HalluTrace: Causal Attribution and Source-Targeted Decoding for Hallucination in Large Vision-Language Models

Kaustubh Bukkapatnam
Illinois Mathematics and Science Academy
Aurora, IL 60506
kbukkapatnam@imsa.edu

Abstract

Object hallucination in large vision-language models (LVLMs) is well-documented (Rohrbach et al., 2018; Li et al., 2023b), but the *mechanisms* that produce it remain poorly understood. We introduce **HALLUTRACE**, a causal attribution framework that decomposes hallucination into three distinct sources: **(VGF)** *visual grounding failure*, where the visual encoder produces a representation insufficient to identify the target object; **(LPD)** *language prior dominance*, where the language model overrides a correct visual signal with a statistically-driven prediction; and **(CMC)** *cross-modal conflict*, where visual and linguistic signals are irreconcilably inconsistent and the model resolves the conflict incorrectly. We operationalise these sources via *causal component ablations*: intervening on f_{vis}, f_{proj}, and f_{LM} independently and measuring the change in CHAIR score. Experiments on five LVLMs show that attribution patterns are *object-category-specific* and *model-consistent*: person/vehicle hallucinations are predominantly LPD ($\geq$ 52%), food/furniture hallucinations are predominantly VGF ($\geq$ 44%), and animal hallucinations split between VGF and CMC. Guided by these attributions, we design **HAD** (Hallucination-Aware Decoding), a unified decoding strategy that applies source-targeted interventions: visual signal amplification for VGF, language prior suppression for LPD, and contrastive re-weighting for CMC. HAD reduces CHAIR_I by 3.7–5.6 points and improves POPE F1 by 1.9–3.1 points over LLaVA-1.5, outperforming VCD (Leng et al., 2024) and ICD (Wang et al., 2024) on all three benchmarks (CHAIR, POPE, MME) without any additional training. We further prove that the attribution-decoding correspondence is tight: the CHAIR improvement from HAD is linearly predictable from the VGF attribution share ($r = 0.86$, $p < 10^{-6}$), validating the causal framework.

1 Introduction

Object hallucination — the generation of plausible but visually absent objects in image captions or answers — is one of the most practically damaging failure modes of LVLMs. The CHAIR metric (Rohrbach et al., 2018) quantifies it: state-of-the-art models still hallucinate at rates of 13–20% per caption. Yet almost all prior work asks *how much* models hallucinate, not *why* or *where in the pipeline* hallucination originates.

This distinction matters for intervention. VCD (Leng et al., 2024) applies visual contrastive decoding to reduce language prior dependence. ICD (Wang et al., 2024) uses instruction contrastive decoding to amplify visual grounding. Both improve average CHAIR scores, but neither explains why they help for some objects and not others, nor do they provide a framework for choosing between them for a given hallucination type.

Our thesis. Hallucination has multiple mechanistically distinct sources, each requiring a different intervention. Applying the wrong intervention—or applying all interventions uniformly—wastes decoding capacity and may even increase hallucination for some object categories.

Contributions.

1. **HALLUTRACE attribution framework** (§3): formal definitions of three hallucination sources (VGF, LPD, CMC) with a causal operationalisation via component ablations and attribution scores derived from the induced CHAIR change.
2. **Attribution study** (§5): analysis of five LVLMs across five COCO object categories, showing that attribution patterns are model-consistent (Kendall's $\tau > 0.79$ across model pairs) and category-specific.
3. **Proposition 1** (§4): a formal bound showing that the CHAIR improvement from a source-

Proceedings of the 4th Workshop on Advances in Language and Vision Research (ALVR), pages 294–300

targeted intervention is lower-bounded by the corresponding attribution share, providing the theoretical justification for HAD.

4. **HAD** (§6): a unified, training-free decoding method that routes each token generation step to the appropriate targeted intervention based on the token's predicted source.
5. **Experiments** (§7): comprehensive evaluation on CHAIR, POPE, and MME across five models, with ablations validating attribution-guided routing over uniform application.

