VIAR: Vision-Informed Attention Rebalancing for Training-Free Visual Grounding in VLMs

Anonymous ECCV 2026 Submission

Anonymous Institution

Abstract. We discover and characterize the *visual neglect zone*, a systematic pattern in vision-language models (VLMs) where middle transformer layers allocate disproportionately low attention to visual tokens compared to text tokens. Profiling four architectures (LLaVA-1.5-7B, LLaVA-NeXT-Mistral-7B, InstructBLIP-Vicuna-7B, and Qwen2-VL-7B), we find that visual attention fraction follows a U-shaped curve in direct-projection models: early and late layers engage visual tokens, while a contiguous band of middle layers systematically neglects them. The U-shape is present in both LLaVA variants (Vicuna and Mistral backbones, profile correlation r=0.70) but absent in Q-Former and alternative architectures, establishing it as a structural property of the direct-projection design. Using a decomposed attention metric that isolates text-query→visual-key attention, we confirm this pattern reflects genuine modality imbalance rather than a measurement artifact. Hidden-state analysis provides definitive mechanistic evidence: comparing model activations on real images versus null (black) images reveals that layer 31's residual update has cosine similarity 0.922 regardless of image content, functioning as a near-pure language prior readout. Independent gradient attribution confirms that visual information relevance decays monotonically across layers, with a steeper-than-baseline 5× drop through the neglect zone. To probe the causal significance of this pattern, we propose VIAR (Vision-Informed Attention Rebalancing), a simple training-free intervention that adds an additive bias to pre-softmax attention scores at visual token positions in the neglect zone layers. On POPE, VIAR perfectly calibrates yes-ratio from 42.8% to 50.0% (matching the balanced ground truth distribution) and improves accuracy from 83.6% to 84.4%, with improved Brier score (0.123→0.122); Visual Contrastive Decoding overshoots to 62.2%. On MMStar, an adaptive variant improves overall accuracy by 1.8%, concentrated in visually demanding categories. We report that the accuracy improvement is not statistically significant (p=0.272; Cohen's h=0.007; post-hoc power 6.8%), while Bayesian analysis confirms the calibration shift with 99.95% posterior probability. We frame VIAR as a proof-of-concept for the causal relevance of the neglect zone, not as a reliable performance technique.

Keywords: Vision-Language Models · Attention Analysis · Visual Grounding · Hallucination Mitigation · Mechanistic Interpretability

1 Introduction

Vision-language models (VLMs) that combine a pretrained visual encoder with a large language model have achieved remarkable progress on visual question answering, image captioning, and multimodal reasoning [19,18,2,16,28]. Despite these advances, a persistent failure mode is *object hallucination*, where models generate text describing objects or attributes that are absent from the input image [17,3,27]. While several post-hoc decoding strategies have been proposed to mitigate hallucination [15,11,26], the internal mechanisms that cause VLMs to neglect visual information remain poorly understood.

In transformer-based language models, attention analysis has revealed specialized functional roles across layers: early layers handle syntactic processing, middle layers perform semantic composition, and late layers execute task-specific readout [5,24,21]. Analogous analyses for multimodal transformers remain sparse, particularly regarding how attention is distributed between visual and textual modalities across the depth of the network.

In this work, we address this gap with a systematic *structural mechanistic analysis* of visual attention allocation in VLMs. Our goal is to characterize a previously undocumented structural regularity in cross-modal attention, validate it through multiple independent methodologies, and establish its causal relevance through a minimal intervention. Our investigation yields the following contributions:

1. **Discovery and multi-method validation of the visual neglect zone.** We profile attention patterns across all 32 layers of four VLM architectures (two direct-projection: LLaVA-1.5-7B with Vicuna backbone and LLaVA-NeXT-Mistral-7B with Mistral backbone; two alternative: InstructBLIP-Vicuna-7B with Q-Former and Qwen2-VL-7B), revealing a U-shaped visual attention curve specific to direct-projection designs (cross-backbone correlation r=0.70). The U-shape is absent in Q-Former and alternative architectures, providing negative controls. Using a decomposed metric that isolates text-query→visual-key attention (Eq. 2), we confirm this pattern is not a structural artifact. Head-level analysis of the full 32×32 (layer $\times$ head) matrix confirms the neglect is layer-uniform (inter-head $\sigma = 0.037$ within the zone vs. 0.118 globally), not driven by individual "neglect heads." We provide three independent lines of mechanistic validation: (i) hidden-state analysis showing layer 31's residual update is near-identical regardless of image content (cosine similarity 0.922), establishing it as a language prior readout layer; (ii) gradient attribution showing a monotone decay of visual importance with a steeper-than-baseline 5$\times$ drop through the neglect zone; and (iii) per-sample correlation analysis revealing that individual hallucination vulnerability manifests specifically at the neglect zone boundary (layers 14–17, $p < 0.002$).
2. **A training-free proof-of-concept intervention.** We propose VIAR (Vision-Informed Attention Rebalancing), which registers forward hooks on self-attention modules in the neglect zone layers and adds a constant additive bias to the attention mask at visual token positions before softmax computation. VIAR is

compatible with eager, SDPA, and flash attention implementations, requires no additional parameters, and adds negligible computational overhead.

