Targeting the Attention Heads Behind Object Hallucination in LLaVA

Armaan Sandhu[*], **Abhilasha Senapati**[*] **& Hima Kammachi**[*]
University of Massachusetts Amherst
{apsandhu, asenapati, hkammachi}@umass.edu
[*]Equal contribution.

Abstract

Vision-language models such as LLaVA-1.5-7B often hallucinate objects absent from the image when generating captions. We ask whether an interpretability diagnosis of this failure can guide a targeted fix, and we measure what that fix actually changes. We rank attention heads by how much their image attention drops around hallucinated object words, then screen the shortlist by ablating candidate heads and measuring the change in hallucination-token log probability, yielding a 32-head set. We restrict two interventions to these heads: a head-sliced LoRA adapter and an inference-time grounding controller. On 400 held-out COCO images, the combined method lowers CHAIRs (the fraction of captions with a hallucinated object) from 0.370 to 0.230 and CHAIRi (the fraction of hallucinated object mentions) from 0.156 to 0.096 ($p < 0.001$, paired sign-flip tests). Two controls sharpen attribution. A random-head LoRA control, matched layer-for-layer and trained identically, performs no better than the matched baseline on a separate 200-image control split, supporting the role of head selection rather than LoRA capacity. Under fixed decoding budgets, the CHAIR reduction persists and grows with budget (23% at 64 tokens to 58% at 128), arguing against a pure max-token or truncation artifact, although the method remains shorter and more conservative. The resulting behavior reduces unsupported object mentions while also lowering object recall (0.78 to 0.70). We present a diagnosis-to-intervention pipeline for object hallucination, and, more importantly, a controlled account of what acting on the diagnostic signal actually does: it localizes intervention sites with real, non-random leverage, reported as a behavioral profile rather than a single score.

1 Introduction

Vision-language models often produce fluent captions that mention objects not present in the image. This object hallucination failure undermines reliability in any setting where generated descriptions feed downstream decisions (Rohrbach et al., 2018). It also makes a useful test case for interpretability: the behavior can be localized to specific model components, measured against ground-truth annotations, and then used to ask a sharper question than whether an internal signal merely correlates with the failure, namely whether acting on that signal actually changes behavior.

Existing mitigation methods usually intervene either at decoding time or through broad model adaptation. Training-free decoding methods such as VCD (Leng et al., 2024) and SPIN (Sarkar et al., 2025) adjust the output distribution or suppress low-image-attention heads at inference time. Parameter-efficient fine-tuning can shift model behavior with small updates (Hu et al., 2022; Dettmers et al., 2023), but is rarely restricted to components implicated by an object-level diagnosis. We instead use an interpretability diagnosis to decide both *where* to adapt and *when* to intervene: the same ablation-screened head set determines a head-restricted LoRA adapter and an inference-time grounding controller.

We study LLaVA-1.5-7B (Liu et al., 2023; 2024a). Our pipeline ranks attention heads by how much their attention to the image drops around hallucinated object words, screens the strongest candidates by ablating them and measuring the change in hallucination-token log probability, and arrives at a 32-head set. We then apply two interventions limited to those heads: a LoRA adapter on the selected heads, and a decoding-time grounding penalty that fires when those heads attend weakly to image tokens.

The combined intervention is conservative: it produces shorter captions with lower object recall. Its CHAIR gains therefore reflect two things at once, genuine hallucination reduction and a general shift toward saying less. We measure both and report the intervention as a behavioral profile rather than a single score.

Our contributions are:

