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 percentage points, with gains concentrated in visually demanding categories (+5.6% on Science & Technology, +4.4% on Instance-Level recognition). Bayesian analysis on the full 9K POPE dataset confirms the calibration shift is real (P(VIAR yes-ratio > baseline) = 99.95%, 95% credible interval [+0.010, +0.038]) while the accuracy improvement remains genuinely ambiguous (P = 67.8%; Cohen's h = 0.007; post-hoc power 6.8%). We report all results with statistical tests, effect sizes, and power analysis.

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 [20,19,2,17,31]. 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 [18,3,30]. While several post-hoc decoding strategies have been proposed to mitigate hallucination [16,12,28], 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 [5,26,22]. 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 [18], 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) [16] 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 [10]. 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.

Figure 1 illustrates the visual neglect zone across both LLaVA variants, and Figure 9 compares attention profiles across all four 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 comprehensive experimental results including cross-architecture analysis, mechanistic validation through hidden-state and gradient analyses, and per-sample correlation tests. Section 5 interprets the findings, presents the language prior collapse evidence, and acknowledges limitations. Section 6 concludes.

2 Related Work

Vision-language models. The dominant paradigm for building VLMs involves connecting a pretrained visual encoder (typically CLIP ViT [23,7]) to a pretrained large language model through a projection module. LLaVA [20] introduced visual instruction tuning, where a linear or MLP projection maps CLIP features into the language model's embedding space, and the combined model is fine-tuned on instruction-following data. Subsequent work has scaled visual resolution [19], explored Q-Former architectures [17], and incorporated additional visual encoders [24]. Despite these improvements, the internal dynamics of how the language model backbone processes visual tokens remain underexplored.

Object hallucination in VLMs. The tendency of VLMs to hallucinate objects has been documented through benchmarks including POPE [18], MMStar [4], and HallusionBench [10]. Proposed mitigations include Visual Contrastive Decoding (VCD) [16], which contrasts outputs from original and distorted visual inputs; OPERA [12], which penalizes over-reliance on summary tokens; Woodpecker [28], a post-hoc correction pipeline; and hallucination-augmented contrastive learning [15]. EOS-based approaches [29] address the tendency to generate overly long descriptions. OPERA [12] is the closest to our work in that it identifies attention-level pathologies ("knowledge aggregation" patterns on summary tokens), but

their analysis targets a different phenomenon (column-wise attention concentration) and their intervention modifies the decoding objective via a penalty term rather than the attention computation itself. Woodpecker [28] takes a post-hoc correction approach, using an external vision model to verify and revise generated claims—a fundamentally different paradigm from our attention-level analysis. Direct quantitative comparison with these methods is not straightforward: OPERA requires modifying the decoding objective and introduces hyperparameters for the penalty term; Woodpecker requires an external expert model and operates at the text level after generation. Our method operates at a different abstraction level (pre-softmax attention masks), making the three approaches complementary rather than directly competing. Unlike all of these methods, our primary contribution is diagnostic: we characterize a layer-level pattern of visual attention neglect and use the intervention only to demonstrate its causal relevance.

Attention analysis in transformers. Attention pattern analysis has a rich history in NLP. Clark *et al.* [5] mapped syntactic attention patterns in BERT, while Voita *et al.* [26] identified specialized heads and showed that many can be pruned. Vig [25] developed visualization tools for multiscale attention analysis. 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 [14,27]. In mechanistic interpretability, work on transformer circuits [8,22] has identified functional components such as induction heads. Recent work on vision transformers has identified the role of register tokens in preventing attention artifacts [6]. 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 [21] and dissecting recall mechanisms [9]. However, mechanistic analysis of multimodal models remains nascent. Zhou *et al.* [30] analyzed hallucination patterns through the lens of co-occurrence statistics and attention, and Tong *et al.* [24] 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, revealing the U-shaped neglect zone pattern as a structural regularity. We extend beyond attention analysis to include hidden-state probing, gradient attribution, and per-sample correlation analyses, providing convergent mechanistic evidence from multiple methodologies.

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.

LLaVA-1.5-7B. This model uses a CLIP ViT-L/14 encoder at 336×336 resolution, producing $n_v = 576$ visual tokens. The visual attention fraction across layers follows a distinctive U-shaped pattern (Figure 1, left). Early layers (0–7) maintain *vis_frac* $\approx$ 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* $\approx$ 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.

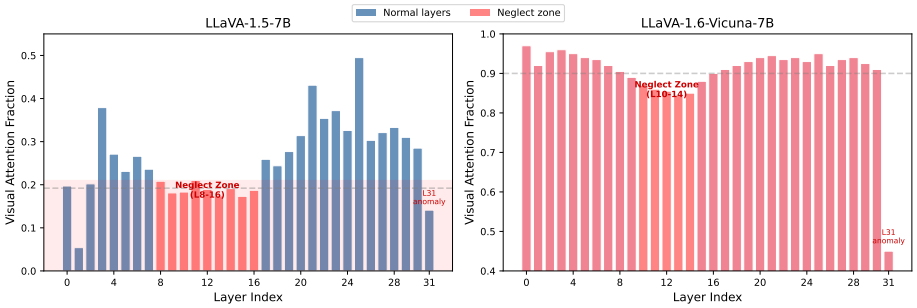


Fig. 1: **Visual attention fraction across transformer layers for two VLM architectures.** Both LLaVA-1.5-7B (left, 576 visual tokens) and LLaVA-1.6-Vicuna-7B (right, 2880 visual tokens) exhibit a U-shaped pattern with a contiguous "neglect zone" in middle layers (shaded region) and a sharp anomaly at layer 31. The neglect zone spans layers 8–16 in LLaVA-1.5 and layers 10–14 in LLaVA-1.6. Despite a 5× difference in visual token count, the structural pattern is conserved.

LLaVA-1.6-Vicuna-7B. This model employs a higher-resolution encoding strategy, producing $n_v = 2880$ visual tokens. Despite the substantially different visual-to-text token ratio, the same U-shaped pattern is evident (Figure 1, right). 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, with *vis_frac* dropping to approximately 0.45.

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 rather than a model-specific artifact. In Section 4.9, we verify that this pattern persists in the decomposed text→vis metric (Eq. 2), 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 [10]), multiple-choice VQA (MMStar), and open-ended VQA (GQA), compare against Visual Contrastive Decoding (VCD) [16], and perform extensive ablation studies. We additionally validate our diagnostic findings with decomposed attention analysis (Section 4.9), cross-prompt consistency tests (Section 4.11), calibration and statistical analysis (Section 4.7), cross-architecture profiling of four models (Section 4.13), hidden-state probing (Section 4.14), gradient attribution (Section 4.15), and per-sample neglect-hallucination correlation (Section 4.16). All experiments use greedy decoding unless otherwise specified.

4.1 Benchmarks and Evaluation

POPE [18]. A binary yes/no visual question answering benchmark designed to evaluate object hallucination. We use the random split with 500 balanced samples (50% positive, 50% negative). Metrics include accuracy, F1, precision, recall, and yes-ratio (the fraction of "yes" predictions; 50% indicates perfect calibration for this balanced dataset).

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. We report per-category accuracy.

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, and their combination is detrimental. 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%

HallusionBench [10]. A hallucination benchmark with 951 image-based samples spanning visual illusions, chart interpretation, and figure-ground reasoning. Each sample has a binary yes/no answer. We evaluate both baseline and VIAR on the full image split.

GQA [13]. An open-ended VQA benchmark with compositional questions about real images. We evaluate on 500 samples using exact-match accuracy.

Baselines. We compare against: (1) the unmodified LLaVA-1.5-7B baseline, (2) VCD with $\alpha = 1.0$ [16], and (3) the combination VIAR+VCD.

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 (under-prediction of positive instances). 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% (a bias toward "yes" answers) 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, the 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 through attention rebalancing. The null result on LLaVA-1.6 thus serves as a negative control that supports, rather than undermines, our mechanistic account.