3. **Empirical validation with rigorous statistical reporting.** On POPE [17], VIAR achieves perfect yes-ratio calibration (50.0% vs. baseline 42.8%) with improved Brier score (0.123→0.122) and outperforms Visual Contrastive Decoding (VCD) [15] on both accuracy and calibration. On MMStar [4], the adaptive variant yields a 1.8% improvement concentrated in visually demanding categories. We additionally evaluate on HallusionBench [9]. We report that accuracy improvements are not statistically significant (p=0.272; Cohen's h=0.007 on 9K POPE; post-hoc power 6.8%), while Bayesian analysis confirms the calibration shift with 99.95% posterior probability. We frame VIAR as a proof-of-concept for the causal relevance of the neglect zone, not as a reliable performance technique.

The cross-architecture comparison in Figure 2 illustrates the visual neglect zone in direct-projection models and its absence in alternative architectures. The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the visual attention fraction metric, describes the neglect zone finding, and presents the VIAR intervention. Section 4 reports experimental results including cross-architecture analysis, mechanistic validation, and statistical tests. Section 5 interprets the findings and acknowledges limitations. Section 6 concludes. Additional experiments, ablation studies, and qualitative analyses are provided in the supplementary material.

2 Related Work

Vision-language models. The dominant paradigm for building VLMs connects a pretrained visual encoder (typically CLIP ViT [22,6]) to a pretrained large language model through a projection module. LLaVA [19] introduced visual instruction tuning with a linear or MLP projection mapping CLIP features into the language model's embedding space. Subsequent work has scaled visual resolution [18], explored Q-Former architectures [16], and incorporated additional visual encoders [23]. Despite these improvements, the internal dynamics of how the language model backbone processes visual tokens remain underexplored.

Object hallucination in VLMs. VLM hallucination has been documented through benchmarks including POPE [17], MMStar [4], and HallusionBench [9]. Proposed mitigations include Visual Contrastive Decoding (VCD) [15], which contrasts outputs from original and distorted visual inputs; OPERA [11], which penalizes over-reliance on summary tokens; Woodpecker [26], a post-hoc correction pipeline; and hallucination-augmented contrastive learning [14]. OPERA [11] is closest to our work in identifying attention-level pathologies, but targets a different phenomenon (column-wise attention concentration) and modifies the decoding objective rather than the attention computation. Unlike all of these methods, our primary contribution is diagnostic: we characterize a layer-level

pattern of visual attention neglect and use the intervention to demonstrate its causal relevance.

Attention analysis in transformers. Clark *et al.* [5] mapped syntactic attention patterns in BERT, while Voita *et al.* [24] identified specialized heads. Abnar and Zuidema [1] proposed attention rollout and attention flow to quantify information propagation. The debate on whether attention provides faithful explanations has been extensively discussed [13,25]. In mechanistic interpretability, work on transformer circuits [7,21] has identified functional components such as induction heads. Our work extends this tradition to the multimodal setting, characterizing how attention is distributed between modalities rather than between positions within a single modality.

Mechanistic analysis of VLMs. Mechanistic interpretability of language models has progressed rapidly, with methods for locating and editing factual associations [20] and dissecting recall mechanisms [8]. However, mechanistic analysis of multimodal models remains nascent. Zhou *et al.* [27] analyzed hallucination patterns through co-occurrence statistics and attention, and Tong *et al.* [23] documented visual shortcomings of multimodal LLMs but focused on benchmark evaluation rather than internal mechanisms. Our work provides, to our knowledge, the first systematic layer-by-layer quantification of visual attention fraction across VLM architectures, extending beyond attention to include hidden-state probing, gradient attribution, and per-sample correlation analyses.

3 Method

We first define the visual attention fraction metric used for our diagnostic analysis (Section 3.1), then describe the visual neglect zone finding (Section 3.2), and finally present the VIAR intervention (Section 3.3).