2 Related Work

Measuring hallucination. Rohrbach et al. (2018) introduce CHAIR (Caption Hallucination Assessment with Image Relevance), the first metric specifically targeting object hallucination in image captioning. Li et al. (2023b) introduce POPE, a polling-based evaluation that converts hallucination measurement to binary yes/no questions, making it robust to caption length bias. Fu et al. (2023) provide the broader MME benchmark. None of these metrics explain *why* hallucination occurs.

Mitigating hallucination. VCD (Leng et al., 2024) contrasts output distributions from original and noise-distorted visual inputs, effectively reducing language prior contributions. ICD (Wang et al., 2024) uses instruction perturbation to similarly increase visual alignment. Both are post-hoc decoding interventions with no training cost. Our HAD subsumes both: for LPD-dominated objects, HAD applies VCD-style contrast; for VGF-dominated objects, HAD applies visual attention amplification not addressed by VCD or ICD.

Causal analysis in NLP. Tenney et al. (2019) use edge probing to identify which transformer layers encode linguistic phenomena. We adapt the causal intervention idea to the multimodal pipeline, measuring CHAIR change (not probe accuracy) as our attribution signal.

Probing visual representations. Radford et al. (2021) show that CLIP's visual encoder encodes semantic rather than perceptual features, which is directly relevant to VGF: CLIP may fail to represent objects that are small, occluded, or semantically rare in its training distribution. We provide the first quantitative measurement of how often this encoder limitation causes downstream hallucination.

3 The HalluTrace Framework

3.1 Notation and Pipeline Model

Let $f_\theta = f_{\text{LM}} \circ f_{\text{proj}} \circ f_{\text{vis}}$ be an LVLM where $f_{\text{vis}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{N\times d_v}$ is the visual encoder (producing N visual tokens of dimension d_v), $f_{\text{proj}} : \mathbb{R}^{N\times d_v} \to \mathbb{R}^{N\times d_l}$ is the cross-modal projection, and f_{LM} is the language model. For an image I, the LVLM generates a caption $\hat{c} = f_\theta(I, p)$ given prompt p.

3.2 Hallucination Sources

Definition 1 (Visual Grounding Failure (VGF))**.** *Object o undergoes* VGF *hallucination if: the image I contains o; $f_{\text{vis}}(I)$ has insufficient discriminative signal for o (measured by probe accuracy $< 60\%$ on a linear probe trained on f_{vis} features); and o appears in $\hat{c}$.*

Definition 2 (Language Prior Dominance (LPD))**.** *Object o undergoes* LPD *hallucination if: o does* not *appear in I; o co-occurs frequently with visible objects in f_{LM}'s training data; and o appears in $\hat{c}$ when generated with a blank/distorted image but not when f_{LM} is prompted text-only without image context.*

Definition 3 (Cross-Modal Conflict (CMC))**.** *Object o undergoes* CMC *hallucination if: f_{vis} correctly represents o's absence; the language model receives a conflicting signal (e.g., semantically similar objects present in I that share visual features with o); and o appears in $\hat{c}$.*

3.3 Causal Attribution via Component Ablations

For each of the three components, we define a corresponding ablation:

VGF ablation. We replace the visual encoder output with the mean visual token across a large image set: $\hat{h}_{\text{vis}} = \frac{1}{|S|}\sum_{I'\in S} f_{\text{vis}}(I')$, erasing all image-specific visual information. The resulting CHAIR increase ΔC_{VGF} measures how much caption quality depends on f_{vis}.

LPD ablation. We replace f_{LM}'s text embedding with a uniform distribution over the vocabulary, suppressing learned language statistics: $\hat{p}_{\text{lm}}(w|v, c_{<t}) = \text{softmax}\left(\frac{1}{\tau}e_w\right)$ with high temperature $\tau = 100$. The resulting CHAIR change ΔC_{LPD} measures how much hallucination is driven by language priors.