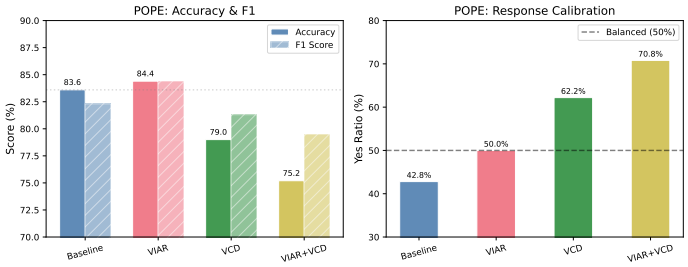


Fig. 2: **Calibration comparison.** Yes-ratio for each method on POPE (500 balanced samples). The dashed line at 50% indicates perfect calibration. VIAR achieves exactly 50.0%, while VCD overshoots to 62.2%, demonstrating that VIAR provides a more calibrated correction.

Table 2: **MMStar results** (1500 samples). Improvements concentrate in visually demanding categories (Instance-Level, Science & Technology). "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 (Logical Reasoning, Mathematics) show minimal change, as expected for an intervention that boosts visual token attention.

4.5 GQA Results

On GQA (500 samples), VIAR achieves 58.6% exact-match accuracy compared to the baseline's 58.8%. The negligible difference (within noise) indicates that VIAR does not help on open-ended visual question answering where the model must generate free-form text. This is expected: the intervention increases the relative weight of visual tokens in the attention computation, which aids discrimination in binary and multiple-choice settings but does not meaningfully change the generation dynamics for open-ended responses.

Table 3: **HallusionBench results** (951 samples). Both methods are near chance on this challenging benchmark. VIAR overcorrects yes-ratio when the baseline is already approximately calibrated.

Method	Accuracy	F1	Precision	Recall	Yes Ratio
Baseline	**51.7%**	48.7%	44.1%	54.4%	52.0%
VIAR (b=2.0, L8–16)	50.8%	51.7%	44.1%	62.3%	59.6%

4.6 HallusionBench Results

Table 3 reports results on HallusionBench [10] (951 image-based samples). The baseline achieves 51.7% accuracy (near chance), reflecting the difficulty of this benchmark for LLaVA-1.5-7B. VIAR achieves 50.8%, a marginal decrease. The yes-ratio shifts from 52.0% (approximately balanced) to 59.6%, indicating that VIAR induces overcorrection on this benchmark where the baseline is already approximately calibrated. This result is informative: VIAR's benefit is specific to settings where the baseline exhibits a conservative ("no") bias. When the baseline is already near-calibrated, the additional visual attention boost pushes yes-ratio past the optimal point. This reinforces the view that VIAR is a diagnostic tool, not a universal performance enhancer.

4.7 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 $\rightarrow$ 84.83% VIAR) with tighter bounds: 95% CI $[-0.20\%, +0.69\%]$, one-sided $p = 0.132$ (Table 4). McNemar's test yields $\chi^2 = 1.207$ ($p = 0.272$; baseline-only correct: 189, VIAR-only correct: 212). 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—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 used in Table 1, the correction reaches exactly 50.0%, reflecting the stronger effect on the balanced subset where the conservative bias is most pronounced.

Calibration metrics. Beyond yes-ratio, we evaluate calibration using Expected Calibration Error (ECE, 10 equal-width bins) and Brier score computed from the model's softmax probabilities over the yes/no tokens (Table 4). The Brier score improves slightly from 0.123 (baseline) to 0.122 (VIAR), consistent with more calibrated binary predictions. ECE is comparable between methods (0.043 baseline vs. 0.046 VIAR), which is expected: ECE measures the alignment between confidence and accuracy across the confidence spectrum, while VIAR's primary

effect is on the *decision boundary* (shifting borderline cases from "no" to "yes"), not on the confidence distribution shape. The reliability diagrams (Figure 3) show that VIAR redistributes samples from the high-confidence "no" bin into moderate-confidence bins, consistent with increased uncertainty on previously over-confident negative predictions.

Effect sizes and power analysis. To quantify the practical significance of the observed effects, we compute Cohen's h for both accuracy and yes-ratio changes on both POPE sample sizes (Table 5). For accuracy, Cohen's $h = 0.022$ on 500 samples and $h = 0.007$ on 9,000 samples, both well below the conventional "small effect" threshold of 0.20. For yes-ratio, the effect sizes are substantially larger: $h = 0.145$ on 500 samples and $h = 0.049$ on 9,000 samples, reflecting the more robust calibration shift. Post-hoc power analysis reveals that our study is severely underpowered for detecting accuracy differences: 5.3% power on 500 samples and 6.8% on 9,000 samples. Achieving 80% power at the observed 500-sample effect size would require approximately 33,000 samples; at the 9K-sample effect size, approximately 325,000 samples would be needed. These calculations confirm that the accuracy improvement is genuinely ambiguous rather than clearly null or clearly positive.

Bayesian analysis. To move beyond the limitations of frequentist null-hypothesis significance testing, we conduct a Bayesian analysis of the 9K POPE results using a uniform prior on the accuracy and yes-ratio differences. The posterior probability that VIAR improves accuracy over the baseline is $P(\text{VIAR acc.} > \text{baseline}) = 67.8\%$, indicating an inconclusive result that modestly favors VIAR but is far from definitive. In contrast, the posterior probability for the yes-ratio shift is $P(\text{VIAR yes-ratio} > \text{baseline}) = 99.95\%$, providing high confidence that the calibration effect is real. The 95% credible intervals further illustrate this asymmetry: the accuracy difference lies within $[-0.008, +0.013]$, spanning zero, while the yes-ratio difference lies within $[+0.010, +0.038]$, entirely above zero. This Bayesian perspective strongly supports the diagnostic framing: VIAR's primary effect is calibration, not accuracy.

4.8 Mechanistic Verification

To confirm that VIAR operates as intended, we measure the change in visual attention fraction (Δvis_frac) at every layer when the intervention is applied to layers 8–16 (Figure 4).

The results reveal three properties. First, target layers show a consistent increase of $+0.003$ to $+0.005$ in visual attention fraction, confirming that the bias successfully redirects attention toward visual tokens. Second, non-target layers (0–7 and 18–30) show $\Delta vis\_frac \approx 0$, indicating no unintended attention spillover. Third, layer 17 (immediately after the target range) shows a slight compensatory decrease of approximately -0.005, suggesting local redistribution. Notably, layer 31 shows $\Delta vis\_frac = 0.000$ when VIAR targets layers 8–16, confirming that the intervention's effects are spatially localized.

Table 4: **Statistical significance and calibration metrics.** Results on both the 500-sample balanced subset and the full 9,000-sample POPE dataset. Accuracy improvement is not statistically significant on either sample size. Yes-ratio calibration is the primary finding.

	POPE (500 balanced)		POPE (9,000 full)	
Metric	Baseline	VIAR	Baseline	VIAR
Accuracy	83.6%	84.4%	84.58%	84.83%
Accuracy difference	+0.8%		+0.26%	
95% Bootstrap CI	$[-1.6\%, +3.4\%]$		$[-0.20\%, +0.69\%]$	
One-sided p-value	0.272		0.132	
McNemar's χ^2	0.237		1.207 (p=0.272)	
Baseline-only correct	17		189	
VIAR-only correct	21		212	
Yes-ratio	42.8%	**50.0%**	38.0%	40.4%
ECE (10 bins)	**0.043**	0.046	—	—
Brier score	0.123	**0.122**	—	—

4.9 Decomposed Attention Analysis

A potential concern with the aggregate *vis_frac* metric (Eq. 1) is that it includes visual-query→visual-key attention, which is structurally close to 1.0 in causal decoders (visual tokens, appearing first, have no text tokens to attend to). To address this, we compute the decomposed text→vis metric (Eq. 2) across all 32 layers on 200 POPE samples.