- **We introduce a diagnosis-to-intervention pipeline for object hallucination.** The pipeline identifies hallucination-linked attention heads, screens them by ablation, and restricts both a LoRA adapter and an inference-time grounding controller to the resulting 32-head set.
- **We show that head selection matters beyond LoRA capacity.** On a separate 200-image control split, a layer-matched random-head LoRA control trained with the identical recipe performs no better than the matched baseline (CHAIRs 0.400), while the selected-head adapter reduces CHAIRs to 0.100 on the same images.
- **We reduce object hallucination on held-out COCO images.** On 400 held-out images, the full LoRA + Grounding method reduces CHAIRs from 0.370 to 0.230 and CHAIRi from 0.156 to 0.096, with paired sign-flip tests showing significant reductions ($p < 0.001$).
- **We test robustness to decoding length.** With the same token budgets, our method still reduces CHAIRs, with relative gains growing from 23% at 64 tokens to 58% at 128 tokens. This rules out a simple max-token truncation explanation, though the method remains shorter and more conservative.
- **We characterize the intervention as a behavioral profile rather than a single score.** The method reduces unsupported object mentions but also shortens captions and lowers object recall, so we report hallucination, length, recall, random-head controls, SPIN/VCD comparisons, and sensitivity checks together.

2 Related Work

LLaVA-style models combine visual encoders with instruction-tuned language models and achieve strong open-ended multimodal generation (Liu et al., 2023; 2024a). However, they can inherit language-prior behavior that produces visually unsupported details. CHAIR formalizes object hallucination in image captioning by measuring whether generated object mentions appear in the image annotations (Rohrbach et al., 2018). Although CHAIR is an incomplete proxy for visual truth, it gives a concrete object-level failure signal suitable for controlled intervention studies.

Attention attribution for hallucination. A growing body of interpretability work links object hallucination to where LVLMs place attention. Jiang et al. (2025) use an attention lens to localize hallucination to the middle layers of LLaVA-style models and show that grounded object tokens receive higher visual attention than hallucinated ones. Liu et al. (2024b) find that image tokens receive little attention during generation and intervene on decoder self-attention to make inference more image-centric. Most directly related, Yang et al. (2025) use causal mediation analysis to identify "hallucination heads" in the middle and deeper layers that over-attend to text, and mitigate hallucination with both a decoding-time adjustment and a head-targeted fine-tuning method.

Our pipeline shares this diagnose-then-target strategy and the use of both a decoding and a training intervention. We differ in what we measure rather than in the targeting idea: we add a layer-matched random-head control that isolates the contribution of head *selection*

from adapter capacity, a fixed-budget analysis that separates hallucination reduction from max-token truncation effects, and an explicit account of the precision-coverage tradeoff the intervention induces. Where prior work reports mitigation, we report what acting on the diagnostic signal does as a full behavioral profile.

Targeted adaptation and training-free decoding. Parameter-efficient adaptation methods such as LoRA and QLoRA show that model behavior can often be shifted with small low-rank updates rather than full fine-tuning (Hu et al., 2022; Dettmers et al., 2023). Preference objectives such as DPO provide a simple way to train from chosen and rejected responses (Rafailov et al., 2023). Our work combines these ideas with a more localized target: only heads implicated by the hallucination diagnosis are allowed to change.

Training-free decoding methods provide complementary comparisons for the inference-time part of our pipeline. Visual Contrastive Decoding (VCD) contrasts output distributions under real and noise-corrupted images to suppress language-prior-driven tokens (Leng et al., 2024). SPIN suppresses dynamically selected low-image-attention heads at inference time (Sarkar et al., 2025), and is our closest decoding baseline since it also uses attention to choose heads. We differ from SPIN in two ways: we use a fixed, ablation-screened head set rather than re-selecting heads at each step, and we train a head-restricted adapter on that set rather than only suppressing at inference. VCD instead applies a distributional contrast between real and corrupted images, without explicitly selecting hallucination-prone heads.

3 Method

3.1 Task and Model

We evaluate open-ended image captioning with `llava-hf/llava-1.5-7b-hf`. All main experiments use COCO val2014 images and object annotations (Lin et al., 2014). The prompt is held fixed across methods: "Describe this image in detail." We report CHAIRs, the fraction of captions with at least one hallucinated object, and CHAIRi, the fraction of object mentions that are hallucinated.

Our pipeline has three sequential stages: head selection by diagnosis (Section 3.2), a LoRA adapter trained on the selected heads, and a grounding controller applied on top of the adapter over the same head set. We evaluate the LoRA stage and the full *LoRA + Grounding* method against the greedy baseline, and we also report a *Grounding only* condition (the controller applied to the base model, without LoRA) as a reference that isolates the controller's effect; it is not part of the pipeline.

