

Mitigating Object Hallucinations in Vision-Language Models via Head-Level Attribution and Counteractive Pruning

Liyan Wu^{1,*}, Xin Luo^{1,*}, Fan Wu^{1,†}, Yingen Liu¹, and Xiaobin Wu²

¹College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

²School of Information Engineering, Changsha Medical University, Changsha, China
{diana,luxinhd,wufan,liuyingen}@hnu.edu.cn, wuxiaobin@csmu.edu.cn

*Equal contribution. †Corresponding author.

Abstract

Large vision-language models (LVLMs) excel at multimodal tasks but often generate instance-level object hallucinations, describing nonexistent objects. Yet most mitigation strategies rely on output-level or coarse-grained interventions, limiting their ability to identify and selectively regulate attention heads associated with faithful versus hallucinatory generation, often at the cost of descriptive coverage. To address this limitation, we propose HACP, a training-free framework that integrates head-level semantic attribution with instance-adaptive counteractive pruning. HACP introduces InfoSpectralScore to summarize spectral changes in head-perturbed representations and identify heads associated with hallucination or grounded generation, which are then selectively suppressed or reinforced during decoding. Across multiple LVLMs and benchmarks, HACP consistently reduces hallucinations while preserving descriptive coverage; on a 1000-image multi-label COCO subset, it lowers CHAIR-s/i from 66.7/18.0 to 43.4/8.6 and improves F1 from 73.0 to 77.0. These findings suggest that head-level attribution provides an actionable signal for fine-grained inference-time hallucination control in LVLMs, while helping preserve descriptive coverage. The official project repository for HACP is available at <https://github.com/Slyang/HACP>.

1 Introduction

Large vision-language models (LVLMs) (Yin et al., 2024a; Liu et al., 2024c,a; Chen et al., 2023) are driving significant progress in image captioning, visual question answering, and multimodal dialogue. However, these models often generate *instance-level object hallucinations*, describing objects that are not present in the input image. Such errors critically undermine their reliability in safety-sensitive applications.

While numerous hallucination mitigation techniques have been proposed, such as instruction

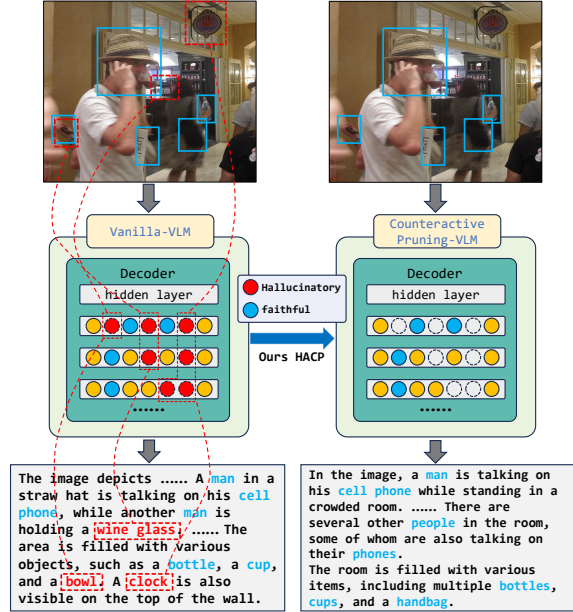


Figure 1: Object Hallucination Mitigation via Head-Level Pruning. HACP reduces hallucinations by pruning attention heads associated with spurious outputs, leading to more faithful captions than unmodified LVLM inference.

tuning, contrastive decoding, retrieval augmentation, and enhanced visual grounding, these methods largely operate at the output level. Consequently, existing approaches often struggle with the longstanding trade-off between hallucination suppression and semantic informativeness (Liu et al., 2024d; He et al., 2025).

Recent advances in mechanistic interpretability suggest that the behavior of large language models can often be connected to specific internal components and computational pathways. In LVLMs, multi-head attention enables cross-modal fusion and semantic abstraction, but most existing approaches treat attention either monolithically or at the coarse layer level. This overlooks a crucial fact: individual attention heads may exhibit distinct semantic functions, and growing evidence shows that

some heads disproportionately contribute to faithful or hallucinatory outputs (Bi et al., 2025; Geng et al., 2023; Jiang et al., 2025; Chen et al., 2024a). Such *functional conflicts* among attention heads may underpin the longstanding trade-off between hallucination suppression and semantic informativeness, yet current approaches rarely provide a principled, fine-grained attribution and intervention framework at the head level to resolve this challenge (Liu et al., 2024d; He et al., 2025).

