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

Anonymous ECCV 2026 Submission

Anonymous Institution

This supplementary material provides additional experimental results, ablation studies, and analysis details referenced in the main paper. We organize the content as follows: GQA results (Section A), HallusionBench results (Section B), calibration details (Section C), mechanistic verification (Section D), decomposed attention analysis (Section E), head-level analysis (Section F), cross-prompt consistency (Section G), per-sample correlation analysis (Section H), threshold-shift analysis (Section I), ablation studies (Section J), addressing potential concerns (Section K), and qualitative analysis (Section L).

A GQA Results

On GQA [2] (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.

B HallusionBench Results

Table B1 reports results on HallusionBench [1] (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.

C Calibration Details

Beyond yes-ratio, we evaluate calibration using Expected Calibration Error (ECE, 10 equal-width bins) and Brier score computed from the model's softmax prob-

Table B1: **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%

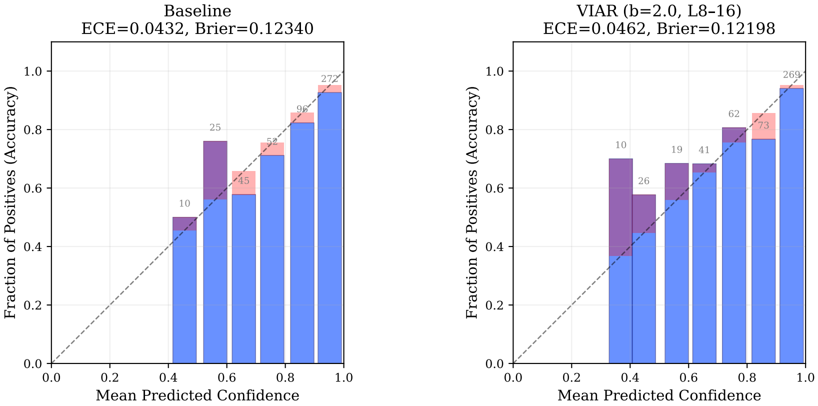


Fig. C1: **Reliability diagrams** for baseline and VIAR on POPE (500 samples). Each bar shows the fraction of positive outcomes per confidence bin; numbers above bars indicate sample counts. The dashed diagonal represents perfect calibration. VIAR shifts samples from high-confidence bins into moderate-confidence bins, reflecting increased uncertainty on borderline cases.

abilities over the yes/no tokens. 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 C1) 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.

D 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 D1).

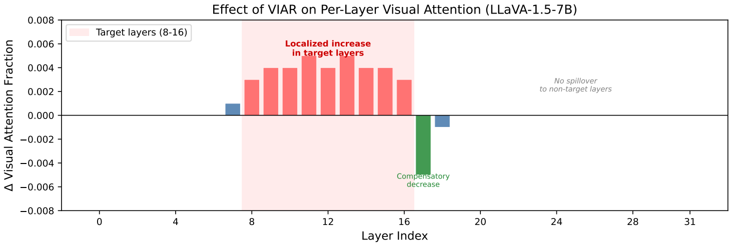


Fig. D1: **Mechanistic verification.** Change in visual attention fraction (Δvis_frac) per layer when VIAR is applied to layers 8–16. The intervention increases visual attention within the target range (+0.003 to +0.005) with no spillover to non-target layers and a slight compensatory decrease at layer 17.

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.

E Decomposed Attention Analysis

A potential concern with the aggregate *vis_frac* metric 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 across all 32 layers on 200 POPE samples.

Figure E1 (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 E1 (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.

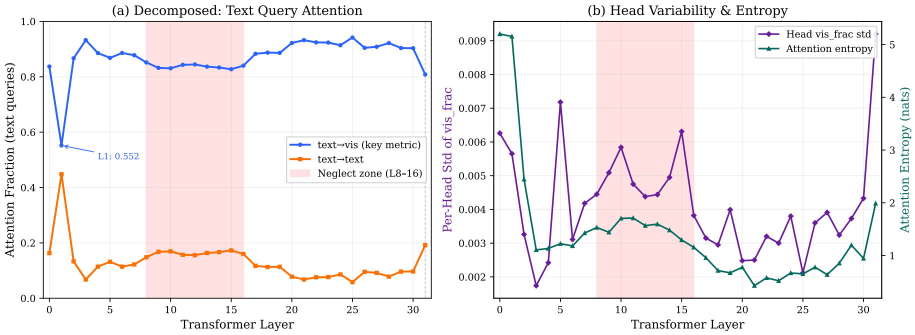


Fig. E1: **Decomposed attention analysis** (200 POPE samples). (a) The text→vis fraction confirms the U-shaped neglect zone independently of structural causal-mask effects. The shaded region marks layers 8–16. (b) Per-head variability (std of vis_frac across 32 heads) and attention entropy both peak in the neglect zone, indicating heterogeneous and diffuse attention patterns.