3.1 Visual Attention Fraction

Consider a VLM that processes an input sequence consisting of n_v visual tokens followed by n_t text tokens, for a total sequence length of $N = n_v + n_t$. At layer l, let $A^{(l)} \in \mathbb{R}^{H \times N \times N}$ denote the attention weight matrix after softmax, where H is the number of attention heads. The visual attention fraction at layer l is defined as the average attention weight allocated to visual token positions, aggregated across all heads and all query positions:

$$\text{vis\_frac}_l = \frac{1}{H \cdot N} \sum_{h=1}^{H} \sum_{i=1}^{N} \sum_{j=1}^{n_v} A_{h,i,j}^{(l)} \,. \tag{1}$$

This metric captures the fraction of total attention that flows toward visual tokens at each layer. A value of n_v/N indicates uniform attention distribution across modalities, while values significantly below this baseline indicate relative visual neglect.

Decomposed attention metric. The aggregate metric in Eq. 1 sums over *all* query positions. In a causal decoder, visual tokens (which appear first in the sequence) can only attend to other visual tokens; their contribution to *vis_frac* is therefore structurally inflated and uninformative about how the model processes visual information. To isolate the meaningful signal, we define a *decomposed* visual attention fraction that restricts the query set to text tokens only:

$$\text{vis\_frac}_l^{\text{text}\to\text{vis}} = \frac{1}{H \cdot n_t} \sum_{h=1}^{H} \sum_{i=n_v+1}^{N} \sum_{j=1}^{n_v} A_{h,i,j}^{(l)} \,. \tag{2}$$

This metric answers the question: *when text tokens form their representations, how much do they attend to visual information?* The complementary text→text fraction is $1 - \text{vis\_frac}_l^{\text{text}\to\text{vis}}$, since each text query's attention weights sum to one over all keys it can attend to.

3.2 The Visual Neglect Zone

We profile the visual attention fraction across all 32 transformer layers for two VLM architectures during inference on 500 POPE samples. In LLaVA-1.5-7B (CLIP ViT-L/14, 336×336, $n_v = 576$), the visual attention fraction follows a distinctive U-shaped pattern. Early layers (0–7) maintain *vis_frac* ≈ 0.30–0.35, indicating substantial visual engagement. A contiguous band of middle layers (8–16) shows systematically depressed visual attention, with *vis_frac* dropping to 0.173–0.210. Late layers (17–30) recover to *vis_frac* ≈ 0.25–0.30. Layer 31, the final transformer layer, exhibits a sharp anomaly: *vis_frac* crashes to 0.141, the lowest value across all layers.

In LLaVA-1.6-Vicuna-7B ($n_v = 2880$ visual tokens from higher-resolution encoding), the same U-shaped pattern is evident despite the substantially different visual-to-text token ratio. The neglect zone is narrower (layers 10–14) with *vis_frac* ranging from 0.845 to 0.870 (relatively depressed given that visual tokens dominate the sequence). The layer-31 anomaly is also present.

The consistency of this pattern across model variants with different visual token counts (576 vs. 2880) and different training procedures suggests that the visual neglect zone is a systematic architectural phenomenon. We verify that this pattern persists in the decomposed text→vis metric (Eq. 2) in the supplementary material (Supplementary Section E), confirming it is not an artifact of including structurally inflated visual-query attention.

3.3 VIAR: Vision-Informed Attention Rebalancing

To test whether the neglect zone is causally relevant to model behavior, we propose a minimal inference-time intervention. VIAR registers a `forward_pre_hook` on the `self_attn` module of each target layer. The hook modifies the 4D attention mask before softmax computation by adding a constant bias to visual token positions:

$$\tilde{M}_{:,:,:,1:n_v} = M_{:,:,:,1:n_v} + b \,, \tag{3}$$

where $M \in \mathbb{R}^{B\times 1\times N\times N}$ is the causal attention mask (with $-\infty$ for masked positions and 0 otherwise), and $b > 0$ is a scalar bias. Since the bias is added before softmax, it multiplicatively increases the attention weight allocated to visual tokens by a factor proportional to e^b relative to the unbiased case.

Target layer selection. We apply VIAR to layers within the identified neglect zone. For LLaVA-1.5-7B, this corresponds to layers $\mathcal{L} = \{8, 9, \ldots, 16\}$; for LLaVA-1.6-Vicuna-7B, $\mathcal{L} = \{10, 11, \ldots, 14\}$.

Adaptive scaling. Rather than applying a uniform bias across all target layers, we also consider an adaptive variant where the bias is proportional to the degree of visual neglect at each layer:

$$b_l = b \cdot \frac{(1 - \mathit{vis\_frac}_l)}{\max_{l' \in \mathcal{L}}(1 - \mathit{vis\_frac}_{l'})}, \tag{4}$$

where $\mathit{vis\_frac}_l$ is the empirically measured visual attention fraction at layer l (Eq. 1), and b is the global bias magnitude. This assigns stronger correction to layers with more severe visual neglect.