To fill this gap, we propose **HACP**, a unified head-level attribution and dynamic pruning framework for LVLMs, inspired by recent advances in the analysis of internal model behaviors. Central to our approach are two key innovations: (1) *InfoSpectralScore*, a semantics-based attribution metric that enables systematic identification of faithful and hallucination-prone attention heads; and (2) a dynamic, training-free counteractive pruning strategy that adaptively suppresses hallucination-prone heads while reinforcing faithful ones during inference. As illustrated in Figure 1, our framework enables automated, fine-grained internal control, resulting in more accurate and faithful captions. Extensive experiments across diverse LVLMs and benchmarks demonstrate that our method reduces hallucinations and better preserves caption informativeness compared to existing approaches.

Our main contributions are as follows:

- We introduce HACP, a training-free inference-time framework for mitigating object hallucinations in LVLMs through fine-grained head-level attribution and intervention.
- We propose InfoSpectralScore, a semantics-oriented attribution metric for identifying faithful and hallucination-prone attention heads, and use it to drive a dynamic counteractive pruning strategy that suppresses harmful heads while reinforcing faithful ones during decoding.
- Extensive experiments across multiple LVLMs and benchmarks show that HACP consistently reduces object hallucinations while largely preserving descriptive coverage and overall multimodal performance.

2 Related Work

Built on multimodal pretraining and visual instruction tuning, LVLMs are designed to generate semantically grounded descriptions of visual inputs (Alayrac et al., 2022; Li et al., 2023a; Zhu et al., 2024; Dai et al., 2023; Liu et al., 2024c; Chen

et al., 2023). Yet recent studies consistently show that they often describe nonexistent objects, a phenomenon known as object hallucination (Rohrbach et al., 2018; Liu et al., 2024b). Such errors occur at the instance (spurious objects), attribute (incorrect properties), and spatial (misplaced relations) levels, undermining reliability in high-stakes domains such as assistive technology and autonomous systems.

Previous attempts to mitigate hallucinations largely relied on output-level or decoding-oriented adjustments. Instruction tuning with curated image–text pairs (Zhang et al., 2024; Liu et al., 2024a; Dai et al., 2023; Li et al., 2025a) and grounding-aware objectives (Favero et al., 2024) improved general visual reasoning but offered limited control over hallucinations (Geigle et al., 2024). Contrastive, grounding-aware, and guided-decoding strategies further constrained generation using external visual signals or exemplar-based guidance (Deng et al., 2024; Leng et al., 2024; Qu et al., 2025; Chen et al., 2024b; Xie et al., 2024; Li et al., 2025b; Zhao et al., 2025). These approaches reduce hallucinations during inference but often compromise output diversity or informativeness. External expert-based systems take a different path: instead of steering the decoding process, they introduce additional correction networks or logic-aware verifiers to post-edit model outputs (Yin et al., 2024b; Wu et al., 2024; Yu et al., 2024). These expert modules often improve reliability but at the cost of increased architectural complexity and inference latency.

Beyond such output-level techniques, advances in mechanistic interpretability enabled the decomposition of deep models into interpretable circuits and, together with activation patching, localized causal associations in language models (Nanda et al., 2023; Elhage et al., 2022; Wang et al., 2023; García-Carrasco et al., 2025; Azaria and Mitchell, 2023; Meng et al., 2022). For vision–language models, attention-based attribution and cross-modal analyses became prominent (Bi et al., 2025; Geng et al., 2023; Gandelsman et al., 2024), yet most strategies relied on simple heuristics or layer-level aggregation (Sarkar et al., 2025; He et al., 2025), lacking a semantics-grounded head-level attribution framework. Recent works further revealed that specific attention heads in intermediate layers are particularly responsible for hallucinations (Jiang et al., 2025; Chen et al., 2024a; Yang et al., 2025), motivating the development of more fine-grained

and interpretable intervention mechanisms.

3 Method

3.1 Motivation

Despite remarkable progress, LVLMs still suffer from instance-level hallucinations, which recent evidence (Yang et al., 2025; He et al., 2025; Sarkar et al., 2025) links to functional conflicts among attention heads: some promote faithful grounding, while others introduce spurious objects. These conflicts appear central to the longstanding trade-off between hallucination suppression and semantic informativeness. Yet current methods lack principled, fine-grained mechanisms for diagnosing and intervening at the head level.