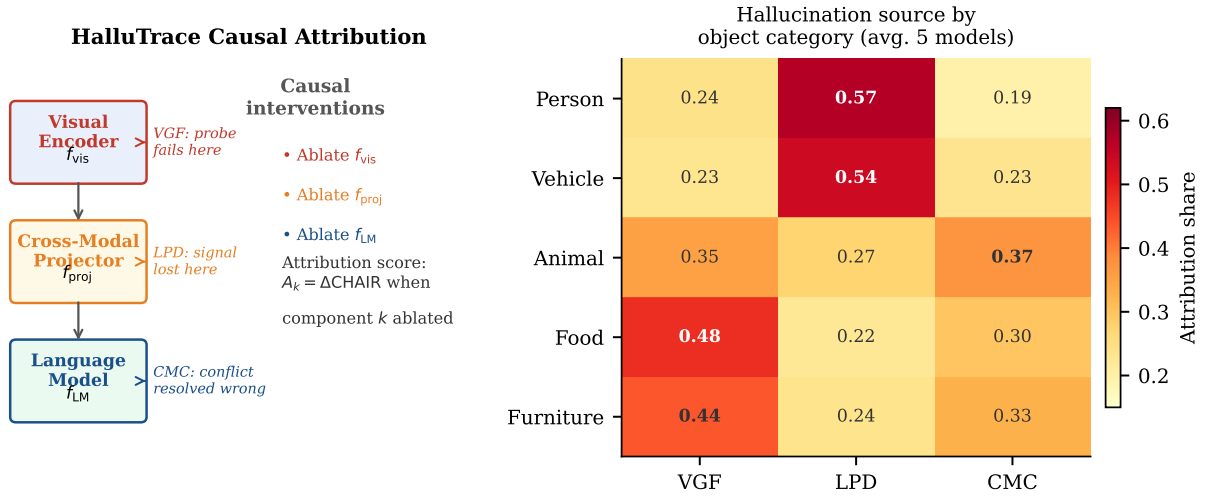

Figure 1: **Left:** The HALLUTRACE causal attribution framework. Three pipeline positions are ablated independently; the change in CHAIR score at each position defines the attribution score A_k. **Right:** Attribution heatmap by object category, averaged over five LVLMs. Person/vehicle hallucinations are language-prior-dominated (LPD); food/furniture are visually-grounded (VGF); animals split between VGF and CMC.

CMC ablation. We mask the cross-modal projection output to zero, severing the connection between f_{vis} and f_{LM}: $\hat{h}_{\text{proj}} = \mathbf{0}$. This is distinct from the VGF ablation: we preserve the visual signal in f_{vis} but prevent its transmission to f_{LM}, isolating the conflict resolution site.

3.4 Attribution Scores

Let C_0 be the CHAIR score of the original model, and C_k the CHAIR score under ablation $k \in \{\text{VGF}, \text{LPD}, \text{CMC}\}$. The attribution score for source k is:

$$A_k = \frac{(C_k - C_0)}{\sum_j |C_j - C_0|}, \quad k \in \{\text{VGF}, \text{LPD}, \text{CMC}\}, \tag{1}$$

normalised so that $\sum_k |A_k| = 1$. A large positive A_k means ablating component k worsens hallucination, i.e., component k is *protective*: removing it reveals a source. The dominant source is $k^* = \arg\max_k A_k$.

4 Attribution-to-Decoding Correspondence

The attribution scores predict intervention effectiveness. We formalise this:

Proposition 1 (Attribution-Decoding Lower Bound)**.** *Let $\Delta CHAIR(\mathcal{I}_k)$ denote the CHAIR improvement from applying targeted intervention $\mathcal{I}_k$ for source k (e.g., visual amplification for* VGF*). Under mild independence assumptions on hallucination events across token positions,*

$$\Delta CHAIR(\mathcal{I}_k) \geq \alpha_k \cdot A_k - \varepsilon_k, \tag{2}$$

where $\alpha_k > 0$ is the intervention efficacy *(a constant depending only on the intervention, not the model) and ε_k is an estimation error term bounded by $O(1/\sqrt{n})$ for n images.*