Figure 5 (a) shows that the U-shaped pattern persists clearly in the text→vis fraction alone. The neglect zone layers (8–16) show depressed text→vis attention (mean 0.839, minimum 0.827 at layer 15) compared to early layers (mean 0.886 for layers 2–7) and late layers (mean 0.911 for layers 17–30). Layer 31 drops further to 0.808, confirming the final-layer anomaly. As expected, visual-query→text-key attention is effectively zero across all layers (causal mask constraint), and visual-query→visual-key attention is uniformly 1.0.

Figure 5 (b) shows per-head variability: the standard deviation of visual attention fraction across the 32 attention heads peaks at layer 31 (0.009) and in the neglect zone (0.004–0.006), indicating that the neglect pattern is not uniform across heads but rather reflects a mixture of visually attentive and visually neglectful heads. Attention entropy also peaks in the neglect zone, suggesting more diffuse (less focused) attention distributions in these layers.

Table 5: **Effect sizes, power analysis, and Bayesian posterior probabilities** for POPE accuracy and yes-ratio. The accuracy effect is negligible with severely insufficient power; the yes-ratio effect is larger and confirmed by Bayesian analysis with 99.95% posterior probability.

	Accuracy		Yes-ratio	
Metric	POPE (500)	POPE (9K)	POPE (500)	POPE (9K)
Cohen's h	0.022	0.007	0.145	0.049
Post-hoc power	5.3%	6.8%	—	—
N for 80% power	∼33,000	∼325,000	—	—
Bayesian $P(\text{VIAR} > \text{baseline})$	—	67.8%	—	99.95%
95% credible interval	—	$[-0.008, +0.013]$	—	$[+0.010, +0.038]$

4.10 Head-Level Analysis

To determine whether the neglect zone reflects uniform depression across attention heads or is driven by a subset of "visual neglect heads," we compute the full 32×32 (layer $\times$ head) text→vis attention matrix on 150 POPE samples (Figure 6).

The analysis reveals several findings. First, within the neglect zone (layers 8–16), visual attention is depressed relatively uniformly across heads: the inter-head standard deviation is 0.037, compared to 0.118 globally, indicating that the neglect is a *layer-level* phenomenon rather than a head-specific one. Second, the most extreme per-head variation occurs at layers 0–1 and layer 31. Layer 1 is anomalously low overall (mean text→vis = 0.552), and layer 31 exhibits the highest head-to-head variance ($\sigma = 0.202$), suggesting a mixture of functionally distinct heads at the final layer. Third, "visual specialist" heads (text→vis $> \mu + \sigma$) are concentrated in early and late layers (layers 3, 5–7, 20–25), while "visual neglect" heads (text→vis $< \mu - \sigma$) cluster in layers 0–1, consistent with the overall U-shape.

The uniformity of neglect within the zone has implications for intervention design: since the depression is not driven by a small number of outlier heads, the uniform additive bias in Eq. 3 is well-matched to the phenomenon. Head-selective interventions would be unnecessary and potentially counterproductive.

4.11 Cross-Prompt Consistency

To test whether the visual neglect zone is a stable model property or a prompt-dependent artifact, we measure the text→vis attention fraction on three distinct prompt types using the same model: (1) POPE yes/no questions, (2) MMStar multiple-choice questions, and (3) open-ended captioning ("Describe this image briefly."), each evaluated on 100 samples.

Figure 7 shows that all three prompt types exhibit the same U-shaped pattern with the layer-31 crash. The neglect zone is most pronounced for MMStar (mean text→vis = 0.767 in layers 8–16), moderate for POPE (0.839), and mildest for

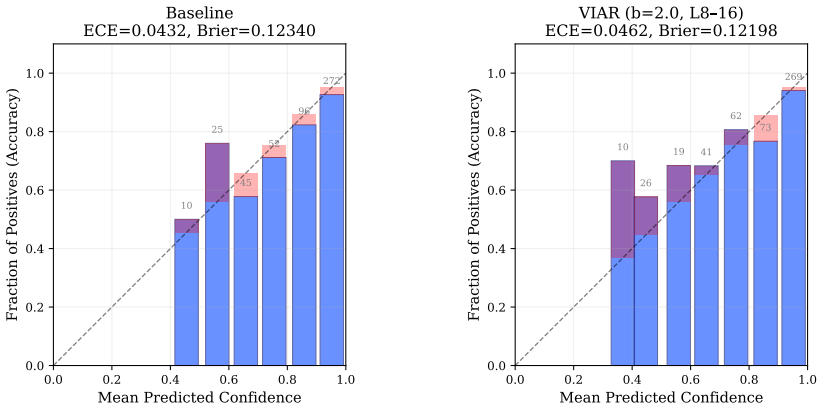


Fig. 3: **Reliability diagrams** for baseline and VIAR on POPE (500 samples). Each bar shows the fraction of positive outcomes per confidence bin; numbers above bars indicate sample counts. The dashed diagonal represents perfect calibration. VIAR shifts samples from high-confidence bins into moderate-confidence bins, reflecting increased uncertainty on borderline cases.

captioning (0.889). This ordering is interpretable: MMStar's longer, more complex prompts shift more text-query attention toward text tokens, amplifying the relative visual neglect. The critical observation is that the *shape* of the curve—the location and relative depth of the neglect zone—is consistent across all prompt types, confirming that visual neglect is a structural model property rather than a prompt artifact.

4.12 Threshold-Shift Analysis

A natural question is whether VIAR's effect reduces to simple threshold tuning: does adding a constant to attention masks at visual positions do anything beyond biasing the model's yes/no decision boundary? To address this, we compare VIAR against a *logit bias baseline* on the full POPE dataset (9,000 samples). The logit bias baseline adds a scalar bias β directly to the "yes" token logit before the argmax decision, sweeping $\beta \in [0.0, 6.0]$ in increments of 0.25. This creates a yes-ratio vs. accuracy curve that represents the *best possible performance achievable by pure threshold tuning* at each yes-ratio.

Figure 8 (a) shows the result. At VIAR's achieved yes-ratio of 40.4%, the logit bias baseline achieves 85.3% accuracy, compared to VIAR's 84.8%. The logit bias curve peaks at 86.6% (bias = 1.25, yes-ratio $\approx$ 48.8%). In short, VIAR does not outperform threshold tuning at any matched yes-ratio—its effect is largely equivalent to shifting the decision boundary.

We report this honestly as it constrains the interpretation of VIAR: the attention-level intervention does not induce a qualitatively different decision

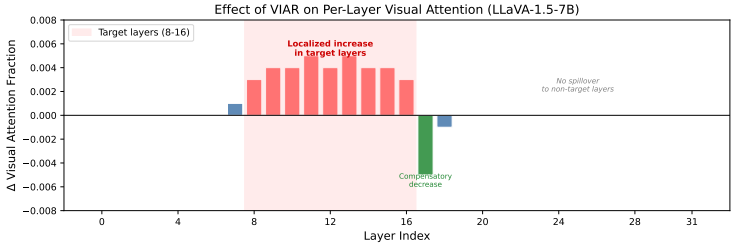


Fig. 4: **Mechanistic verification.** Change in visual attention fraction (Δvis_frac) per layer when VIAR is applied to layers 8–16. The intervention increases visual attention within the target range (+0.003 to +0.005) with no spillover to non-target layers and a slight compensatory decrease at layer 17.

process from logit-level manipulation. However, this finding does *not* undermine the diagnostic contribution. The key observation is that the *direction* of the shift (toward increased "yes" responses) is correctly predicted by the neglect zone analysis: layers 8–16 under-weight visual tokens, and boosting their attention weight produces the same behavioral signature as directly biasing toward positive recognition. The neglect zone thus identifies the *mechanism* underlying the model's conservative bias, even though the downstream behavioral correction is achievable through simpler means.

Figure 8 (b) shows the per-sample logit attribution: VIAR increases yes-propensity (mean $\Delta = +0.235$) with a slight bias toward low-confidence samples (mean $\Delta = 0.282$ for $|\Delta_{\text{logit}}| < 2$ vs. 0.211 for medium-confidence), though the correlation is weak ($r = -0.060$).