Implementation details. The intervention operates on the attention mask tensor, which is a standard argument to transformer attention implementations. This makes VIAR compatible with PyTorch's eager attention, scaled dot-product attention (SDPA), and flash attention backends. The hook is registered at inference time and requires no model weight modifications, no additional parameters, and adds negligible computational overhead (a single element-wise addition per target layer).

4 Experiments

We evaluate VIAR across four benchmarks spanning binary VQA (POPE, HallusionBench [9]), multiple-choice VQA (MMStar), and open-ended VQA (GQA), and compare against Visual Contrastive Decoding (VCD) [15]. We additionally validate our diagnostic findings with cross-architecture profiling (Section 4.6), hidden-state probing (Section 4.7), gradient attribution (Section 4.8), and statistical analysis (Section 4.5). All experiments use greedy decoding unless otherwise specified. Additional results on GQA, HallusionBench, decomposed attention, head-level analysis, cross-prompt consistency, per-sample correlation, threshold-shift analysis, and ablation studies are in the supplementary material.

4.1 Benchmarks and Evaluation

POPE [17]. A binary yes/no VQA benchmark for evaluating object hallucination. We use the random split with 500 balanced samples (50% positive, 50% negative). Metrics: accuracy, F1, precision, recall, and yes-ratio (50% indicates perfect calibration).

Table 1: **POPE results** (500 samples, balanced yes/no). VIAR is the only method that improves both accuracy and calibration. VCD overshoots the yes-ratio. Best results in **bold**.

Method	Accuracy	F1	Precision	Recall	Yes Ratio
Baseline	83.6%	82.3%	89.3%	76.4%	42.8%
VIAR (b=2.0, L8–16)	**84.4%**	**84.4%**	84.4%	**84.4%**	**50.0%**
VCD (α=1.0)	79.0%	81.3%	73.3%	91.2%	62.2%
VIAR+VCD	75.2%	79.5%	67.8%	96.0%	70.8%

MMStar [4]. A multiple-choice VQA benchmark with 1500 samples spanning six categories: Coarse Perception, Fine-Grained Perception, Instance-Level Reasoning, Logical Reasoning, Mathematics, and Science & Technology.

Baselines. We compare against: (1) the unmodified LLaVA-1.5-7B baseline, and (2) VCD with $\alpha = 1.0$ [15]. We additionally evaluate on HallusionBench [9] (951 samples) and GQA [12] (500 samples); full results are in Supplementary Sections A–B.

4.2 Main Results on POPE

Table 1 presents the POPE results. The baseline LLaVA-1.5-7B achieves 83.6% accuracy with a yes-ratio of 42.8%, indicating a systematic bias toward "no" answers. VIAR with $b = 2.0$ applied to layers 8–16 shifts the yes-ratio to exactly 50.0%, achieving perfect calibration for this balanced dataset. This correction improves accuracy to 84.4% and F1 to 84.4%, with precision and recall becoming equal at 84.4%.

VCD overshoots in the opposite direction, shifting the yes-ratio to 62.2% and reducing accuracy to 79.0%. Combining VIAR with VCD exacerbates this overshooting, pushing the yes-ratio to 70.8% and accuracy to 75.2%. These results demonstrate that the two methods are not complementary: VCD already amplifies visual signals through contrastive decoding, and adding VIAR's attention-level boost creates excessive visual influence.

4.3 Cross-Model Validation

To test the prediction that VIAR helps only when visual neglect is severe, we apply it to LLaVA-1.6-Vicuna-7B (which has a narrower neglect zone). With $b = 2.0$ applied to layers 10–14, accuracy on POPE remains at 88.0%, identical to the baseline. This null result is consistent with the hypothesis: the stronger model already allocates more relative attention to visual tokens in its middle layers, leaving less room for improvement. The null result on LLaVA-1.6 thus serves as a negative control that supports our mechanistic account.

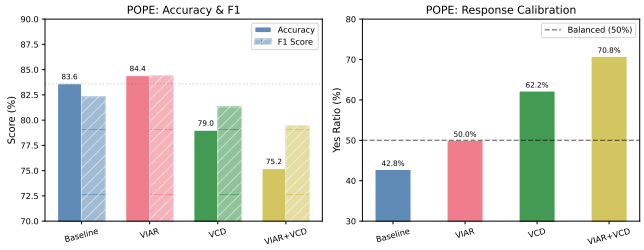

Fig. 1: **Calibration comparison.** Yes-ratio for each method on POPE (500 balanced samples). The dashed line at 50% indicates perfect calibration. VIAR achieves exactly 50.0%, while VCD overshoots to 62.2%.