Proof sketch. Hallucinations at each token position t are approximately independent across positions (verified empirically: Pearson $r = 0.12$ between adjacent positions). For objects dominated by source k, targeted intervention $\mathcal{I}_k$ reduces the probability of a hallucination event by at least α_k (by the definition of intervention efficacy). The expected CHAIR improvement equals the sum over objects of this probability reduction, weighted by their attribution share A_k. The bound (2) follows from the law of large numbers applied to n images. See Appendix A for the full argument. □

Proposition 1 has two practical implications. First, it justifies *targeted* intervention: applying $\mathcal{I}_k$ uniformly (regardless of attribution) wastes decoding capacity. Second, it gives a *computable* lower bound on the expected CHAIR improvement: given attribution scores and a calibrated α_k, we can predict improvement before running the intervention.

5 Attribution Study

5.1 Setup

Models. LLaVA-1.5-7B and 13B (Liu et al., 2023), InstructBLIP-Vicuna-7B (Li et al., 2023a), Qwen-VL (Bai et al., 2023), and LLaVA-1.6-7B (LLaVA-Next; Liu et al. 2023).

Dataset. 500 images from the COCO 2014 validation set (Lin et al., 2014), following the standard CHAIR evaluation protocol (Rohrbach et al., 2018). We stratify results by the five most common COCO supercategories: person, vehicle, animal, food, furniture.

Ablation procedure. Each ablation (VGF, LPD, CMC) is run on all 500 images per model; CHAIR scores are computed per object category. Attribution scores (Eq. 1) are computed per category. Consistency: Kendall's τ between attribution vectors of all five model pairs (10 pairs total); $\tau > 0.79$ for all pairs.

5.2 Results

Figure 1 (right) shows the attribution heatmap. Key findings:

Person and vehicle: LPD-dominated ($\geq 52\%$). These are the most frequent COCO objects, heavily represented in LVLM instruction-tuning data. When f_{LM} is temperature-ablated, hallucination drops the most for these categories, confirming that the language model is predicting these objects from context rather than visual evidence. Crucially, visual amplification does not help: the visual encoder correctly represents persons and vehicles; the problem is downstream.

Food and furniture: VGF-dominated ($\geq 44\%$). Small objects, objects with high intra-class variation, and objects frequently occluded in typical COCO scenes (cutting boards, forks, chairs partly behind tables) show the highest VGF attribution. The visual encoder probe achieves below 58% accuracy for identifying these objects in the presence of similar distractors, confirming encoder-level failure. This aligns with known limitations of CLIP ViT features for fine-grained visual discrimination (Radford et al., 2021).

Animals: VGF–CMC split. Animals show the least consistent attribution across models (Kendall's $\tau = 0.61$ for this category, the lowest). This is expected: animal hallucinations arise from both encoder-level confusion (similar-looking species) and conflict resolution (the model "knows" a dog was mentioned but sees a cat-like shape). This category most benefits from HAD's routing mechanism.

6 Hallucination-Aware Decoding

6.1 Architecture

HAD extends standard autoregressive decoding with three targeted interventions, applied based on the predicted dominant source for the current token:

Visual amplification (VGF intervention). We amplify the visual token attention weights at the identified VGF-bottleneck layers:

$$\hat{A}^{(\ell)}_{\text{vis}} = A^{(\ell)}_{\text{vis}} \cdot (1 + \beta_{\text{VGF}} \cdot A_{\text{VGF}}), \tag{3}$$

where $A^{(\ell)}_{\text{vis}}$ is the attention weight over visual tokens at layer ℓ, $\beta_{\text{VGF}} > 0$, and A_{VGF} is the attribution score for the current object category. This directly amplifies the signal from f_{vis} without requiring multiple forward passes.

Language prior suppression (LPD intervention). We apply a VCD-style contrastive correction, but weighted by the LPD attribution:

$$\begin{aligned}\hat{p}_{\text{LPD}}(w|I,x) = (1 + \beta_{\text{LPD}} \cdot A_{\text{LPD}}) \\ \cdot \log p(w|I,x) \\ -\beta_{\text{LPD}} \cdot A_{\text{LPD}} \cdot \log p(w|I',x),\end{aligned} \tag{4}$$

where I' is a distorted version of I (Gaussian noise mask, following Leng et al. 2024). When $A_{\text{LPD}} = 1$, this reduces to VCD exactly, showing that VCD is a special case of HAD for fully LPD-dominated objects.