F 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 × head) text→vis attention matrix on 150 POPE samples (Figure F1).

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 is well-matched to the phenomenon. Head-selective interventions would be unnecessary and potentially counterproductive.

G 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

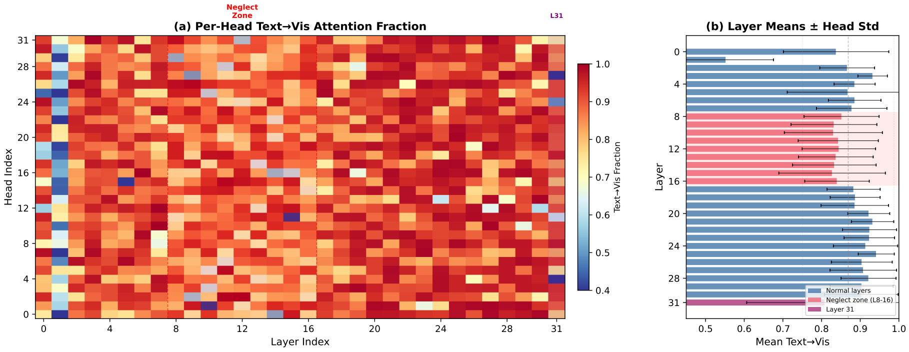


Fig. F1: **Head-level attention analysis** (150 POPE samples). (a) Full 32×32 text→vis fraction matrix. The neglect zone (dashed red, layers 8–16) shows uniformly depressed visual attention across heads, while layer 31 (purple) exhibits high inter-head variance. (b) Layer means with head-level standard deviation bars.

multiple-choice questions, and (3) open-ended captioning ("Describe this image briefly."), each evaluated on 100 samples.

Figure G1 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 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.

H Per-Sample Correlation Analysis

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.

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

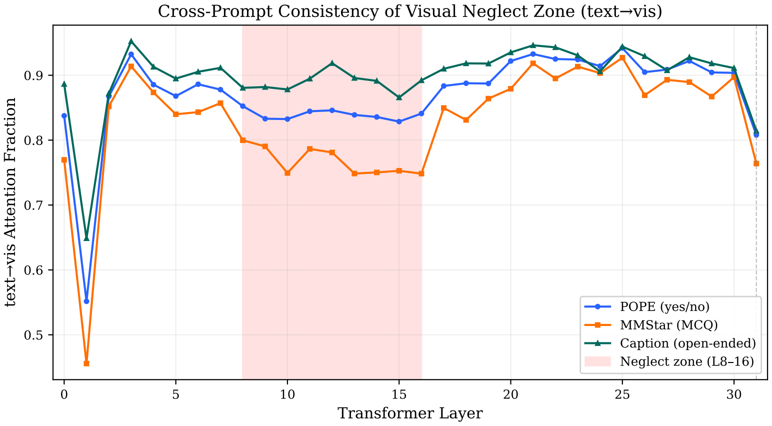


Fig. G1: **Cross-prompt consistency of the visual neglect zone.** The text→vis attention fraction follows the same U-shaped pattern across three distinct prompt types: POPE (yes/no), MMStar (multiple-choice), and open captioning. The neglect zone location (layers 8–16) and layer-31 anomaly are invariant to prompt structure.

(Figure H1). 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 neglect zone is a structural property of the model's computation, not a sample-varying phenomenon.

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 H1a): 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,