First, existing approaches rarely provide semantics-grounded metrics to attribute hallucinations to specific attention heads. Output-level measures, such as Kullback–Leibler (KL) divergence between token distributions p and q over the vocabulary $\mathcal{V}$ (e.g., before and after perturbing a head under the same context),

$$D_{\text{KL}}(p \parallel q) = \sum_{i \in \mathcal{V}} p_i \log \left(\frac{p_i}{q_i} \right), \quad (1)$$

where p_i and q_i are the probabilities assigned to token i , only capture distributional shifts after ablation but cannot reveal whether the change improves semantic alignment. Both faithful and hallucination-prone heads may yield high KL scores, limiting diagnostic utility.

Inspired by research in LLMs (Chen et al., 2024a) and empirical analyses of hallucination-associated heads in LVLMs (Jiang et al., 2025), we propose *InfoSpectralScore*: a semantics-based, head-level attribution metric that systematically quantifies the semantic contribution of each attention head in targeted intermediate layers, enabling precise identification of both hallucination-prone and faithful heads.

Second, conventional head-pruning strategies (e.g., zero or mean ablation) are often too severe, over-penalizing the model and degrading semantic expressivity. To address this, we develop training-free counteractive pruning: a dynamic, inference-time intervention that suppresses hallucination-inducing heads while reinforcing faithful ones, achieving hallucination mitigation without compromising output quality or semantic richness.

After addressing these challenges, we introduce a unified pipeline for head-level attribution and

inference-time hallucination mitigation (Figure 2). Building on this foundation, we further design an automated, instance-adaptive mechanism, with implementation details given in Section 3.4.

3.2 Semantic Attribution via InfoSpectralScore

Algorithm 1 Head-Level Attribution via Sequential Head Pruning

Require: Model M , attribution set $\mathcal{D}$, target layers $\mathcal{L}$, per-layer head counts $\{n_h(l)\}_{l \in \mathcal{L}}$, rounds R

Ensure: Attribution scores $\mathcal{S}_{l,j}$; per-layer faithful/hallucinatory head sets $\mathcal{H}_{\text{faith}}^{(l)}, \mathcal{H}_{\text{hallu}}^{(l)}$

```

1: for round  $r = 1$  to  $R$  do
2:   for each layer  $l \in \mathcal{L}$  do
3:     for head index  $j = 1$  to  $n_h(l)$  do
4:       Prune head  $(l, j)$ , run inference on  $\mathcal{D}$ 
5:       Compute  $\text{InfoSpectralScore}(l, j)$ 
6:       Append  $\text{InfoSpectralScore}(l, j)$  to  $\mathcal{S}_{l,j}$ 
7:     end for
8:   end for
9: end for
10: Aggregate  $\mathcal{S}_{l,j}$  across rounds
11: Classify heads by thresholds to obtain  $\mathcal{H}_{\text{faith}}^{(l)}, \mathcal{H}_{\text{hallu}}^{(l)}$ 
12: return  $\{\mathcal{S}_{l,j}\}, \{\mathcal{H}_{\text{faith}}^{(l)}\}, \{\mathcal{H}_{\text{hallu}}^{(l)}\}$ 

```

Let $\mathcal{D}$ denote the attribution set of image–prompt pairs used for head-level attribution, with $|\mathcal{D}|$ denoting its size. To isolate the interventional effect of each head, we apply mean-replacement perturbations on $\mathcal{D}$ and collect the resulting final-token semantic representations. The Sequential Head Pruning Module (SHPM) traverses all heads in target layers $\mathcal{L}$ (see Algorithm 1). Stacking these for K samples yields $Z \in \mathbb{R}^{K \times d}$. The covariance matrix is

$$\Sigma = \frac{1}{K-1} (Z - \bar{Z})^\top (Z - \bar{Z}) + \epsilon I, \quad (2)$$

with eigenvalues $\lambda_1 \geq \dots \geq \lambda_d$. The *EigenScore* is defined as

$$\text{EigenScore} = \sum_{i=1}^{k^*} \lambda_i, \quad (3)$$

where k^* is the smallest integer such that $\sum_{i=1}^{k^*} \lambda_i \geq \tau \sum_{j=1}^d \lambda_j$ (e.g., $\tau = 0.9$), capturing

