVA-OneVision, we hypothesize that the combination of not removing an image-focused head and benefiting from such second-order redistribution effects contributes to its robustness and improved performance under PIH ablation.

For all three models, we do not observe a clear, isolated peak of attention on the generated number token, so while some of the heatmap are suggestive of modality-specific routing, none are conclusive about how the numeric value is produced. This inconclusiveness is reasonable, as interpreting the internal function of attention heads from visualization alone remains a well-known challenge despite their demonstrated causal effects [59, 133, 11, 44].

Before PIH Knockout	After PIH Knockout
Qwen-VL [leftmargin=*, itemsep=0.6em, topsep=0pt]The image shows a set of 52 glasses, which appear to be stemless wine glasses. They are filled with red wine, and the glasses are arranged in a grid pattern, with 13 rows and 4 columns. The image shows three pairs of leggings. The first pair is light pink, the second pair is white with a small pattern, and the third pair is also white but without any pattern. The image shows a total of four wineglasses. Two of them are filled with red wine and are placed side by side on a wooden table. The other two wineglasses are empty and are also placed on the table.	[leftmargin=*, itemsep=0.6em, topsep=0pt]The image shows two wine glasses. The glasses are stemless and have a modern, minimalist design. They are filled with red wine, and the glasses appear to be made of clear glass. The image shows two pairs of leggings. The pair on the left is light pink, and the pair on the right is white with a subtle pattern. The image shows two wine glasses filled with red wine. The glasses are placed on a wooden table, and there is a bottle of red wine and a basket of grapes in the background.
Janus-Pro [leftmargin=*, itemsep=0.6em, topsep=0pt]The image shows four kids standing together outdoors. They are all wearing casual clothing, with one of them in a plaid shirt over a yellow t-shirt. The other three are in solid-colored shirts. The image shows four croissants, which are golden brown in color and have a flaky, layered texture. They are arranged on a wooden cutting board, with two croissants placed side by side and the other two slightly overlapping each other. The image shows 9 planes flying in formation. They are all blue with white markings and have the word "Canada" written on them. The planes are arranged in a tight, symmetrical pattern, with four planes on the top row and five on the bottom.	[leftmargin=*, itemsep=0.6em, topsep=0pt]The image shows two children standing close together outdoors. The child on the left has curly hair and is wearing a bright orange shirt with a red and blue checkered shirt over it. The child on the right has short hair and is wearing a dark shirt. The image shows two croissants placed on a wooden cutting board. They are golden brown, flaky, and have a slightly twisted shape, typical of traditional croissants. The image shows four blue aerobatic planes flying in formation. The planes are marked with the letters "AC" and have a distinctive design with white accents.
LLaVA-OneVision [leftmargin=*, itemsep=0.6em, topsep=0pt]The image displays five different illustrations of classic convertible cars arranged in a two-by-three grid. Each car is depicted in a different color and style, showcasing a variety of vintage designs. The image shows a set of eight forks arranged in two rows against a blue background. Each fork has a polished, reflective surface, indicating they are made of a shiny metal, likely silver or a silver-plated material. The image displays a collection of 17 plastic water bottles arranged in a row against a white background. Each bottle appears to be filled with water, and they vary in shape and size.	[leftmargin=*, itemsep=0.6em, topsep=0pt]The image presents a collection of four vintage cars, each distinct in color and model, arranged in a two-by-two grid. The image shows a set of six ornate silver forks arranged in a row against a blue background. Each fork has a polished finish and features intricate designs on the handles. The image displays a collection of seven plastic water bottles arranged in a row against a white background, with cap colors pink, white, blue, white, blue, white, and white.

Table A11: Qualitative examples of generations before and after PIH attention head knockout. Before intervention, models not only copy the incorrect count implied by the misaligned prompt, but also hallucinate non-existing details. After ablating PIH heads, generations more consistently reflect the true number of objects present in the image across all three models.

A.20 Output Examples After PIH Knockouts

Table [A11](#) shows examples of generated texts after PIH-head ablations.