3.2 Object-Level Diagnosis

We generate captions on a screening set of 200 COCO val2014 images and label each object mention as grounded or hallucinated with a COCO-derived CHAIR scorer. For every object token we record per-head attention to image tokens across the 1024 language-model attention heads of LLaVA-1.5-7B (32 layers, 32 heads). Heads are ranked by how much their image attention decreases around hallucinated object words relative to grounded object words, producing a fast diagnostic shortlist.

We then ablation-screen the strongest candidates: we ablate heads one at a time and measure the change in hallucination-token log probability. To avoid selecting on the same signal twice, the screen runs over both shortlisted and non-shortlisted heads, so a candidate can be rejected if its ablation effect is small. We retain the 32 heads with the largest ablation effect, spanning 19 transformer layers. This set is fixed and used by both later stages.

3.3 LoRA Adaptation

The LoRA stage updates only the layers and head slices associated with the 32 selected heads. LoRA modules are attached to the Q/K/V projections in layers containing selected heads, and gradient masking restricts learning to the selected head dimensions. Training

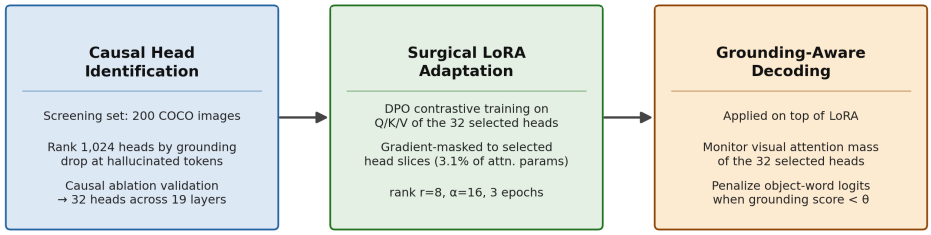

Figure 1: Diagnosis-to-intervention pipeline. Head selection identifies hallucination-linked heads, the LoRA stage adapts only those head slices, and the grounding controller is applied on top over the same head set.

examples are contrastive caption pairs: the chosen response is a COCO ground-truth caption and the rejected response is the model's own hallucinated caption. We optimize a DPO-style objective (Rafailov et al., 2023) with rank $r = 8$, $\alpha = 16$, dropout 0.05, for three epochs. Because the chosen responses are COCO captions, which are shorter than the model's free-running output, this objective also biases the adapter toward brevity; we analyze the resulting length effect in Section 4.5.

3.4 Inference-Time Grounding

The grounding controller complements the LoRA adapter rather than operating as a standalone intervention. It applies a conservative guard at inference time using the same head set that guides adaptation.

At each generation step it computes a grounding score from the visual attention mass assigned by the selected heads:

$$g_t = \text{mean}_{h \in H} \frac{\sum_{i \in \text{img}} \alpha_{h,i}}{\sum_j \alpha_{h,j}}.$$

Here H is the ablation-screened 32-head set, i indexes image token positions, j indexes all attended positions at the current step (image, prompt, and previously generated tokens), and $\alpha_{h,i}$ is the attention weight from the current step to position i under head h. Thus g_t is the fraction of head-H attention mass placed on image tokens.

When g_t falls below a threshold θ, object-word logits from a COCO-derived object vocabulary receive a penalty:

$$\ell_v \leftarrow \ell_v + \lambda \, \min(0, \, g_t - \theta) \quad \text{for object token } v,$$

with $\lambda > 0$, so the penalty activates only when head-H image attention is below threshold and is inactive otherwise. Main results use $\theta = 0.08$ and $\lambda = 10$. The controller is deliberately lightweight: it does not prove grounding at each step, but operationalizes the diagnostic signal as a conservative guard against object mentions when the selected heads attend weakly to the image. The full method applies this controller on top of the LoRA adapter, over the same head set.

4 Experiments

4.1 Protocol

The main evaluation uses 400 held-out COCO val2014 images and four conditions: the greedy LLaVA baseline, LoRA, Grounding only, and the full LoRA + Grounding method. We report bootstrap 95% confidence intervals over images and paired sign-flip permutation tests against the baseline on the same images.