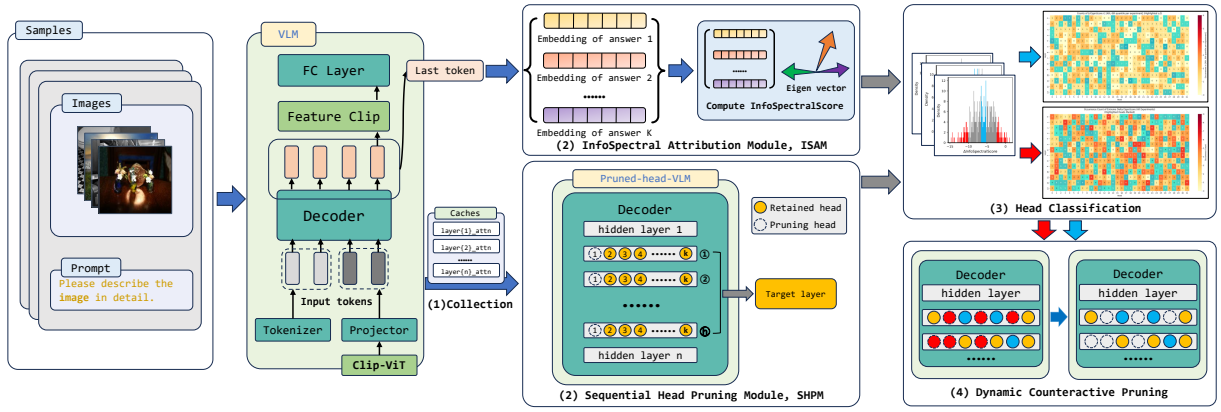


Figure 2: **Unified pipeline for head-level attribution and inference-time hallucination mitigation.** Our framework consists of four stages: (1) activation patching and cache collection, (2) head-level attribution via InfoSpectralScore, (3) faithful vs. hallucinatory head classification, and (4) dynamic counteractive pruning during inference.

the dominant variance scale of the representation under head-level intervention.

However, EigenScore alone may miss pathologies such as *anisotropy* and *sparsity* in hallucination-prone heads. We thus introduce the InfoSpectral Attribution Module (ISAM), which augments EigenScore with two regularizers:

$$\text{SpectralVar} = \text{Var}\left(\{\log \lambda_i\}_{i=1}^d\right). \quad (4)$$

which penalizes anisotropy by measuring dispersion in the log-spectrum, where d is the total number of eigenvalues.

$$p_i = \frac{\lambda_i}{\sum_{j=1}^d \lambda_j}, \quad (5)$$

$$\text{SpectralEntropy} = - \sum_{i=1}^d p_i \log(p_i + \epsilon), \quad (6)$$

which characterizes the concentration of the normalized spectrum via its entropy and measures how evenly semantic variance is distributed across dimensions, where ϵ is a small constant (e.g., 10^{-8}) for numerical stability.

Combining these, the InfoSpectralScore is defined as:

$$\begin{aligned} \text{InfoSpectralScore} = & \text{EigenScore} \\ & - \alpha \text{SpectralVar} \\ & + \gamma \text{SpectralEntropy}, \end{aligned} \quad (7)$$

where α and γ are regularization weights. Hyperparameters are selected via grid search (see Appendix D.2).

Algorithm 2 Dynamic Counterbalance of Attention Heads

Require: For each layer: faithful set $\mathcal{H}_{\text{faith}}$, hallucinatory set $\mathcal{H}_{\text{hallu}}$, head outputs $\{\mathbf{h}_j\}$; modulation μ , smoothing λ