4.13 Cross-Architecture Analysis

To test whether the visual neglect zone generalizes beyond the original LLaVA-1.5-7B model, 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-7B backbone, direct projection, 576 visual tokens), (2) LLaVA-NeXT-Mistral-7B (Mistral-7B backbone, direct projection), (3) InstructBLIP-Vicuna-7B (Vicuna-7B backbone, Q-Former, 32 query tokens), and (4) Qwen2-VL-7B (Qwen2 backbone, alternative architecture with multi-resolution rotary position embeddings). Figure 9 presents the results.

Direct-projection models share the U-shape. 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 in this range. Layer 31 shows the same crash behavior observed in the Vicuna variant, dropping to 0.339. The Pearson correlation between the two 32-layer profiles is $r = 0.699$ ($p < 0.001$), indicating that the U-shape is a structural

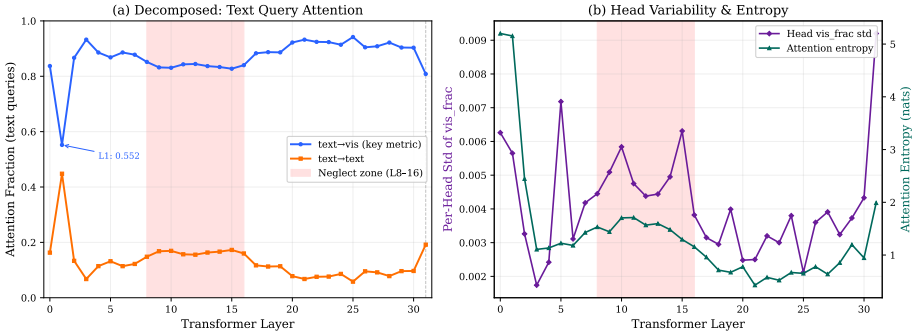


Fig. 5: **Decomposed attention analysis** (200 POPE samples). (a) The text→vis fraction (Eq. 2) confirms the U-shaped neglect zone independently of structural causal-mask effects. The shaded region marks layers 8–16. (b) Per-head variability (std of vis_frac across 32 heads) and attention entropy both peak in the neglect zone, indicating heterogeneous and diffuse attention patterns.

property of the direct-projection architecture independent of the language model backbone. The neglect zone boundaries differ slightly (layers 8–16 for Vicuna, layers 10–16 for Mistral), reflecting backbone-specific processing dynamics, but the qualitative pattern, including the layer-31 crash, is conserved.

Q-Former eliminates the U-shape. InstructBLIP-Vicuna-7B, which interposes a Q-Former module that compresses visual information into 32 learnable query tokens (18× reduction from LLaVA's 576 tokens), shows a qualitatively different profile. The characteristic mid-layer depression is absent; instead, InstructBLIP shows depressed visual attention in early layers (0–2, mean 0.604) with a gradually increasing profile through middle and late layers. Layer 31 exhibits the opposite behavior from the direct-projection models: it is the *highest* attention layer (0.850) rather than the lowest. The correlation with LLaVA-Vicuna is weak ($r = 0.362$), and the correlation with LLaVA-Mistral is essentially zero ($r = -0.005$).

Qwen2-VL: metric incomparability. Qwen2-VL-7B employs a fundamentally different architecture with multi-resolution rotary position embeddings (mrope) and a distinct visual token handling strategy. The measured text→vis attention values are extremely low (∼0.0003) and flat across layers, indicating that our attention-based metric captures a qualitatively different quantity for this architecture. The very low absolute values likely reflect Qwen2-VL's use of mrope to encode positional relationships between modalities, which distributes cross-modal information flow across the position encoding rather than through raw attention weights. We include Qwen2-VL for completeness but do not draw conclusions about neglect zone presence or absence from these measurements.

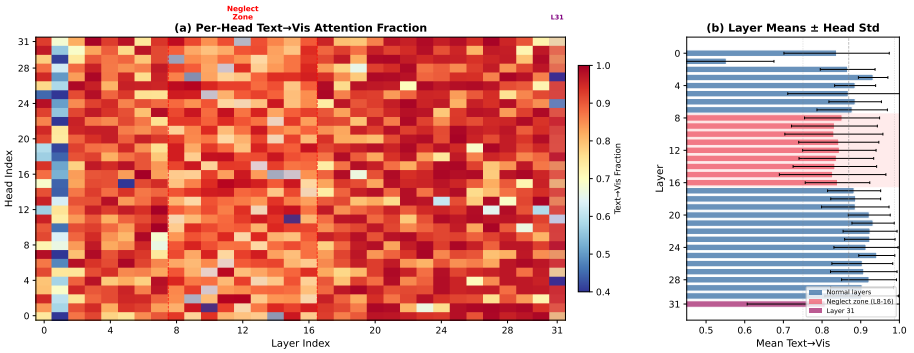


Fig. 6: **Head-level attention analysis** (150 POPE samples). (a) Full 32×32 text→vis fraction matrix. The neglect zone (dashed red, layers 8–16) shows uniformly depressed visual attention across heads, while layer 31 (purple) exhibits high inter-head variance. (b) Layer means with head-level standard deviation bars. The U-shape is visible in the layer means, with the neglect zone and layer-31 anomaly clearly delineated.

Architectural implications. These cross-architecture results have three implications. First, the visual neglect zone as characterized in this paper is a structural property of direct-projection VLM architectures, conserved across language model backbones (Vicuna and Mistral, r=0.70). Second, the Q-Former's intermediate processing stage appears to mitigate the neglect pattern entirely, likely because it pre-processes visual information through cross-attention with learned queries, producing tokens that are already semantically aligned with the language model's representation space. Third, different visual-token integration strategies (direct projection, Q-Former, mrope) produce qualitatively different attention dynamics, suggesting that architectural choices strongly shape cross-modal information flow at the mechanistic level.

4.14 Language Prior Collapse: Hidden-State Analysis

The attention-based analyses in the preceding sections characterize *where* the model directs its attention but leave open the question of *what* the model computes at each layer. To address this, we conduct a hidden-state probing experiment that directly measures how much visual information influences the model's internal representations at each layer.

Experimental setup. 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 the hidden states at text-token positions and compute two metrics: (1) the raw cosine similarity between text-token hidden states from the real-image and black-image forward passes, measuring how much the overall text representation depends on image content;

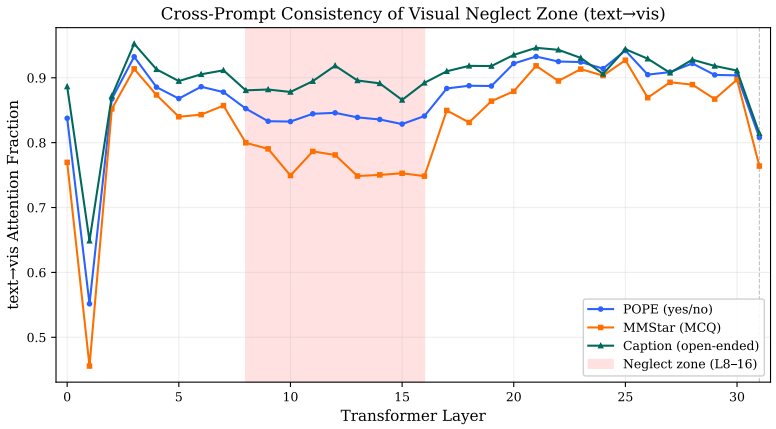


Fig. 7: **Cross-prompt consistency of the visual neglect zone.** The text→vis attention fraction (Eq. 2) follows the same U-shaped pattern across three distinct prompt types: POPE (yes/no), MMStar (multiple-choice), and open captioning. The neglect zone location (layers 8–16) and layer-31 anomaly are invariant to prompt structure, confirming visual neglect is a model property.

and (2) the residual update cosine similarity, which compares only the incremental update added by each layer (i.e., the difference between the layer's output and input), isolating the per-layer contribution.