We also compare all conditions against SPIN and VCD under fixed decoding budgets of 64, 80, and 128 new tokens on the same 400 images. Regenerating every method under the same

Table 1: Main 400-image held-out comparison with bootstrap 95% confidence intervals. Lower CHAIRs and CHAIRi are better.

Method	CHAIRs	CHAIRi	Avg. len.
Baseline	0.3700 [0.3225, 0.4175]	0.1558 [0.1327, 0.1800]	61.1
LoRA	0.2650 [0.2225, 0.3075]	0.1043 [0.0857, 0.1245]	27.6
Grounding only	0.3100 [0.2650, 0.3575]	0.1407 [0.1174, 0.1654]	62.1
LoRA + Grounding (Ours)	**0.2300 [0.1900, 0.2700]**	**0.0958 [0.0761, 0.1163]**	29.6

Table 2: Paired deltas versus the baseline on the same 400 held-out images. Negative deltas are better. P-values use paired sign-flip permutation tests with 10000 resamples.

Method	Δ CHAIRs	Δ CHAIRi
LoRA	-0.1050 [-0.1575, -0.0550], $p = 0.0002$	-0.0515 [-0.0738, -0.0295], $p = 0.0001$
Grounding only	-0.0600 [-0.1125, -0.0075], $p = 0.0287$	-0.0151 [-0.0386, +0.0082], $p = 0.2096$
LoRA + Grounding (Ours)	**-0.1400 [-0.1900, -0.0900],** $p = 0.0001$	**-0.0601 [-0.0837, -0.0369],** $p = 0.0001$

budget controls for the strong effect of caption length on CHAIR. VCD uses its published hyperparameters ($\alpha = 1.0$, $\beta = 0.1$, noise step 500), implemented as a custom autoregressive loop on the same HF model for identical tokenization and prompt handling (Leng et al., 2024). All fixed-budget results appear in one table (Table 3).

Finally, we run a head-selection control on a 200-image held-out split. It uses the same DPO pairs, LoRA configuration, and training recipe as our LoRA stage, but replaces the ablation-screened 32 heads with a random set matched to the same per-layer head counts, excluding the selected heads. We compare CHAIR on the same images.

4.2 Main Results

Table 1 shows that LoRA alone reduces hallucination significantly, and the grounding controller reduces it further when applied on top; the full method performs best. Grounding alone does not produce a significant CHAIRi reduction (Table 2, $p = 0.2096$), consistent with its design as a modifier rather than a standalone intervention. The full method reduces CHAIRs by 37.8% and CHAIRi by 38.5% relative to the baseline. This headline comparison is not length-matched, since the method also shortens captions substantially; we therefore read it as a conservative intervention that changes the targeted failure mode, and defer broader caption-quality claims to length-matched and human evaluations.

We next present the head-selection control, the most direct test of whether the diagnostic signal, rather than LoRA capacity, drives the change.

4.3 Random-Head Control

On the 200-image control split, the layer-matched random-head LoRA control is indistinguishable from the baseline (CHAIRs 0.400, CHAIRi 0.131), while the selected-head adapter reduces CHAIRs to 0.100 and CHAIRi to 0.054 on the same images. Replacing the selected heads with a layer-matched random set does not reproduce the adapter's behavioral change, so the diagnostic signal identifies intervention sites with distinct leverage. This control does not by itself establish improved grounding, since the adapter's reduction remains subject to the length and conservatism effects we examine next.

4.4 Tradeoff: Fewer Hallucinations, Shorter Captions

The full method is substantially more conservative than the baseline: average caption length drops from 61.1 to 29.6 words and object recall from 0.78 to 0.70 (Figure 2). This matters because CHAIR can be reduced by mentioning fewer objects. Our reading is therefore conditional: the method reduces object hallucination under CHAIR but does not uniformly improve caption quality. The fixed-budget analysis below addresses the strongest form of