- 1: Remove overlaps: $\mathcal{H}'_{\text{hallu}} \leftarrow \mathcal{H}_{\text{hallu}} \setminus \mathcal{H}_{\text{faith}}$
- 2: **for** each layer **do**
- 3: **if** $|\mathcal{H}_{\text{faith}}| < 2$ or $|\mathcal{H}'_{\text{hallu}}| < 2$ **then**
- 4: **continue**
- 5: **end if**
- 6: $\mathbf{v}_f \leftarrow \frac{1}{|\mathcal{H}_{\text{faith}}|} \sum_{j \in \mathcal{H}_{\text{faith}}} \mathbf{h}_j$
- 7: $\mathbf{v}_h \leftarrow \frac{1}{|\mathcal{H}'_{\text{hallu}}|} \sum_{j \in \mathcal{H}'_{\text{hallu}}} \mathbf{h}_j$
- 8: $D \leftarrow \|\mathbf{v}_h\| / (\|\mathbf{v}_f\| + \epsilon)$
- 9: $\tilde{D} \leftarrow \lambda \tilde{D}_{\text{prev}} + (1 - \lambda) D$
- 10: **for** each head j **do**
- 11: **if** $j \in \mathcal{H}_{\text{faith}}$ **then**
- 12: $\mathbf{h}_j \leftarrow \mathbf{h}_j + \mu \tilde{D} \mathbf{v}_f$
- 13: **else if** $j \in \mathcal{H}'_{\text{hallu}}$ **then**
- 14: $\mathbf{h}_j \leftarrow \mathbf{h}_j - \mu \tilde{D} \mathbf{v}_h$
- 15: **end if**
- 16: **end for**
- 17: **end for**

3.3 Counterbalancing Attention Heads via Directional Pruning

To suppress hallucination-inducing behaviors while preserving faithful generation, we introduce a dynamic counteractive pruning mechanism (see Algorithm 2). Let $\mathcal{H}_{\text{faith}}$ and $\mathcal{H}_{\text{hallu}}$ denote the sets of faithful and hallucinatory heads identified by ISAM. We first remove overlaps to avoid ambiguous signals:

$$\mathcal{H}'_{\text{hallu}} = \mathcal{H}_{\text{hallu}} \setminus \mathcal{H}_{\text{faith}}. \quad (8)$$

For each layer, let $\mathbf{h}_j$ be the output vector of head j at the final token position. We compute the per-set mean directions

$$\begin{aligned}\mathbf{v}_f &= \frac{1}{|\mathcal{H}_{\text{faith}}|} \sum_{j \in \mathcal{H}_{\text{faith}}} \mathbf{h}_j, \\ \mathbf{v}_h &= \frac{1}{|\mathcal{H}'_{\text{hallu}}|} \sum_{j \in \mathcal{H}'_{\text{hallu}}} \mathbf{h}_j.\end{aligned}\quad (9)$$

The modulation ratio is

$$D = \frac{\|\mathbf{v}_h\|}{\|\mathbf{v}_f\| + \epsilon}, \quad (10)$$

and its exponentially smoothed version is

$$\tilde{D} = \lambda \tilde{D}_{\text{prev}} + (1 - \lambda) D. \quad (11)$$

Finally, each head j is reweighted as

$$\mathbf{h}'_j = \begin{cases} \mathbf{h}_j + \mu \tilde{D} \mathbf{v}_f, & j \in \mathcal{H}_{\text{faith}}, \\ \mathbf{h}_j - \mu \tilde{D} \mathbf{v}_h, & j \in \mathcal{H}'_{\text{hallu}}, \\ \mathbf{h}_j, & \text{otherwise.} \end{cases} \quad (12)$$

We skip any layer with fewer than two heads in either set. We implement the mechanism via forward hooks at inference, reshaping the post-attention tensor to $[B, T, H, D]$, computing $\mathbf{v}_f, \mathbf{v}_h$, updating $\tilde{D}$, and applying Eq. (12) on GPU without extra parameters. Hyperparameters μ and λ are selected via grid search (Appendix D.3).

3.4 Unified Automation and Task-Specific Adaptation Pipeline

In practical LVLM deployments, hallucination behavior varies across image–prompt pairs, motivating an instance-adaptive pruning pipeline that integrates the attribution module (Section 3.2) with the counterbalance mechanism (Section 3.3). Algorithm 3 outlines this procedure. In this task-specific setting, the attribution set in Algorithm 1 contains only the current image–prompt pair, i.e., $\mathcal{D} = \{(I, P)\}$ and $|\mathcal{D}| = 1$. The pipeline then identifies faithful and hallucinatory directions, runs a small Bayesian optimization loop over (μ, λ) under a hallucination-aware objective, and applies the best configuration to generate the final caption. Appendix B.1 details the two engineering improvements that reduce attribution and search overhead in practice.