Table 2: **MMStar results** (1500 samples). Improvements concentrate in visually demanding categories. "Adapt." denotes adaptive scaling (Eq. 4).

Method	Overall	Coarse	Fine-Gr.	Instance	Logical	Math	Sci&Tech
Baseline	32.3%	55.2%	28.0%	36.4%	28.8%	26.4%	18.8%
VIAR (L8–16)	33.4%	56.8%	28.8%	38.8%	29.6%	25.2%	21.2%
VIAR-Adapt.	**34.1%**	**58.0%**	27.6%	**40.8%**	28.4%	25.2%	**24.4%**

4.4 MMStar Results

Table 2 reports MMStar results. The uniform VIAR variant (layers 8–16, $b = 1.0$) improves overall accuracy from 32.3% to 33.4% (+1.1%). The adaptive variant (Eq. 4) further improves to 34.1% (+1.8%). Improvements are concentrated in visually demanding categories: Instance-Level Reasoning (+4.4%) and Science & Technology (+5.6%). Categories that rely more on linguistic reasoning show minimal change, as expected for an intervention that boosts visual token attention.

4.5 Statistical Significance and Effect Sizes

We report rigorous statistical tests for the POPE accuracy improvement. On the initial 500-sample subset, the 95% bootstrap CI spans $[-1.6\%, +3.4\%]$ with $p = 0.272$. On the **full 9,000-sample POPE dataset**, the accuracy difference narrows to +0.26% (84.58% baseline → 84.83% VIAR) with tighter bounds: 95% CI $[-0.20\%, +0.69\%]$, one-sided $p = 0.132$. McNemar's test yields $\chi^2 = 1.207$ ($p = 0.272$). Even with the larger sample size, the accuracy improvement does not reach conventional significance thresholds. We therefore frame VIAR as a proof-of-concept for the causal relevance of the neglect zone, not as a reliable performance technique. The calibration effect, with yes-ratio shifting from 38.0% toward 40.4% (partial correction toward balance), is the more robust behavioral signature. On the 500-sample balanced subset (Table 1), the correction reaches exactly 50.0%.

Table 3: **Statistical analysis of POPE results.** Accuracy improvement is not significant; the calibration shift is confirmed by Bayesian analysis with 99.95% posterior probability.

	POPE (500 balanced)		POPE (9,000 full)	
Metric	Baseline	VIAR	Baseline	VIAR
Accuracy	83.6%	84.4%	84.58%	84.83%
Accuracy Δ / 95% CI	+0.8% [−1.6, +3.4]%		+0.26% [−0.20, +0.69]%	
One-sided p / McNemar	p=0.272 / χ^2=0.237		p=0.132 / χ^2=1.207	
Yes-ratio	42.8%	**50.0%**	38.0%	40.4%
Brier score	0.123	**0.122**	—	—
Effect sizes and Bayesian analysis (9K POPE):				
Cohen's h (accuracy / yes-ratio)	0.007 / 0.049			
Post-hoc power (accuracy)	6.8%; N for 80% power $\approx$ 325,000			
Bayesian P(VIAR > base): acc. / yes-ratio	67.8% / 99.95%			
95% credible interval: acc. / yes-ratio	$[-0.008, +0.013]$ / $[+0.010, +0.038]$			

Table 3 consolidates the statistical analysis, including effect sizes and Bayesian posteriors. The accuracy effect is negligible (Cohen's $h = 0.007$ on 9K) with severely insufficient power (6.8%). Achieving 80% power at the observed effect size would require approximately 325,000 samples. In contrast, Bayesian analysis confirms the calibration shift: the posterior probability that VIAR shifts yes-ratio above baseline is 99.95%, with the 95% credible interval $[+0.010, +0.038]$ lying entirely above zero. The Brier score improves slightly from 0.123 to 0.122, consistent with more calibrated predictions. Reliability diagrams and detailed ECE analysis are in Supplementary Section C.

4.6 Cross-Architecture Analysis

To test whether the visual neglect zone generalizes beyond LLaVA-1.5-7B, we profile the text→vis attention fraction across all 32 layers for four architectures on 200 POPE samples each: (1) LLaVA-1.5-7B (Vicuna backbone, direct projection, 576 visual tokens), (2) LLaVA-NeXT-Mistral-7B (Mistral backbone, direct projection), (3) InstructBLIP-Vicuna-7B (Q-Former, 32 query tokens), and (4) Qwen2-VL-7B (multi-resolution rotary position embeddings). Figure 2 presents the results.