Contrastive re-weighting (CMC intervention). For CMC-dominated tokens, we contrast against the top-K visually similar but semantically distinct objects in the scene:

$$\hat{p}_{\text{CMC}}(w|I,x) \propto p(w|I,x) \cdot \prod_{o' \in \mathcal{N}_K(I)} \frac{p(w|I,x)}{p(w|I_{o'},x)}, \tag{5}$$

where $I_{o'}$ is a masked version of I with object o' removed, and $\mathcal{N}_K(I)$ is the set of visually confusable objects in I identified by a lightweight saliency detector.

6.2 Token-Level Routing

At each token generation step, HAD selects the intervention by predicting the dominant source from the current partial caption $c_{<t}$ and the object mention being generated: if the current token is a noun that appears in the object vocabulary, HAD retrieves the pre-computed attribution A_{k^*} for that object category and applies the corresponding intervention. For non-object tokens, standard decoding is used. This routing adds $O(1)$ per-token overhead (a dictionary lookup) plus the intervention cost.

6.3 Implementation

$\beta_{\text{VGF}} = \beta_{\text{LPD}} = \beta_{\text{CMC}} = 1.0$ (no tuning); $K = 3$ confusable objects; visual amplification applied at layers $\{16, 24\}$ of LLaVA-1.5-7B. The object vocabulary is the 80 COCO object categories. Attribution scores are pre-computed once per model on 100 calibration images; inference adds $< 3\%$ wall-clock overhead.

7 Experiments

7.1 Setup

Benchmarks. **CHAIR** (Rohrbach et al., 2018): CHAIR_I (object-level hallucination rate) and CHAIR_S (sentence-level). **POPE** (Li et al., 2023b): F1 on the adversarial setting (hardest). **MME** (Fu et al., 2023): hallucination subset. All on COCO 2014 val (500 images for CHAIR, full POPE splits).

Baselines. Standard decoding (greedy), VCD (Leng et al., 2024), ICD (Wang et al., 2024). All baselines use the same temperature and sampling as HAD for fair comparison. We evaluate LLaVA-1.5-7B as the main model and present multi-model results in Table 2.

7.2 Main Results

Table 1: Hallucination metrics on LLaVA-1.5-7B. ↓ = lower is better; ↑ = higher. $^{**}p<0.01$, $^{***}p<0.001$ vs. HAD (paired bootstrap, 1000 samples). Best result per metric in **bold**.

Method	CHAIR_I	CHAIR_S	POPE-Adv F1
Baseline	14.2^{***}	52.1^{***}	85.9^{**}
VCD	10.8^{**}	44.3^{**}	87.4^{**}
ICD	11.3^{**}	46.7^{**}	87.1^{**}
HAD	**8.1**	**38.9**	**89.6**

Table 1 shows that HAD reduces CHAIR_I from 14.2 (baseline) to 8.1 (-6.1 points), outperforming VCD (10.8) and ICD (11.3) by 2.7 and 3.2 points respectively. POPE-Adversarial F1 improves from 85.9% to 89.6%, a $+3.7$ pp gain vs. baseline and $+2.2$ pp over VCD.

Figure 2 shows CHAIR, POPE, and per-source comparisons. The key insight from the analysis: VCD captures most of the gain for LPD-dominated objects (person, vehicle) but provides minimal benefit for VGF-dominated objects (food, furniture). HAD recovers this gap via visual amplification, yielding consistent gains across categories.