Algorithm 3 Task-Specific Pruning Pipeline for a Single Image–Prompt Pair

Require: LVLM M , image I , prompt P , BO budget T

Ensure: Pruned caption $\hat{y}_p$

- 1: **Attribution (Algorithm 1):** run head-level ablations on (I, P) to obtain InfoSpectralScore for all heads and classify faithful/hallucinatory sets.
 - 2: Run Vanilla inference to obtain $\hat{y}_0$ and metrics m_0 .
 - 3: **for** $t = 1$ to T **do**
 - 4: Use Bayesian optimization to propose pruning parameters (μ_t, λ_t) .
 - 5: Instantiate the counterbalance module (Algorithm 2) using (μ_t, λ_t) and the attribution-derived head sets.
 - 6: Run M with this configuration to obtain $\hat{y}_t$ and metrics m_t .
 - 7: Update BO surrogate using hallucination-aware objective $J(m_t, m_0)$.
 - 8: **end for**
 - 9: Let (μ^*, λ^*) be the best BO parameters.
 - 10: Apply Algorithm 2 with (μ^*, λ^*) to produce the final caption $\hat{y}_p$.
-

4 Experiments

4.1 Experimental Setup

We evaluate our approach on three representative LVLMs: LLaVA-1.5 (7B, 13B), Shikra-7B, and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). For terminology, *Vanilla* denotes the original LVLM inference under the default model parameters and decoding setup, without any hallucination-control intervention; competing baselines refer to competitive training-free hallucination-mitigation methods, including VCD (Leng et al., 2024), OPERA (Huang et al., 2024), PAI (Liu et al., 2024d), SPIN (Sarkar et al., 2025), MLIH (Jiang et al., 2025), and VHR (He et al., 2025), each evaluated using its officially recommended settings. Diagnostic analyses in Section 4.2 and part of Section 4.3 use the generalization pipeline, whereas comparisons against competing baselines use *Ours*, the full task-specific pruning pipeline described in Algorithm 3. *Ours-fixed* denotes the same pipeline under a single fixed hyperparameter configuration.

For evaluation, we follow the CHAIR protocol (Rohrbach et al., 2018) on the COCO 2014 dataset (Lin et al., 2014), reporting sentence-level

Table 1: Relative improvement (%) from Vanilla to Pruned on hallucination metrics. **Bold** indicates the best result within each model and evaluation-set block.

Group	Validation Δ (%)			Hallucination Δ (%)			Generalization Δ (%)		
	$C_s \downarrow$	$C_i \downarrow$	F1 $\uparrow$	$C_s \downarrow$	$C_i \downarrow$	F1 $\uparrow$	$C_s \downarrow$	$C_i \downarrow$	F1 $\uparrow$
Model: LLaVA-7B									
[5,18]	-6.85	-1.96	+0.6	-11.3	-3.81	+1.31	-4.11	-5.72	+0.37
[19,26]	-8.58	-7.27	+1.35	-5.24	-3.16	+0.73	-3.61	-5.02	+1.63
Merged	-8.05	-3.29	-0.04	-9.77	-4.34	-0.07	-6.25	-5.55	+0.55
Model: LLaVA-13B									
[5,18]	-7.41	-12.0	+2.27	-7.89	-2.57	-0.61	-5.57	-6.57	-0.19
[19,26]	-12.3	-5.54	+2.59	-4.37	-5.32	+0.77	-5.48	-6.54	+0.77
Merged	-9.66	-2.23	+3.92	-7.99	-2.8	-0.79	-8.8	-1.24	+1.87
Model: Shikra-7B									
[3,13]	-5.46	-3.05	-0.06	-10.7	-4.06	+2.43	-1.96	-8.27	+0.69
[14,28]	-10.4	-5.13	+0.9	-9.26	-8.94	+1.29	-8.11	-5.13	+0.33
Merged	-2.17	-8.67	+1.91	-10.3	-6.13	+3.34	-6.32	-10.9	+3.44

and instance-level CHAIR scores (CHAIR_s and CHAIR_i , denoted as C_s and C_i), along with recall, precision, F1, and caption length.

On POPE (Li et al., 2023b), we report accuracy, precision, recall, and F1. To assess broader multimodal capabilities beyond object hallucination, we further evaluate on MME (Fu et al., 2025), applying our method to LLaVA-1.5-7B. We also include supplementary caption-level evaluations using ALOHa (Petryk et al., 2024) and GPT-4V-aided pairwise assessment in Appendix C.2 and Appendix C.3.

Appendix A provides implementation details, runtime and cost details, and configurations for our method and all competing baselines.