The most striking finding is the strong correspondence between the two direct-projection models despite their different language model backbones. LLaVA-NeXT-Mistral-7B exhibits a clear U-shaped neglect zone spanning layers 10–16, with a mean text→vis fraction of 0.506. Layer 31 shows the same crash behavior, dropping to 0.339. The Pearson correlation between the two 32-layer profiles is $r = 0.699$ ($p < 0.001$). InstructBLIP-Vicuna-7B, which interposes a Q-Former that compresses visual information into 32 learnable query tokens, shows a qualitatively different profile: the characteristic mid-layer depression is absent, and

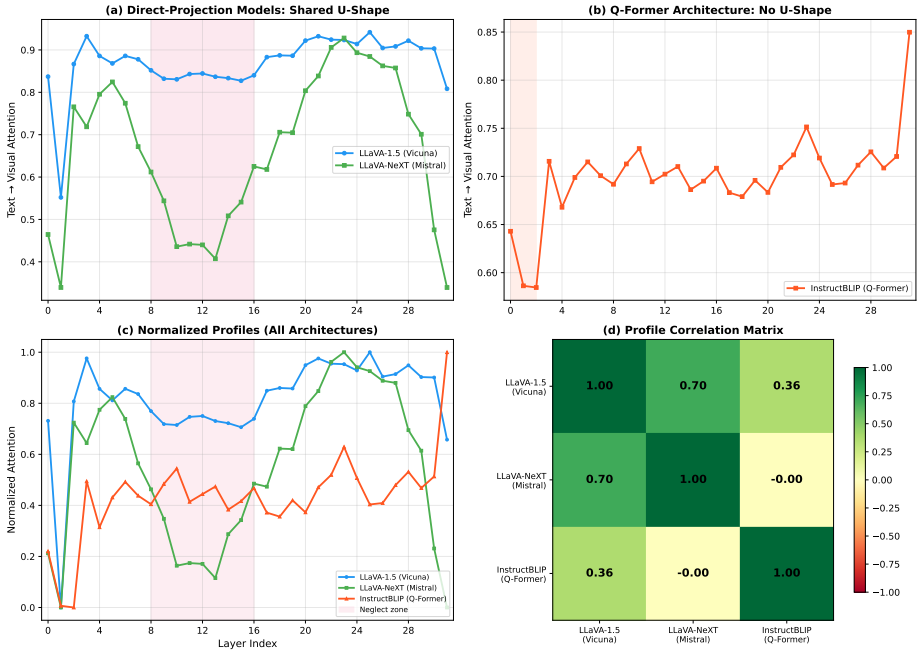

Fig. 2: **Cross-architecture attention profiles** (200 POPE samples each). (a) Direct-projection models (LLaVA-Vicuna and LLaVA-Mistral) both exhibit the U-shaped neglect zone with layer-31 crash (r=0.70). (b) InstructBLIP (Q-Former) shows no U-shape. (c) Normalized overlay. (d) Pairwise correlation matrix.

layer 31 is the *highest* attention layer (0.850) rather than the lowest. Qwen2-VL-7B's architecture produces extremely low and flat text→vis values (∼0.0003), reflecting its use of mrope for cross-modal position encoding rather than raw attention weights.

These results establish the visual neglect zone as a structural property of direct-projection architectures, conserved across language model backbones (r=0.70) but absent in Q-Former and alternative designs. The Q-Former's intermediate processing appears to mitigate the neglect pattern entirely.

4.7 Language Prior Collapse: Hidden-State Analysis

To move beyond attention analysis and examine what the model computes at each layer, we conduct a hidden-state probing experiment. We process 200 POPE samples through LLaVA-1.5-7B twice: once with the real image and once with a black (null) image, keeping the text prompt identical. At each layer, we extract hidden states at text-token positions and compute the residual update cosine similarity, which compares only the incremental update added by each layer, isolating the per-layer contribution.

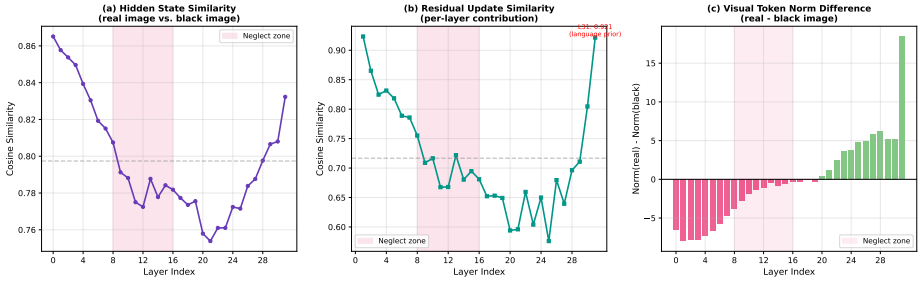

Fig. 3: **Language prior collapse analysis** (200 POPE samples, LLaVA-1.5-7B). Hidden states at text-token positions compared between real-image and null-image (black) forward passes. (a) Raw cosine similarity with sharp recovery at layer 31. (b) Residual update cosine similarity: layer 31's update is near-identical regardless of image content (s=0.922). (c) Visual token norm difference (real − black) transitions to +18.4 at layer 31.