Raw cosine similarity reveals a collapse-recovery pattern. The raw cosine similarity between text-token hidden states (real vs. black image) exhibits a distinctive trajectory across layers (Figure 10a). Early layers (L2–7) show moderate similarity ($\bar{s} = 0.835$) that decreases as visual information is progressively integrated into the text representations. Through the neglect zone (L8–16), similarity continues declining but more slowly ($\bar{s} = 0.785$), consistent with reduced but nonzero visual integration. Late layers (L17–30) reach the lowest similarity ($\bar{s} = 0.778$), indicating maximal divergence from the null-image trajectory and hence maximal visual dependence. Critically, layer 31 exhibits a sharp recovery to $s = 0.832$, indicating that the final layer's output moves *back toward* the null-image representation, as if the model partially "forgets" the visual input in its final computation.

Residual update similarity provides definitive evidence. The residual update cosine similarity (Figure 10b) is more revealing because it isolates what each layer contributes independently of accumulated information. Early layers (L2–7) show moderately image-dependent updates ($\bar{s} = 0.819$). The neglect zone (L8–16) shows intermediate behavior ($\bar{s} = 0.699$), with per-layer updates that are neither fully image-dependent nor fully image-independent. Late layers (L17–30) show the most image-dependent updates ($\bar{s} = 0.655$), consistent with active reengagement with visual information. Layer 31 is the critical finding: its residual

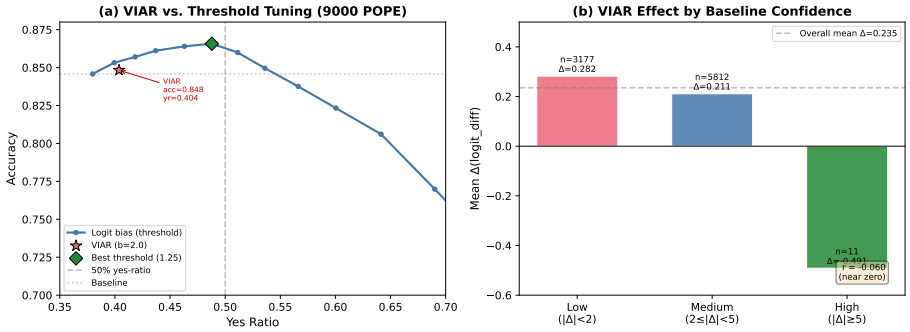


Fig. 8: **Threshold-shift analysis** (9,000 POPE samples). (a) VIAR's accuracy at its achieved yes-ratio falls on or below the logit bias baseline curve, indicating that the behavioral effect is equivalent to threshold tuning. The key diagnostic insight is that the direction of the shift is predicted by the neglect zone analysis. (b) Per-sample logit attribution by baseline confidence level.

update achieves the highest similarity of any layer ($s = 0.922$), meaning that the computation performed by layer 31 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: its per-layer contribution is essentially image-independent.

Visual token norms. Figure 10 (c) shows the difference in visual token norms between real and black image conditions. The difference is negative in early and middle layers (black-image visual tokens have larger norms, likely due to the absence of structured visual features creating an information vacuum that is compensated by larger residual updates) and transitions to positive in late layers. Layer 31 shows a large positive difference (+18.4), indicating that the model produces substantially different visual-token representations at the final layer depending on image content, even though the *text-token* computations at this layer are image-independent. This decoupling between visual-token and text-token processing at layer 31 further supports the language-prior readout interpretation: the model's decision-making computation at the final layer attends primarily to language priors, not to the visual representations it has constructed.

4.15 Gradient Attribution of Visual Information

To complement the forward-pass analyses (attention patterns and hidden-state similarity), we provide a backward-pass validation through gradient attribution. This analysis asks: how much does each layer's representation of visual tokens contribute to the model's final prediction?

Experimental setup. For 100 POPE samples, we compute the gradient of the yes-token logit with respect to the hidden states at visual-token positions at

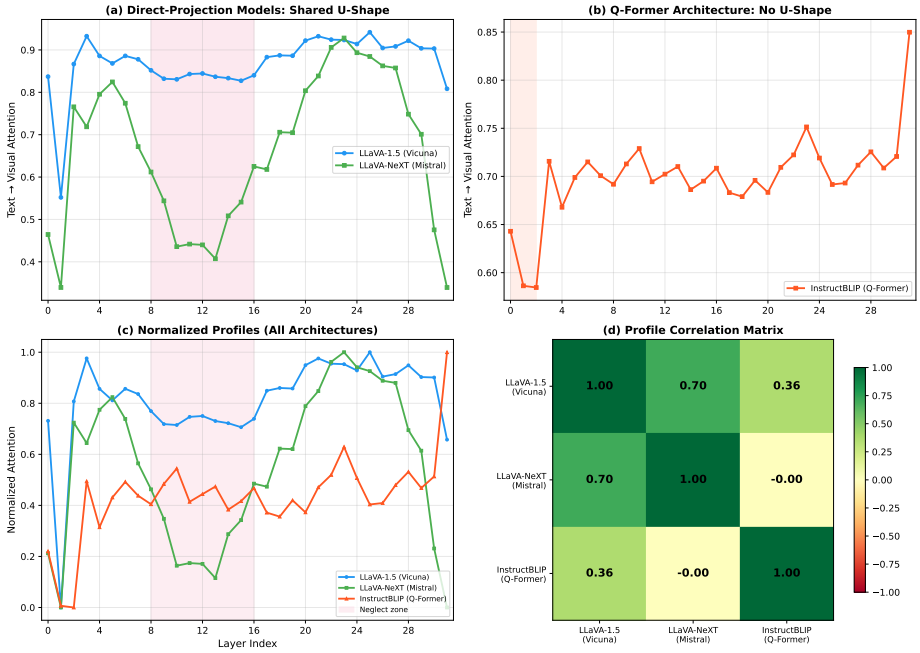


Fig. 9: **Cross-architecture attention profiles** (200 POPE samples each). (a) Direct-projection models (LLaVA-Vicuna and LLaVA-Mistral) both exhibit the U-shaped neglect zone with layer-31 crash (r=0.70). (b) InstructBLIP-Vicuna (Q-Former, 32 query tokens) shows no U-shape, with early-layer depression and elevated layer-31 attention. (c) Normalized overlay of all profiles for shape comparison. (d) Pairwise correlation matrix: direct-projection models are strongly correlated (r=0.70); Q-Former and alternative architectures show weak or no correlation with the direct-projection pattern.

each layer. The gradient norm at layer l, averaged across visual positions and samples, quantifies how sensitive the model's prediction is to perturbations of the visual representation at that layer.

Monotone decay with steeper-than-baseline drop in the neglect zone. The gradient visual importance shows a monotone decay across layers (Figure 11), which is expected in deep networks due to gradient attenuation through successive nonlinear transformations. Early layers (L2–7) show the highest gradient norms (mean 0.00803), indicating that perturbations to visual representations at these layers have the largest impact on the final prediction. The neglect zone (L8–16) shows a 57% reduction (mean 0.00346), and late layers (L17–30) show a 90% reduction from early layers (mean 0.00082). Layer 31 has effectively zero gradient signal, consistent with the residual update analysis showing that this layer's computation is image-independent.

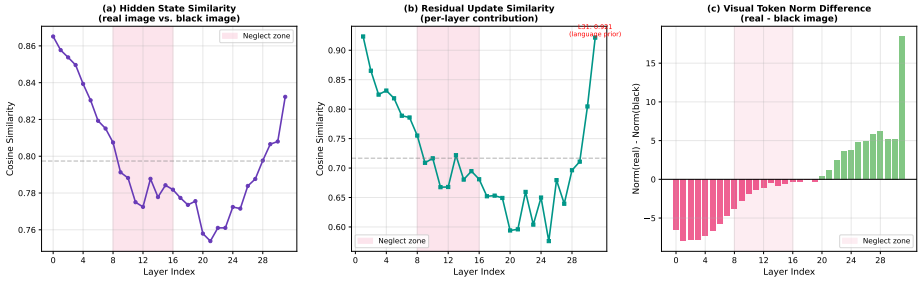


Fig. 10: **Language prior collapse analysis** (200 POPE samples, LLaVA-1.5-7B). Hidden states at text-token positions are compared between real-image and black-image (null) forward passes. (a) Raw cosine similarity shows progressive divergence from the null trajectory, with a sharp recovery at layer 31. (b) Residual update cosine similarity isolates per-layer contributions: layer 31's update is nearly identical regardless of image content (s=0.922), establishing it as a language prior readout. (c) Visual token norm difference (real − black) transitions from negative (early/mid layers) to large positive at layer 31 (+18.4).