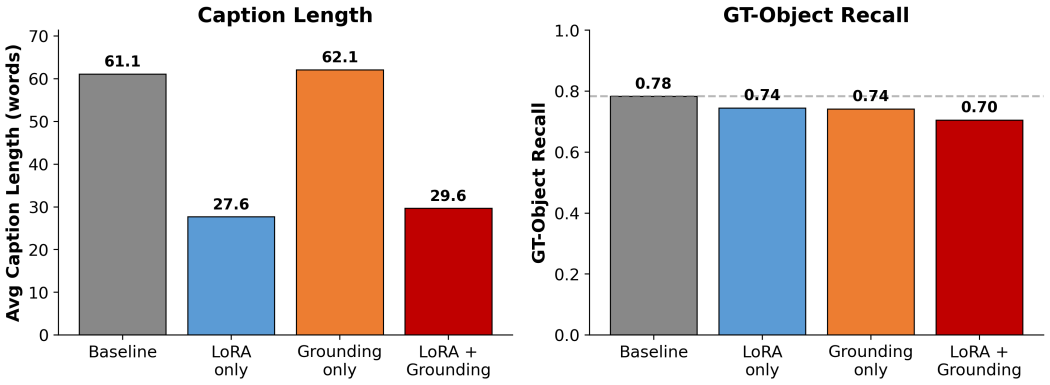

Figure 2: Hallucination reduction is accompanied by shorter captions and lower object recall. This tradeoff is central to evaluating actionability.

this concern; a fully length-matched evaluation with human quality assessment remains the most important missing control.

4.5 Fixed-Budget Comparison and Robustness

Table 3 reports all methods at fixed decoding budgets of 64, 80, and 128 new tokens on the same 400 images, with bootstrap 95% confidence intervals for Baseline, SPIN, and VCD. It serves two purposes: comparison against training-free baselines under a controlled length budget, and a test of whether our CHAIR reduction is a shorter-caption artifact.

Comparison to SPIN and VCD. Neither training-free baseline consistently reduces hallucination: SPIN is numerically worse than its matched baseline at 64 and 80 tokens, and VCD increases CHAIRi at all three budgets. At 128 tokens SPIN improves CHAIRs but leaves CHAIRi essentially unchanged. We read these as protocol-specific comparisons, not general claims about SPIN or VCD.

Robustness to caption length. Within a fixed budget, both methods share the same maximum length, removing the free-running length confound. At a 64-token budget that forces baseline truncation (baseline length 61.1 to 49.5 words), our method still reduces CHAIRs from 0.280 to 0.215, a 23.2% relative reduction. The gain therefore is not purely a max-token or truncation artifact, though the method still generates shorter captions within the same budget. The gap also *grows* with budget: baseline CHAIRs rises from 0.280 to 0.512 as the budget triples, while our method stays near 0.215 throughout, so the relative reduction grows from 23% at 64 tokens to 58% at 128 (Figure 3). We read this as the baseline becoming progressively miscalibrated to the image as more tokens must be filled, a drift the targeted intervention does not show. This does not achieve full length matching, since our method still generates shorter captions within each budget (27.9 vs. 49.5 words at 64 tokens); what it rules out is the strong claim that the entire reduction comes from generating less.

4.6 Sensitivity Checks

Figure 4 shows an adapter-scale sweep on 200 held-out images: the grounding controller improves CHAIRs at every LoRA scale, including the no-adapter and full-adapter settings, so it contributes beyond the LoRA update even though its standalone 400-image CHAIRi reduction is not significant. The controller is also insensitive to threshold over $\theta \in \{0.04, 0.06, 0.08, 0.10, 0.12\}$ (CHAIRs stays in 0.24–0.26), and top-16 versus top-32 heads gives nearly identical CHAIRs (0.24) and CHAIRi (0.099 vs. 0.098). With the random-head control (Section 4.3), these checks support the head-selection story but do not substitute for length-matched 400-image controls.

Table 3: All methods at fixed decoding budgets on the same 400 held-out images. Baseline, SPIN, and VCD include bootstrap 95% confidence intervals; our conditions report point estimates, with bootstrap intervals for these fixed-budget runs left to appendix evaluation. Lower CHAIRs and CHAIRi are better. LoRA + Grounding (Ours) outperforms both training-free baselines at every budget; the relative reduction over the matched baseline grows from 23% at 64 tokens to 58% at 128 tokens.