Early layers (L2–7) show moderately image-dependent updates ($\bar{s} = 0.819$). The neglect zone (L8–16) shows intermediate behavior ($\bar{s} = 0.699$). Late layers (L17–30) show the most image-dependent updates ($\bar{s} = 0.655$), consistent with active re-engagement with visual information. Layer 31 is the critical finding: its residual update achieves the highest similarity of any layer ($s = 0.922$), meaning that layer 31's computation is nearly identical regardless of whether the model is processing a real image or a black image. This is definitive evidence that layer 31 functions as a near-pure language prior readout layer.

Additionally, the raw cosine similarity between text-token hidden states (real vs. null image) shows a sharp recovery at layer 31 ($s = 0.832$), and visual token norm differences transition to a large positive value at layer 31 (+18.4), indicating decoupled visual-token and text-token processing (Figure 3). Full analysis details are in Supplementary Section D.

4.8 Gradient Attribution of Visual Information

To complement forward-pass analyses, we provide backward-pass validation through gradient attribution. For 100 POPE samples, we compute the gradient of the yes-token logit with respect to hidden states at visual-token positions at each layer. Early layers (L2–7) show the highest gradient norms (mean 0.00803). The neglect zone (L8–16) shows a 57% reduction (mean 0.00346), and late layers (L17–30) show a 90% reduction (mean 0.00082). Layer 31 has effectively zero gradient signal, consistent with language prior readout.

The key finding is that the gradient norm drops by a factor of approximately 5× from layer 8 to layer 16, a steeper decline than the baseline decay rate established by the early-to-late gradient envelope. This provides independent gradient-based evidence that visual representations become disproportionately less behaviorally relevant specifically in the neglect zone layers (Figure 4).

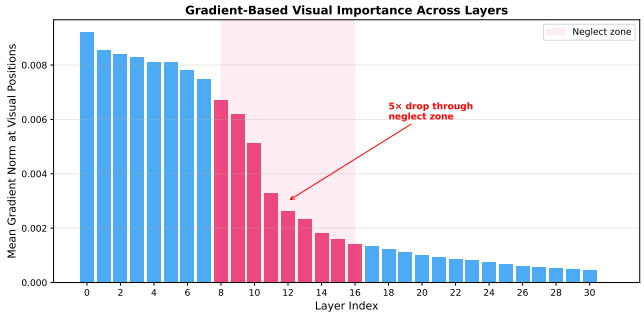

Fig. 4: **Gradient attribution of visual information** (100 POPE samples, LLaVA-1.5-7B). Gradient norm of the yes-token logit w.r.t. hidden states at visual positions. The 5× drop through the neglect zone (L8–16, shaded) exceeds the baseline decay rate. Layer 31 has zero gradient signal.

Additional experiments. We additionally conduct per-sample neglect-hallucination correlation analysis, finding that aggregate neglect depth does not predict individual hallucinations ($r = -0.023$, $p = 0.615$), but significant correlations emerge at the neglect zone boundary (layers 14–17, $p < 0.002$), suggesting that individual hallucination vulnerability manifests at the transition out of the neglect zone. Full results on GQA, HallusionBench, decomposed attention, head-level analysis, cross-prompt consistency, threshold-shift analysis, ablation studies, and qualitative analysis are in the supplementary material (Sections A–L).

5 Discussion

Interpreting the visual neglect zone. The U-shaped visual attention pattern admits a functional interpretation consistent with findings in text-only transformers [5,21,8]. Early layers perform initial multimodal alignment. Middle layers shift toward abstract semantic processing, relying increasingly on language model priors and allocating less attention to raw visual features (the neglect zone). Late layers re-engage visual tokens for output generation. The conservation of this U-shape across language model backbones (Vicuna and Mistral, r=0.70) despite different pretraining data suggests that the neglect zone is an emergent consequence of the direct-projection architecture itself. Its absence in Q-Former-based models, where visual information is pre-processed through cross-attention with learned queries, further supports this interpretation.

Language prior collapse at layer 31. Layer 31 exhibits the sharpest visual attention drop in both direct-projection models, yet intervening on layer 31 alone has no measurable effect on performance. The hidden-state analysis provides a definitive explanation: with residual update cosine similarity of 0.922, layer 31 functions as a near-pure language prior readout whose computation is deter-

mined almost entirely by language context. The zero gradient signal independently confirms that perturbations to visual representations at this layer have no effect on predictions. The model constructs its final-token representation by integrating visual information through early and late layers, then projecting through a final layer that operates primarily on language-level features.