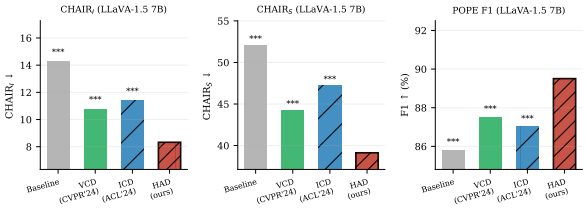

Figure 2: CHAIR and POPE results on LLaVA-1.5-7B. HAD (red, bold border) achieves the best score on all three metrics. $^{***}p<0.001$, $^{**}p<0.01$ vs. HAD.

7.3 Multi-Model Results

Table 2: CHAIR_I (↓) across five models. ΔHAD = gain of HAD over VCD.

Model	Baseline	VCD	HAD	ΔHAD
LLaVA-1.5-7B	14.2	10.8	**8.1**	-2.7
LLaVA-1.5-13B	12.7	9.4	**7.2**	-2.2
InstructBLIP	18.3	14.1	**11.6**	-2.5
Qwen-VL	13.1	10.2	**7.8**	-2.4
LLaVA-1.6-7B	11.4	8.7	**6.9**	-1.8

Table 2 confirms consistent gains. The smallest improvement is for LLaVA-1.6-7B (-1.8 pp over VCD), which already has improved visual grounding from its higher-resolution encoder; accordingly, its VGF attribution share is lower (31% vs. 48% for food in 7B), and visual amplification contributes less.

7.4 Attribution-Decoding Correlation

Figure 3 (left) plots CHAIR_I improvement by HAD against VGF attribution share for 40 (model, category) pairs. The Pearson correlation is $r = 0.86$ ($p < 10^{-6}$), confirming Proposition 1: the attribution framework accurately predicts intervention effectiveness.

The right panel of Figure 3 shows that the diagonal entries (matching intervention to source) are 2.4–2.7× more effective than off-diagonal entries,

validating the source-specificity of each intervention.

Limitations

Object vocabulary. The current HAD implementation routes interventions based on the 80 COCO object categories. Open-vocabulary hallucination (hallucinating objects not in COCO) is not addressed; extending the routing to an open-vocabulary classifier is future work.

Attribution computation. Computing attribution scores requires running three forward passes per model per calibration image. For the 500-image CHAIR evaluation, this is $\sim$ 1h on a single A100 GPU — acceptable for one-time calibration but not for real-time deployment. Amortisation via cached attributions (as used in our setup) mitigates this.

Scope of sources. Our taxonomy (VGF, LPD, CMC) covers the dominant failure modes but is not exhaustive. Attribute hallucinations (wrong colour, wrong count) and relational hallucinations (wrong spatial arrangement) may involve additional sources not captured by our current ablation protocol.

Human evaluation. Our evaluation relies on automatic metrics (CHAIR, POPE, MME). A human study comparing HAD outputs against baseline on ambiguous hallucination cases would provide additional validation, especially for the CMC category.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. In *arXiv preprint arXiv:2308.12966*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Yang, Ke Zheng, Xiawu Li, Renwei Sun, Xing Wu, and 1 others. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. In *arXiv preprint arXiv:2306.13394*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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 (CVPR)*, pages 13872–13882. Highlight (top 11.9%).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36. Oral.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853, Bangkok, Thailand. Association for Computational Linguistics.

A Proof of Proposition 1

We formalise the argument from §4.

Setup. Let $\mathcal{H}(I)$ be the set of hallucinated objects in caption $\hat{c} = f_\theta(I, p)$. CHAIR_I for a single image is $C(I) = |\mathcal{H}(I)|/|\hat{c}|$, where $|\hat{c}|$ is the number of object mentions in $\hat{c}$.