Budget	Method	CHAIRs	CHAIRi	Avg. len.
64	Baseline	0.2800 [0.2375, 0.3225]	0.1066 [0.0863, 0.1266]	49.5
64	SPIN	0.3025 [0.2575, 0.3475]	0.1216 [0.1000, 0.1445]	48.8
64	VCD	0.2900 [0.2475, 0.3350]	0.1301 [0.1069, 0.1559]	48.8
64	LoRA	0.2225	0.0940	28.4
64	Grounding only	0.2675	0.1053	49.3
64	LoRA + Grounding (Ours)	**0.2150**	**0.0909**	27.9
80	Baseline	0.3650 [0.3175, 0.4125]	0.1382 [0.1155, 0.1601]	60.8
80	SPIN	0.3675 [0.3200, 0.4150]	0.1419 [0.1194, 0.1657]	58.4
80	VCD	0.3850 [0.3375, 0.4325]	0.1479 [0.1255, 0.1719]	60.4
80	LoRA	0.2650	0.1043	27.6
80	Grounding only	0.3100	0.1407	62.1
80	LoRA + Grounding (Ours)	**0.2300**	**0.0958**	29.6
128	Baseline	0.5125 [0.4650, 0.5600]	0.1794 [0.1565, 0.2040]	84.2
128	SPIN	0.4750 [0.4250, 0.5225]	0.1786 [0.1551, 0.2033]	78.2
128	VCD	0.5100 [0.4625, 0.5575]	0.1951 [0.1709, 0.2214]	84.0
128	LoRA	0.2225	0.0947	45.0
128	Grounding only	0.4925	0.1734	83.9
128	LoRA + Grounding (Ours)	**0.2150**	**0.0917**	43.7

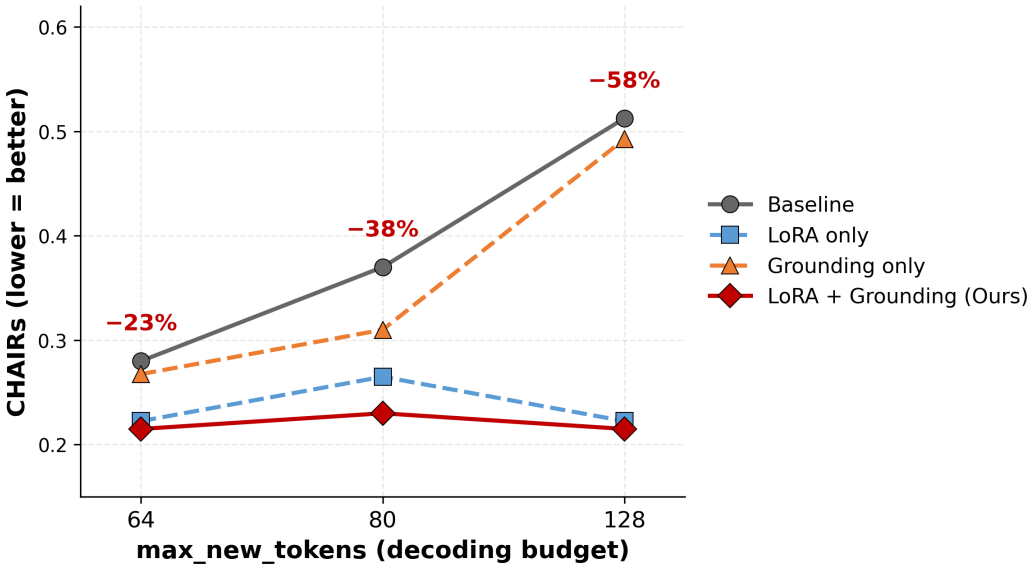

Figure 3: Caption-level CHAIR versus decoding budget on 400 held-out COCO images. Baseline CHAIRs rises sharply (0.28 → 0.37 → 0.51) as the budget grows, while LoRA + Grounding (Ours) stays near 0.22 at all budgets. The relative reduction grows from 23% at 64 tokens to 58% at 128, indicating the baseline becomes progressively miscalibrated to the image under larger budgets while the targeted intervention does not.