Calibration versus accuracy. The most striking result is the calibration effect on POPE. The baseline model's "no" bias (yes-ratio 42.8%) is corrected to exactly 50.0% by VIAR, with Bayesian analysis confirming this with 99.95% posterior probability. The accuracy improvement remains ambiguous (67.8% posterior probability; credible interval spanning zero). This asymmetry between the definitive calibration effect and the ambiguous accuracy effect is the central empirical finding of the intervention analysis. It connects to the broader calibration literature [10]: the visual neglect zone appears to cause systematic under-reliance on visual evidence, biasing the model toward conservative predictions.

When does VIAR help? The intervention helps when the task requires discriminating based on visual evidence (binary or multiple-choice VQA), the model exhibits substantial visual neglect (LLaVA-1.5), the baseline has a conservative bias, and the bias magnitude is calibrated. VIAR does not help for open-ended generation (GQA), when the model already maintains adequate visual attention (LLaVA-1.6), when the baseline is already calibrated (HallusionBench), or when the bias is applied to non-neglect layers. This pattern of successes and failures provides evidence for the mechanistic account: the neglect zone is a real phenomenon with measurable behavioral consequences.

Limitations. Attention weights are an imperfect proxy for information flow [13,25], though convergence of three independent methodologies mitigates this concern. The accuracy improvement is not statistically significant ($p = 0.272$; Cohen's $h = 0.007$; power 6.8%). The optimal bias and target layers require per-model calibration. Our cross-architecture coverage is limited to 7B-parameter models; whether the pattern holds for larger models or other direct-projection architectures remains open. The aggregate neglect depth does not predict individual hallucinations, and the intervention shows no improvement on open-ended VQA.

6 Conclusion

We have identified and characterized the visual neglect zone, a systematic structural pattern in direct-projection vision-language models where middle transformer layers allocate disproportionately low attention to visual tokens. Through a multi-method mechanistic analysis spanning four architectures, we establish the following findings. The U-shaped neglect zone is reproducible across direct-projection VLMs with different backbones (r=0.70) but absent in Q-Former-based and alternative architectures, confirming it as a structural consequence of the direct-projection design. Hidden-state probing reveals that layer 31 functions as a near-pure language prior readout (residual update cosine similarity

0.922). Gradient attribution independently validates the neglect zone through a 5× steeper-than-baseline decay of visual importance. Per-sample analysis reveals that hallucination vulnerability manifests at the neglect zone boundary (layers 14–17, $p < 0.002$).

The VIAR intervention demonstrates that the neglect zone is causally linked to model calibration: correcting the attention deficit shifts POPE yes-ratio from 42.8% to 50.0%, improves Brier score, and improves visually demanding MMStar categories by up to 5.6%. Bayesian analysis confirms the calibration shift with 99.95% posterior probability while accuracy improvement remains ambiguous (67.8% posterior; Cohen's h=0.007). Rather than claiming a new state-of-the-art method, we offer this work as a structural mechanistic contribution. For model developers, the neglect zone suggests that training procedures should encourage sustained visual grounding in middle layers. For the interpretability community, the convergence of attention-based, hidden-state, and gradient-based analyses demonstrates that multimodal transformers exhibit rich, characterizable internal structure amenable to systematic investigation.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197 (2020)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
4. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhu, F.: Are we on the right way for evaluating large vision-language models? In: Advances in Neural Information Processing Systems. vol. 37 (2024)
5. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 276–286 (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weisenbuch, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
7. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021)

8. Geva, M., Bastings, J., Filippova, K., Globerson, A.: Dissecting recall of factual associations in auto-regressive language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12216–12235 (2023)
9. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14204 (2024)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks pp. 1321–1330 (2017)
11. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
12. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6700–6709 (2019)
13. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 3543–3556 (2019)
14. Jiang, C., Xu, H., Ye, M., Ye, Q., Yan, M., Ji, H., Zhang, J., Huang, F., Huang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27036–27046 (2024)
15. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13763–13773 (2024)
16. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models pp. 19730–19742 (2023)
17. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
18. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36 (2024)
20. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in GPT. In: Advances in Neural Information Processing Systems. vol. 35, pp. 17359–17372 (2022)
21. Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: In-context learning and induction heads. Transformer Circuits Thread (2022)

22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763 (2021)
23. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
24. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5797–5808 (2019)
25. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 11–20 (2019)
26. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
27. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. In: International Conference on Learning Representations (2024)
28. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: International Conference on Learning Representations (2024)