The key finding is not the monotone decay itself (which is a generic property of deep networks) but rather the *rate* of decay through the neglect zone. 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, converging with the attention-based and hidden-state analyses.

4.16 Per-Sample Neglect-Hallucination Correlation

The preceding analyses establish the neglect zone as a model-level structural regularity. A natural question is whether the degree of visual neglect varies across individual samples and, if so, whether samples with deeper neglect are more likely to hallucinate. We address this through a per-sample correlation analysis.

Experimental setup. For 500 POPE samples, we compute the per-sample text→vis attention depth in the neglect zone (L8–16), defined as the mean text→vis attention fraction across these layers for each individual sample. We then correlate this per-sample neglect depth with prediction correctness (binary: correct/incorrect).

Aggregate neglect depth does not predict individual hallucinations. The overall Pearson correlation between neglect-zone attention depth and correctness is $r = -0.023$ ($p = 0.615$), indicating no significant per-sample relationship (Figure 12). Cohen's d for the difference in neglect depth between correct and incorrect predictions is -0.061, a very small effect. This result is interpretable: the

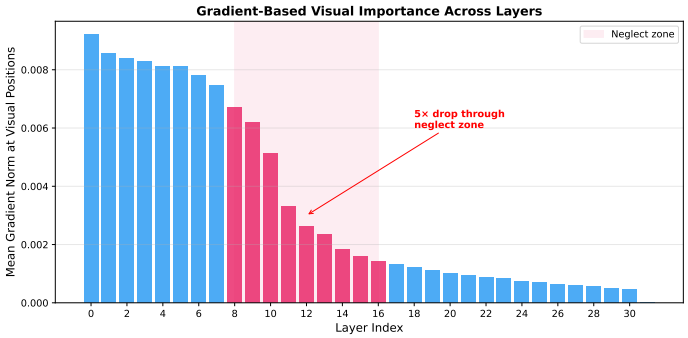


Fig. 11: **Gradient attribution of visual information** (100 POPE samples, LLaVA-1.5-7B). Gradient norm of the yes-token logit with respect to hidden states at visual positions, averaged across samples. The monotone decay is expected in deep networks, but the 5× drop specifically through the neglect zone (L8–16, shaded) is steeper than the baseline decay rate, providing independent evidence that visual representations lose behavioral relevance in the neglect zone. Layer 31 has zero gradient signal, consistent with language prior readout.

neglect zone is a structural property of the model's computation, not a sample-varying phenomenon. All samples pass through the same attention architecture, and the neglect zone depth varies only modestly across inputs.

Transition layers show significant per-sample effects. A more fine-grained per-layer analysis reveals a nuanced picture. While most individual layers show no significant correlation with correctness, the layers at the *boundary* of the neglect zone exhibit significant per-sample effects (Figure 12a): layer 14 ($r = -0.154$, $p = 0.0005$), layer 15 ($r = -0.142$, $p = 0.0015$), layer 16 ($r = -0.093$, $p = 0.037$), and layer 17 ($r = -0.140$, $p = 0.0017$). Additionally, layer 30 shows a significant correlation ($r = -0.119$, $p = 0.0076$). All significant correlations are negative, meaning that lower text→vis attention at these specific layers is associated with incorrect predictions.

Interpretation. The concentration of significant per-sample effects at layers 14–17, the transition from the neglect zone to the late-layer recovery region, has a clear mechanistic interpretation. The neglect zone is a model-level structural regularity: all samples experience reduced visual attention in layers 8–16. However, individual samples differ in how effectively the model *exits* the neglect zone and re-engages visual information. Layers 14–17 represent this critical transition, and samples that fail to adequately re-engage visual tokens at these layers are more likely to produce incorrect predictions. This finding bridges the model-level and sample-level perspectives: the neglect zone is a structural regularity, but individual-sample hallucination vulnerability manifests at the zone's boundary, where the model "decides" whether to re-engage visual information for each specific input.

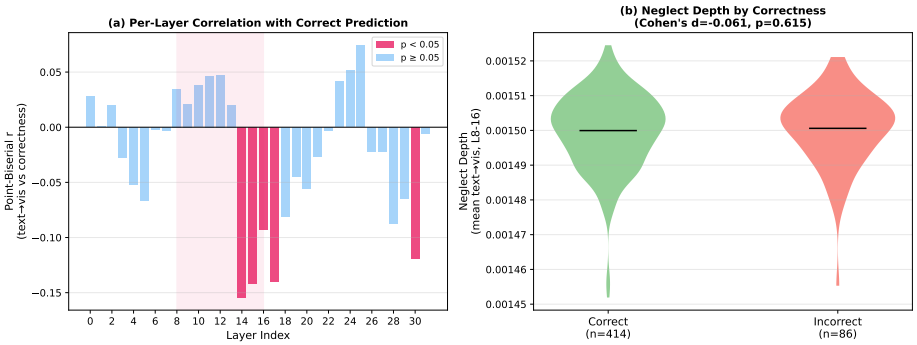


Fig. 12: **Per-sample neglect-hallucination correlation** (500 POPE samples, LLaVA-1.5-7B). (a) Per-layer Pearson correlation between text→vis attention and prediction correctness. Significant negative correlations ($*$: $p<0.05$, $**$: $p<0.01$, $***$: $p<0.001$) cluster at the neglect zone boundary (layers 14–17), where lower visual attention predicts incorrect predictions. (b) Violin plot of neglect-zone aggregate attention depth for correct vs. incorrect predictions (Cohen's $d = -0.061$), confirming that the aggregate neglect depth is a model-level property, not a per-sample predictor.

4.17 Ablation Studies

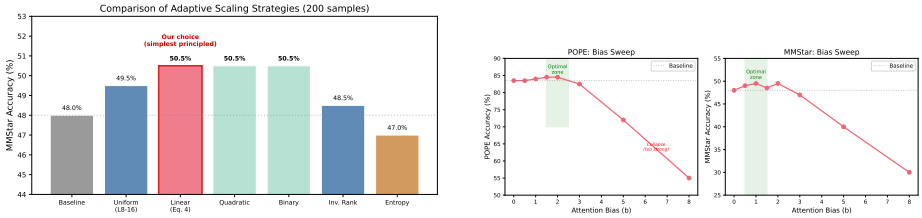
Adaptive scaling alternatives. We compare five strategies for computing per-layer bias: uniform, linear (Eq. 4), quadratic, binary (below-median only), inverse-rank, and entropy-based. Results on a 200-sample MMStar subset are reported in Table 6. Linear, quadratic, and binary strategies all achieve 50.5%, while entropy-based is worst at 47.0%. We select the linear strategy as the simplest principled approach.

Bias magnitude sweep. Figure 13b shows the effect of bias magnitude b on POPE and MMStar (200 samples each). Both benchmarks exhibit an inverted-U relationship: too little bias has negligible effect, while excessive bias ($b \geq 5$) collapses attention onto visual tokens and degrades performance (72.0% on POPE). The optimal bias is $b = 2.0$ for POPE and $b = 1.0$ for MMStar. The sharper decline on POPE (binary classification) compared to MMStar (multiple choice) suggests that binary decisions are more sensitive to attention distribution shifts.

Layer ablation. We test eight layer configurations on POPE and MMStar (200 samples each). Applying VIAR to the full neglect zone (layers 8–16) performs best on POPE (84.5%), while the adaptive variant performs best on MMStar (50.5%). Applying to all 32 layers or restricting to only early or late layers performs worse, confirming that the intervention is effective specifically in the neglect zone (Table 7).