A.21 PIH Color Task

To investigate whether PIH heads generalize beyond the counting task, we extend our framework to a color prediction task using the Visual CounterFact dataset [\[40\]](#). This setting replaces numerical offsets with color offsets while preserving the overall experimental structure used in counting. We use "What color is the [object] in the image?" as the baseline prompt, and "Describe the $C + k$ [object]" as the misaligned prompt, where C denotes the ground-truth color of the object in the image and $k \in \{1, 2, 3\}$ denotes the perceptual "difference" from the C on the color wheel. In particular, $k = 1$ indicates a highly similar color and $k = 3$ indicates the

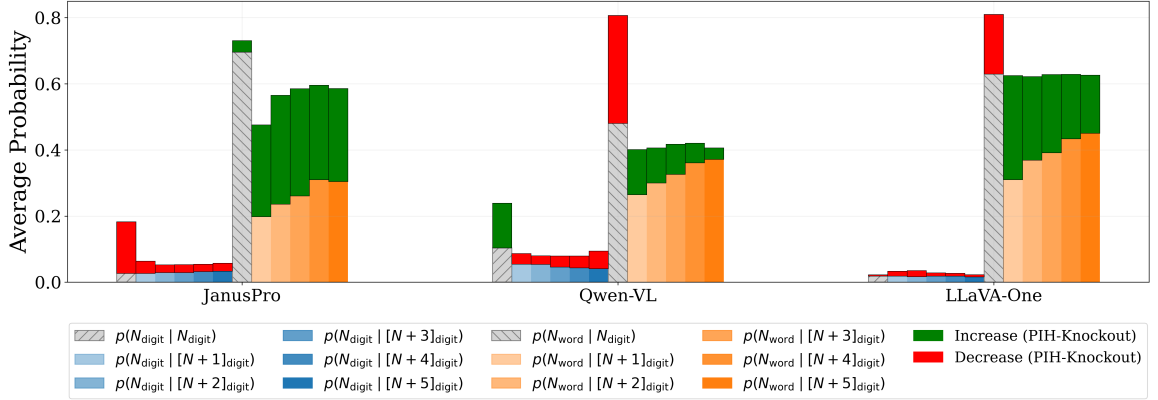


Figure A22: Impact of knockout on the probability of the ground-truth answer in both digit (blue) and word form (orange) given the prompt “Describe the $[N + k]_{\text{digit}}$ [objects]” for count. Note that when $k = 0$, there is no modality conflict in the prompt.

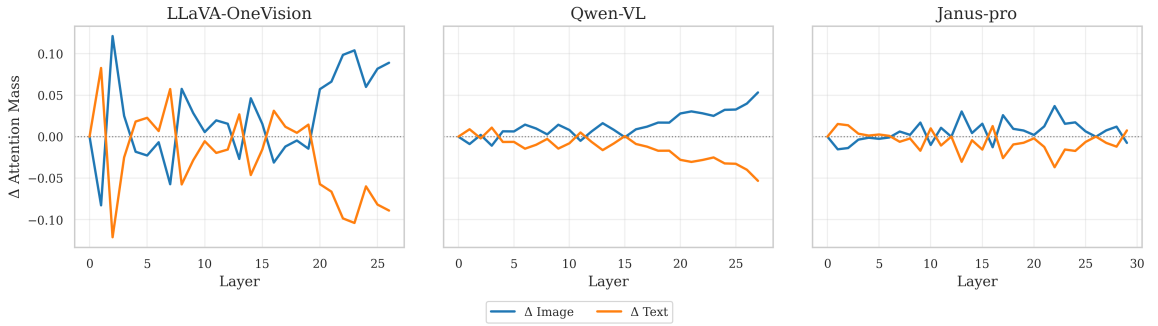


Figure A23: **Layer-wise change in attention mass after PIH head ablation.** For each model, we plot the change in attention mass to image tokens (blue) and text tokens (orange) as a function of transformer layer. Positive values indicate increased attention allocation relative to the baseline. Across all models, PIH ablation consistently shifts attention from text to image tokens, with LLaVA-OneVision exhibiting the strongest and earliest effect.

contrasting color which is the further distance on the color wheel from C . For example, given $C = \text{red}$, we have orange, yellow, and green for $k = 1, 2, 3$, respectively.

A.22 Layer-wise Attention Mass Shifts After PIH Ablation

While Figure 4.4 focuses on the single layer exhibiting the largest intervention-induced shift in attention mass, we additionally examine how PIH head ablation affects attention allocation *across all transformer layers*. Specifically, for each layer we compute the change in attention mass to image tokens relative to the baseline, with the complementary change applied to text tokens (since attention mass sums to one).

Figure A23 shows the layer-wise change in image and text attention mass for LLaVA-OneVision, Qwen-VL, and Janus-pro. Across all models, PIH ablation consistently increases attention to image tokens while reducing attention to text tokens, indicating a systematic rebalancing of multimodal attention rather than a

highly localized effect.

Notably, the magnitude and depth of this shift varies across architectures. LLaVA-OneVision exhibits a pronounced increase in image attention in early layers, whereas Qwen-VL and Janus-pro show smaller but more distributed changes concentrated in later layers. These trends suggest that PIH head ablation alters the global flow of multimodal information throughout the network.