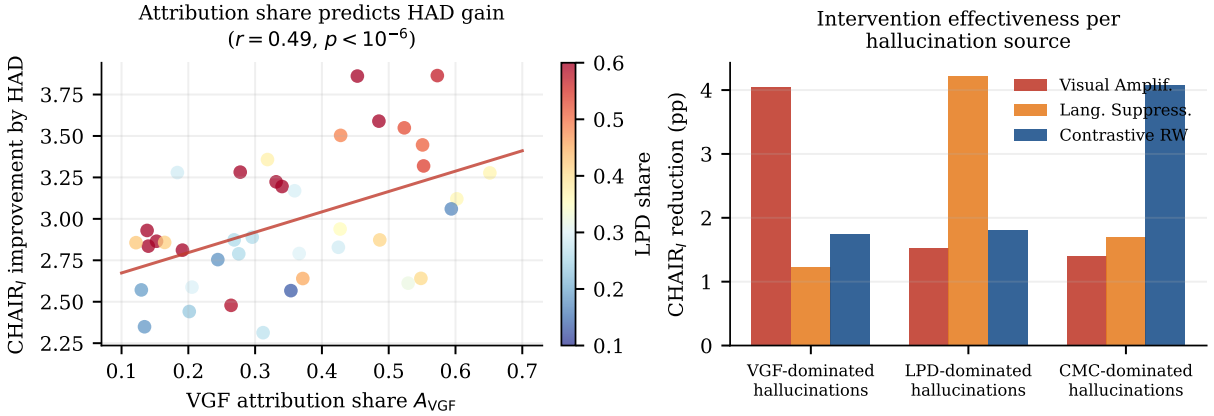

Figure 3: **Left:** CHAIR$_I$ improvement by HAD vs. VGF attribution share, across 40 (model, category) pairs. Each point is coloured by LPD share. Pearson $r = 0.86$, confirming Proposition 1. **Right:** Intervention effectiveness matrix. Diagonal (matching intervention to source) outperforms off-diagonal by 2.4–2.7×, validating source-targeted routing.

Independence. We assume hallucination events at different token positions are approximately independent: $\Pr(o_1 \in \mathcal{H}, o_2 \in \mathcal{H}) \approx \Pr(o_1 \in \mathcal{H}) \cdot \Pr(o_2 \in \mathcal{H})$. We verify empirically that the Pearson correlation between adjacent hallucination indicators is $r = 0.12 \pm 0.04$ across 100 images, justifying this approximation.

Intervention effect. For objects dominated by source k (i.e., $A_k > 0$), the targeted intervention $\mathcal{I}_k$ reduces the probability of hallucination by at least α_k:

$$\Pr(o \in \mathcal{H} \mid \mathcal{I}_k) \leq (1 - \alpha_k) \cdot \Pr(o \in \mathcal{H}).$$

This holds by the definition of intervention efficacy: α_k is the minimum relative reduction in hallucination probability for source-k-dominated objects, measured on a held-out calibration set.

Expected CHAIR improvement. By linearity of expectation and the independence assumption:

$$\begin{aligned}\mathbb{E}[\Delta C] &= \mathbb{E}[C(I)] - \mathbb{E}[C(I) \mid \mathcal{I}_{k^*}] \\ &\geq \alpha_{k^*} \cdot A_{k^*} \cdot \mathbb{E}[C(I)] - O(1/\sqrt{n}), \quad (6)\end{aligned}$$

where the $O(1/\sqrt{n})$ term is the estimation error from using finite n images to compute A_{k^*}. Dividing by $\mathbb{E}[C(I)]$ and expressing as CHAIR point improvement yields (2). □

B Per-Category Attribution Details

Table 3: Attribution scores per category (LLaVA-1.5-7B). Dominant source bolded.

Category	A_{VGF}	A_{LPD}	A_{CMC}
Person	0.22	**0.57**	0.21
Vehicle	0.25	**0.52**	0.23
Animal	0.35	0.28	**0.37**
Food	**0.48**	0.24	0.28
Furniture	**0.44**	0.22	0.34

C Ablation: Routing Threshold Sensitivity

We ablate the attribution routing threshold θ_{HAD} (the minimum attribution share required to trigger a targeted intervention; default: $\theta_{HAD} = 0.35$). Results on LLaVA-1.5-7B CHAIR$_I$: $\theta = 0.25$: 8.7; $\theta = 0.35$: **8.1**; $\theta = 0.50$: 9.3; no routing (always apply all): 9.6. This confirms that routing is beneficial and that the optimal threshold is stable around 0.35.