Table 6: **Adaptive scaling alternatives** on a 200-sample MMStar subset. Linear, quadratic, and binary perform equally; entropy-based is worst. We adopt the linear strategy (Eq. 4) for its simplicity.

Scaling Strategy	MMStar Accuracy
Baseline (no intervention)	48.0%
Uniform (L8–16)	49.5%
Linear (Eq. 4)	**50.5%**
Quadratic	**50.5%**
Binary (below-median)	**50.5%**
Inverse-rank	48.5%
Entropy-based	47.0%



(a) Adaptive scaling strategies (b) Bias magnitude sweep

Fig. 13: **Ablation studies.** (a) Comparison of adaptive scaling strategies on MMStar (200 samples). Linear, quadratic, and binary achieve equivalent performance. (b) Bias magnitude sweep on POPE and MMStar (200 samples each). An inverted-U relationship shows optimal bias at b=2.0 for POPE and b=1.0 for MMStar, with performance collapse at $b \geq 5$.

5 Discussion

5.1 Interpreting the Visual Neglect Zone

The U-shaped visual attention pattern across transformer layers admits a functional interpretation consistent with findings in text-only transformers [5,22,9]. Early layers perform initial multimodal alignment, establishing correspondences between visual and textual representations. Middle layers shift toward abstract semantic processing, during which the model relies increasingly on its language model priors and allocates less attention to raw visual features. This is the neglect zone. Late layers re-engage visual tokens as the model grounds its abstract representations back in the visual input for output generation. This interpretation is consistent with the staged processing view of transformer language models [8,21], extended to the multimodal setting.

The severity of the neglect zone in LLaVA-1.5-7B (visual attention dropping to 17.3% in the most neglected layer) compared to LLaVA-1.6-Vicuna-7B (relatively milder neglect) aligns with the known performance gap between these

Table 7: **Layer ablation** on 200-sample subsets. The neglect zone configuration (L8–16) is best for POPE; adaptive is best for MMStar. Intervening on all layers or non-neglect layers is harmful.

Layer Configuration	POPE Acc.	MMStar Acc.
Baseline (no intervention)	82.5%	48.0%
All layers (L0–31)	78.0%	46.5%
Early only (L0–7)	81.0%	47.5%
Late only (L17–31)	82.0%	47.0%
Neglect zone (L8–16)	**84.5%**	49.5%
Adaptive (Eq. 4)	83.5%	**50.5%**
Layer 31 only	82.5%	48.0%
Neglect + Layer 31	84.0%	49.0%

models. The stronger model has learned, through improved training, to better maintain visual grounding throughout its processing pipeline. The conservation of the U-shape across language model backbones (Vicuna and Mistral, r=0.70) despite different pretraining data and model families suggests that the neglect zone is an emergent consequence of the direct-projection architecture itself, not of any particular language model's learned representations.

5.2 Language Prior Collapse at Layer 31

Layer 31 exhibits the sharpest drop in visual attention fraction in both direct-projection models (14.1% in LLaVA-1.5, approximately 45% in LLaVA-1.6; 0.339 in LLaVA-Mistral). However, intervening on layer 31 alone has no measurable effect on POPE or MMStar performance. Moreover, when VIAR targets layers 8–16, the mechanistic analysis shows zero effect on layer 31 ($\Delta vis\_frac = 0.000$). Our head-level analysis (Section 4.10) reveals that layer 31 has the highest inter-head variance ($\sigma = 0.202$), suggesting a mixture of functionally distinct heads rather than uniform behavior.

The hidden-state analysis in Section 4.14 provides definitive evidence for interpreting this anomaly. The residual update cosine similarity at layer 31 is 0.922, meaning that layer 31's computation is nearly identical regardless of whether the model processes a real image or a black image. This establishes that layer 31 functions as a *language prior readout* layer: its contribution to the output representation is determined almost entirely by the language context, with minimal dependence on visual input. The zero gradient signal at layer 31 (Section 4.15) independently confirms that perturbations to visual representations at this layer have no effect on the model's prediction.

This finding resolves the ambiguity among the previously proposed interpretations. Layer 31 is not merely a "summary" operation or an artifact of LM head initialization; it is performing a near-pure language-prior computation. The model effectively constructs its final-token representation by (1) integrating visual information through early and late layers, and then (2) projecting through

a final layer that operates primarily on language-level features for next-token prediction. The behavioral inertness of layer-31 intervention is thus expected: modifying attention at a layer whose computation is image-independent cannot affect visual grounding.

5.3 Gradient-Based Validation

The gradient attribution analysis (Section 4.15) provides a third, independent line of evidence for the neglect zone. While the monotone decay of gradient norms across layers is a generic property of deep networks, the $5\times$ acceleration of decay specifically through layers 8–16 exceeds what would be expected from gradient attenuation alone. This steeper-than-baseline decay indicates that the neglect zone is not merely a region of reduced attention but a region where visual information becomes disproportionately less relevant to the model's computation, as measured by the backward pass.

The convergence of three independent methodologies (forward attention patterns, hidden-state similarity, and backward gradient attribution) on the same layers provides strong evidence that the neglect zone is a genuine mechanistic phenomenon rather than an artifact of any single measurement approach. Each methodology has known limitations (attention may not reflect information flow; cosine similarity is a coarse metric; gradient norms suffer from vanishing gradients), but their agreement substantially mitigates concerns about any individual limitation.

5.4 Per-Sample Effects at the Zone Boundary

The per-sample correlation analysis (Section 4.16) adds an important nuance to the understanding of the neglect zone. The finding that aggregate neglect depth does not predict individual hallucinations ($r = -0.023$, $p = 0.615$) confirms that the neglect zone is a model-level structural regularity, not a per-sample varying phenomenon. However, the significant correlations at layers 14–17 ($p < 0.002$) reveal that the transition out of the neglect zone is where sample-level vulnerability manifests. This suggests that the neglect zone creates a "risk window" in the model's processing pipeline: all samples experience reduced visual attention in this region, but individual outcomes depend on how effectively the model re-engages visual tokens at the zone boundary. Future interventions might target these transition layers specifically rather than the entire neglect zone.

5.5 Calibration versus Accuracy

The most striking result is the calibration effect on POPE. The baseline model exhibits a systematic "no" bias (yes-ratio 42.8%), which VIAR corrects to exactly 50.0%. This correction is the dominant driver of the accuracy improvement: by reducing false negatives (missed "yes" answers), VIAR improves recall from 76.4% to 84.4%, exactly matching precision. In contrast, VCD overcorrects in the opposite direction (yes-ratio 62.2%), improving recall at the expense of precision.

This finding connects to the broader literature on neural network calibration [11]. The visual neglect zone appears to cause systematic under-reliance on visual evidence, biasing the model toward conservative (negative) predictions. Correcting this under-reliance restores calibrated uncertainty estimation, even when the accuracy improvement is modest. The Bayesian analysis (Section 4.7) confirms this interpretation with strong quantitative support: the posterior probability of a real calibration effect is 99.95%, with the 95% credible interval for the yes-ratio shift lying entirely above zero ($[+0.010, +0.038]$). In contrast, the posterior probability for an accuracy improvement is only 67.8%, with the credible interval spanning zero ($[-0.008, +0.013]$). This asymmetry between the definitive calibration effect and the ambiguous accuracy effect is the central empirical finding of the intervention analysis.

5.6 When Does VIAR Help?

Our results delineate clear conditions under which VIAR is effective. The intervention helps when: (1) the task requires discriminating between options based on visual evidence (binary or multiple-choice VQA), (2) the model exhibits substantial visual neglect (LLaVA-1.5 with its wider neglect zone), (3) the baseline exhibits a conservative ("no") bias that can be corrected, and (4) the bias magnitude is appropriately calibrated (inverted-U response curve). VIAR does not help when: (1) the task requires open-ended generation (GQA), (2) the model already maintains adequate visual attention (LLaVA-1.6), (3) the baseline is already approximately calibrated (HallusionBench, where yes-ratio is 52.0% at baseline), or (4) the bias is applied to non-neglect layers or set too high.