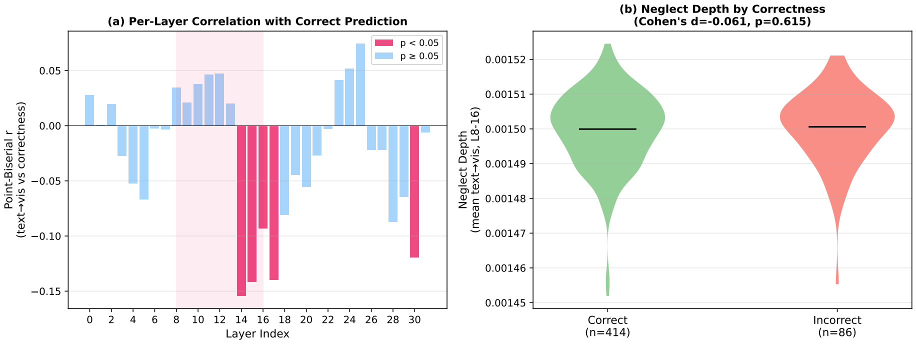


Fig. H1: **Per-sample neglect-hallucination correlation** (500 POPE samples, LLaVA-1.5-7B). (a) Per-layer Pearson correlation between text→vis attention and prediction correctness. Significant negative correlations ($*$: p<0.05, $**$: p<0.01, $* * *$: p<0.001) cluster at the neglect zone boundary (layers 14–17). (b) Violin plot of neglect-zone aggregate attention depth for correct vs. incorrect predictions (Cohen's $d = -0.061$).

but individual-sample hallucination vulnerability manifests at the zone's boundary.

I 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 I1 (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 ≈ 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 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

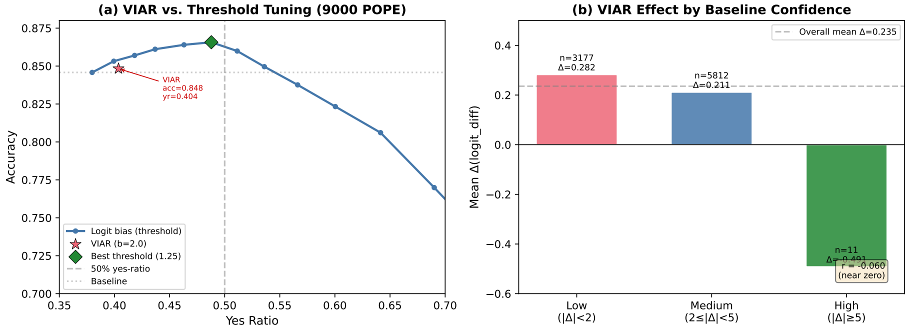


Fig. I1: **Threshold-shift analysis** (9,000 POPE samples). (a) VIAR's accuracy at its achieved yes-ratio falls on or below the logit bias baseline curve, indicating that the behavioral effect is equivalent to threshold tuning. The key diagnostic insight is that the direction of the shift is predicted by the neglect zone analysis. (b) Per-sample logit attribution by baseline confidence level.

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 I1 (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 ($r = -0.060$).

J Ablation Studies

J.1 Adaptive Scaling Alternatives

We compare five strategies for computing per-layer bias: uniform, linear (from the main paper, Eq. 4), quadratic, binary (below-median only), inverse-rank, and entropy-based. Results on a 200-sample MMStar subset are reported in Table J1. 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.

J.2 Bias Magnitude Sweep

Figure J1 (b) 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.

Table J1: **Adaptive scaling alternatives** on a 200-sample MMStar subset. Linear, quadratic, and binary perform equally; entropy-based is worst. We adopt the linear strategy for its simplicity.

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

Table J2: **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 in main paper)	83.5%	**50.5%**
Layer 31 only	82.5%	48.0%
Neglect + Layer 31	84.0%	49.0%

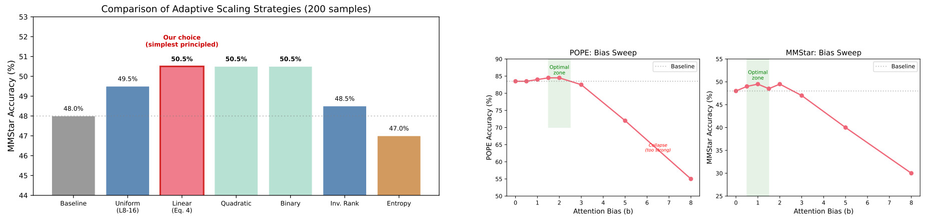
J.3 Layer Ablation

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 J2).

K 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.



(a) Adaptive scaling strategies

(b) Bias magnitude sweep

Fig. J1: **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 $b \geq 5$.

Does the U-shape persist across different tasks/prompts? Yes. Section G 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 E introduces the decomposed metric, 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 C), 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 F 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: 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 (main paper, Section 4.6). This specificity strengthens rather than weakens the finding: the neglect zone is tied to the direct-projection design, not an artifact of the metric.

L Qualitative Analysis

Per-sample analysis on POPE reveals that VIAR corrects 21 samples that the baseline misclassifies, while introducing 17 new errors, resulting in 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.

References

1. 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)
2. 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)