4.7 Qualitative Example

Figure 5 shows a held-out kitchen example. The baseline correctly mentions the refrigerator but hallucinates an unsupported oven, and the grounding-only method repeats the same hallucination. LoRA-only removes the oven but introduces another unsupported object mention. The combined LoRA + grounding method retains grounded objects such as the

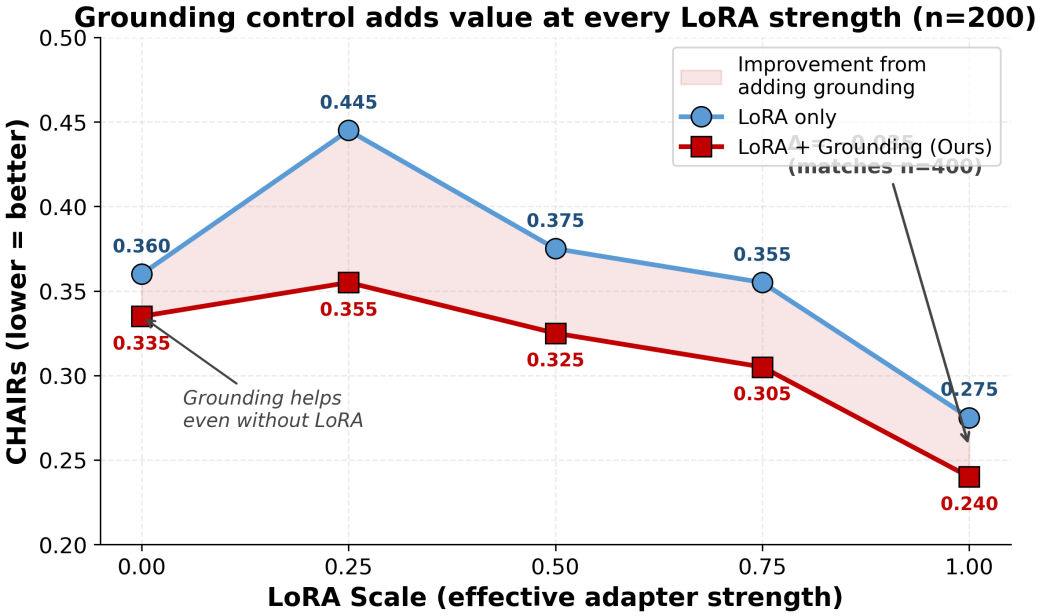

Figure 4: Adapter-scale sweep on 200 held-out images. Grounding control improves CHAIRs across LoRA strengths, but does not resolve the separate length-matching and head-selection questions.

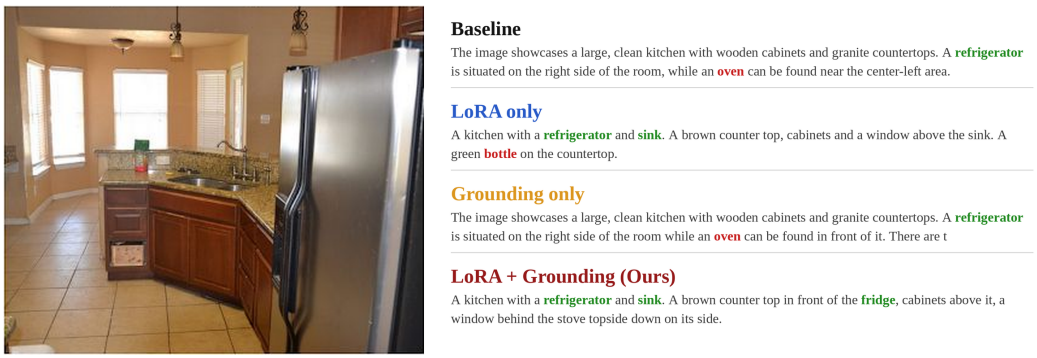

Figure 5: Qualitative comparison on a held-out kitchen image. Green object mentions are supported by COCO annotations; red mentions are counted as hallucinated by CHAIR. The combined LoRA + grounding method suppresses the oven hallucination while retaining grounded objects, but yields a shorter, more conservative caption.

refrigerator and sink while suppressing the oven hallucination, but produces a shorter and less fluent caption. This mirrors the aggregate pattern: fewer unsupported object mentions, with a tradeoff toward more conservative captions.