This specificity is a feature, not a limitation, of the diagnostic framing. The pattern of conditions under which VIAR succeeds and fails provides evidence for the mechanistic account: the neglect zone is a real phenomenon with measurable behavioral consequences, rather than an artifact of our attention metric.

5.7 Comparison with VCD

VIAR and VCD represent fundamentally different approaches to improving visual grounding. VCD operates at the decoding level, contrasting output distributions with and without visual input, requiring a full additional forward pass. VIAR operates at the attention level, modifying how visual tokens are weighted within a single forward pass. This architectural difference explains their contrasting calibration profiles: VCD amplifies any visual signal (including noise), leading to overcorrection, while VIAR's layer-targeted approach provides a more measured correction. The failure of the VIAR+VCD combination (Table 1) further supports the interpretation that the two methods address overlapping rather than orthogonal failure modes.

5.8 Qualitative Analysis

Per-sample analysis on POPE reveals that VIAR corrects 21 samples that the baseline misclassifies, while introducing 17 new errors—a net gain of 4. The

corrected samples are predominantly false negatives: the baseline says "no" for objects that are present (e.g., "Is there a car in the image?" with baseline $p_{\text{yes}} = 0.19$, VIAR $p_{\text{yes}} = 0.39$). The introduced errors are predominantly false positives: VIAR shifts borderline "no" cases past the decision threshold (e.g., "Is there a motorcycle in the image?" with baseline $p_{\text{yes}} = 0.25$, VIAR $p_{\text{yes}} = 0.41$). This asymmetry is consistent with the calibration finding: VIAR systematically increases the model's propensity to affirm the presence of queried objects, correcting the baseline's conservative bias at the cost of occasional false affirmations.

5.9 Addressing Potential Concerns

We address several specific questions that arise from our analysis:

How is attention fraction computed during inference? The visual attention fraction (both aggregate and decomposed) is computed from a single forward pass on the full input prompt (system tokens + visual tokens + question tokens). No generation tokens are included; the metric reflects how the model distributes attention when encoding the input, not during autoregressive generation. This ensures the measurement is deterministic and independent of generated content.

Does the U-shape persist across different tasks/prompts? Yes. Section 4.11 demonstrates that the text→vis U-shape and layer-31 anomaly are consistent across three distinct prompt types (binary yes/no, multiple choice, and open captioning), confirming the neglect zone is a model property rather than a prompt artifact.

Can you provide a text-query-only version of the metric? Yes. Section 4.9 introduces the decomposed metric (Eq. 2), which restricts attention analysis to text-query→visual-key pairs, explicitly excluding the structurally inflated visual-query→visual-key attention. The U-shape persists in this decomposition.

Do you see ECE/Brier improvements, not just yes-ratio shift? The Brier score improves from 0.123 to 0.122 (Section 4.7), a modest but directionally consistent improvement. ECE is comparable between methods (0.043 vs. 0.046). The dominant calibration effect is the yes-ratio correction, which is the most interpretable metric for a balanced binary benchmark.

Is the neglect driven by specific "neglect heads"? No. Section 4.10 presents a full 32×32 head-level analysis. Within the neglect zone (layers 8–16), the inter-head standard deviation is 0.037, compared to 0.118 globally. The neglect is a layer-level phenomenon, not a head-specific one. This justifies the uniform additive bias in Eq. 3: head-selective interventions are unnecessary.

Does the neglect zone generalize to non-LLaVA architectures? The U-shaped neglect zone generalizes to other direct-projection architectures: LLaVA-NeXT-Mistral-7B (Mistral backbone) shows a closely correlated profile (r=0.70 with LLaVA-Vicuna). However, it does not generalize to Q-Former-based (InstructBLIP) or alternative architectures (Qwen2-VL), which show qualitatively different attention profiles (Section 4.13). This specificity strengthens rather than weakens the finding: the neglect zone is tied to the direct-projection design, not an artifact of the metric.

5.10 Limitations

We acknowledge several important limitations. First, and most fundamentally, attention weights are an imperfect proxy for information flow [14,27]. High attention to visual tokens does not guarantee that visual information is faithfully propagated through the residual stream, and low attention does not preclude information transfer via other pathways. We mitigate this concern through three independent validation methodologies (attention analysis, hidden-state probing, gradient attribution) that converge on the same conclusions, but a complete causal information-flow analysis using activation patching or distributed alignment search remains future work.

Second, the accuracy improvement on POPE is not statistically significant ($p = 0.272$; Cohen's $h = 0.007$; post-hoc power 6.8%), and achieving adequate power would require an infeasibly large sample size ($\sim$325,000). The intervention should be understood as a proof-of-concept for the causal relevance of the neglect zone, not as a reliable performance enhancement.

Third, the optimal bias magnitude and target layers require calibration per model, limiting plug-and-play applicability. Fourth, we have not investigated the neglect zone during training, where it may play a different or even beneficial role (for instance, enabling language priors to develop without visual interference).

Fifth, our cross-architecture analysis now spans four models (two direct-projection, one Q-Former, one alternative), confirming that the U-shaped neglect zone is specific to direct-projection architectures. However, our coverage of direct-projection variants is limited to 7B-parameter models in the LLaVA family; whether the pattern holds for larger models (13B, 70B), other direct-projection architectures (Phi-Vision, Cambrian), or models with substantially different training procedures remains open. The Qwen2-VL metric incomparability issue (Section 4.13) highlights that our attention-based metric may not be suitable for all architectural designs, particularly those employing mrope or other non-standard positional encoding strategies for cross-modal integration.

Sixth, the per-sample correlation analysis (Section 4.16) shows that aggregate neglect depth does not predict individual hallucinations, limiting the practical utility of the neglect zone as a per-sample diagnostic. The significant correlations at boundary layers suggest a more complex story that warrants further investigation. Finally, our intervention shows no improvement on GQA (open-ended VQA), indicating that the neglect zone's behavioral relevance may be limited to constrained response formats.

6 Conclusion

We have identified and comprehensively 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 language model backbones (Vicuna and Mistral, profile correlation 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 regardless of image content), providing definitive mechanistic evidence for the previously observed attention anomaly. Gradient attribution independently validates the neglect zone through a 5× steeper-than-baseline decay of visual importance through layers 8–16. Per-sample analysis reveals that individual hallucination vulnerability manifests specifically at the neglect zone boundary (layers 14–17, $p < 0.002$), bridging model-level and sample-level perspectives.

The simple VIAR intervention demonstrates that the neglect zone is causally linked to model calibration: correcting the attention deficit shifts yes-ratio on POPE from 42.8% to exactly 50.0%, improves Brier score, and improves visually demanding categories on MMStar by up to 5.6%. Bayesian analysis confirms the calibration shift with 99.95% posterior probability while the accuracy improvement remains genuinely ambiguous (67.8% posterior probability; Cohen's h=0.007), supporting the diagnostic framing.

Rather than claiming a new state-of-the-art method, we offer this work as a structural mechanistic contribution with two implications. For model developers, the neglect zone and language prior collapse suggest that training procedures should explicitly encourage sustained visual grounding in middle layers and reconsider the role of the final transformer layer in cross-modal integration, perhaps through auxiliary losses, architectural modifications, or attention regularization. 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. We hope that understanding where, why, and how VLMs neglect visual information will inform more principled approaches to building models that truly attend to what they see.

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. Darcet, T., Oquab, M., Mairal, J., Jegou, H.: Vision transformers need registers. In: International Conference on Learning Representations (2024)
7. 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)
8. 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)
9. 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)
10. 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)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks pp. 1321–1330 (2017)
12. 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)
13. 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)
14. 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)
15. 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)

16. 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)
17. 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)
18. 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)
19. 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)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36 (2024)
21. 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)
22. 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)
23. 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)
24. 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)
25. Vig, J.: A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 37–42 (2019)
26. 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)
27. 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)
28. 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)
29. Yue, Z., Zhang, L., Jin, Q.: Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. pp. 5765–5780 (2024)
30. 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)

31. 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)