5 Discussion

Object-level diagnosis can be converted into an actionable intervention. The ablation-screened 32-head set supports both a targeted adapter and an inference-time control signal, and the full method gives the strongest paired reduction in hallucination. The random-head control (Section 4.3) suggests that this leverage depends on the selected heads rather than LoRA capacity alone, while the LoRA-scale sweep shows that the grounding controller adds value across adapter strengths.

The main tradeoff is that hallucination reduction comes with more conservative generation. CHAIR measures an important object-level failure mode, but reducing unsupported

mentions can also reduce informativeness and object coverage. The fixed-budget analysis (Section 4.5) addresses the strongest version of the length-artifact concern: the CHAIR reduction persists when methods share the same token budgets and grows as more tokens are allowed. At the same time, the method still produces shorter captions within those budgets. We therefore interpret the result as targeted hallucination reduction, not as a uniform improvement in caption quality.

6 Limitations and Future Work

This study focuses on one model family, one captioning prompt, and COCO-style object annotations. CHAIR provides a useful controlled signal, but it depends on annotation coverage and synonym matching, so valid but unannotated objects may be counted as hallucinations. The grounding score also uses attention as a lightweight proxy for visual grounding, even though attention weights are not guaranteed to be faithful explanations of model behavior (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), and the selected heads are screened by ablation.

Several follow-up controls would sharpen the attribution. A fully length-matched comparison could force the baseline to match the average output length of our method, separating reduced hallucination from reduced object coverage more directly. An all-head LoRA baseline and a plain DPO baseline at matched length would test whether head restriction gives advantages beyond targeted capacity control. Human evaluation would also help measure fluency, informativeness, and faithfulness beyond CHAIR. Finally, extending the pipeline to additional VLMs, prompts, and datasets would test whether hallucination-linked attention heads provide stable intervention targets beyond LLaVA on COCO captions.

7 Conclusion

We presented a compact diagnosis-to-intervention pipeline for object hallucination in LLaVA captions: identify hallucination-linked attention heads, adapt only those head slices, and add a grounding-aware decoding controller over the same set. The full method significantly reduces CHAIRs and CHAIRi on 400 held-out COCO images (both $p < 0.001$), while a layer-matched random-head control supports the relevance of the selected heads rather than LoRA capacity alone. The budget analysis further shows that the reduction persists under fixed decoding budgets and grows as more tokens are generated.

The resulting captions are shorter and more conservative, so the gains are best understood as targeted reductions in object hallucination rather than as a general improvement in caption quality. Still, the result is encouraging for mechanistic interpretability in vision-language models: an internal diagnostic signal can identify intervention sites with measurable behavioral leverage. More broadly, object hallucination offers a useful testbed for moving from interpretability as post-hoc explanation toward interpretability as a guide for localized model editing and controlled behavioral change.

References

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.14314.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3543–3556, 2019. URL https://aclanthology.org/N19-1357/.

Z. Jiang, J. Chen, B. Zhu, T. Luo, Y. Shen, and X. Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25004–25014, 2025. URL https://arxiv.org/abs/2411.16724.

S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.16922.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2304.08485.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024a. URL https://arxiv.org/abs/2310.03744.

S. Liu, K. Zheng, and W. Chen. Paying more attention to image: A training-free method for alleviating hallucination in large vision-language models. *arXiv preprint arXiv:2407.21771*, 2024b. URL https://arxiv.org/abs/2407.21771.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2305.18290.

A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045, 2018. URL https://arxiv.org/abs/1809.02156.

S. Sarkar, Y. Che, A. Gavin, P. A. Beerel, and S. Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12481–12500, 2025. URL https://arxiv.org/abs/2505.16411.

S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019. URL https://aclanthology.org/D19-1002/.

T. Yang, Z. Li, J. Cao, and C. Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Bjq4W7P2Us.