4.2 Generalization Pipeline: Systematic Cross-Model Evaluation

The generalization pipeline evaluates whether head-level attribution and pruning remain effective across different model architectures and data distributions. This analysis focuses on the *consistency* and *stability* of pruning effects across models and evaluation splits. The central questions are: (1) Can our attribution and pruning strategy consistently suppress hallucination across diverse model architectures? (2) Does this suppression generalize to out-of-distribution data splits?

For each model and layer group, we use three controlled sets. The **Validation Set** is the attribution set $\mathcal{D}$ used to identify faithful and hallucinatory head partitions, with $|\mathcal{D}| = 160$ for each LLaVA-7B/13B group and $|\mathcal{D}| = 320$ for each Shikra-7B group. The **Hallucination Set** contains images that produce fully hallucinatory Vanilla captions ($C_s = 1$), while the **Generalization Set** contains

Table 2: Relative improvement (%) for LLaVA-7B under random and hallucination-focused attribution. **Bold** indicates the best result within each group and evaluation-set block.

Group	Validation Δ (%)			Hallucination Δ (%)			Generalization Δ (%)		
	$C_s \downarrow$	$C_i \downarrow$	F1 $\uparrow$	$C_s \downarrow$	$C_i \downarrow$	F1 $\uparrow$	$C_s \downarrow$	$C_i \downarrow$	F1 $\uparrow$
Random attribution (LLaVA-7B)									
[5,18]	-6.85	-1.96	+0.6	-11.3	-3.81	+1.31	-4.11	-5.72	+0.37
[19,26]	-8.58	-7.27	+1.35	-5.24	-3.16	+0.73	-3.61	-5.02	+1.63
Merged	-8.05	-3.29	-0.04	-9.77	-4.34	-0.07	-6.25	-5.55	+0.55
Hallucination-focused attribution (LLaVA-7B-hallu)									
[5,18]	-9.03	-6.12	+3.2	-7.55	-1.35	+0.7	-11.0	-14.1	+1.41
[19,26]	-3.18	-1.49	+1.97	-3.87	+1.36	+0.63	-9.6	-2.81	+1.01
Merged	-10.9	-7.4	+2.19	-8.32	-9.13	+0.57	-15.3	-20.0	+3.36

held-out images outside the corresponding Validation Set. Both evaluation sets are size-matched to their corresponding Validation Set. For merged groups, each set is the de-duplicated union of the corresponding primary groups; additional protocol details are provided in Appendix A.1. Following the empirical findings of (Jiang et al., 2025; Li et al., 2025b), for each model, we define three attribution layer groups: LLaVA uses [5,18], [19,26], and a merged [5,26]; Shikra-7B uses [3,13], [14,28], and a merged [3,28]. Grid search for hyperparameters (μ , λ) is conducted independently on each split, with settings detailed in Appendix D.3.

As shown in Table 1, our attribution-guided pruning consistently reduces both C_s and C_i across all models and evaluation splits, demonstrating stable suppression of hallucinations under varying architectural and distributional conditions. Notably, F1 remains stable or improves in nearly all cases, underscoring the robustness and generalizability of our method in mitigating hallucinations without sacrificing informativeness.

4.3 Hallucination-Focused Attribution: An Empirical Exploration

Hallucination cases provide attribution signals that are rarely visible in ordinary validation images, exposing larger variations in attention patterns and InfoSpectralScore responses. This motivates an alternative setting in which attribution is computed on intentionally challenging inputs, with the expectation that such cases may better reveal heads associated with hallucination and thus support more effective pruning. Concretely, for each model, we identify images that produce hallucinated objects under Vanilla inference and use these hallucination-prone cases as challenging attribution inputs. For the task-specific comparison in Table 3, we eval-

Table 3: Comparison of hallucination-mitigation methods on hallucination-prone evaluation sets with 500 COCO validation images per model. Best results are in **bold** and second best are underlined within each model block, excluding Vanilla.

Method	$C_s \downarrow$	$C_i \downarrow$	Recall $\uparrow$	Precision $\uparrow$	F1 $\uparrow$	Len
Model: LLaVA-7B						
Vanilla	94.4	31.0	84.9	58.5	67.0	93.56
VCD	73.1	25.9	76.0	64.3	69.7	92.33
OPERA	73.1	23.4	82.7	68.3	<u>74.8</u>	95.99
PAI	55.6	<u>19.7</u>	72.8	70.4	71.6	87.41
MLIH	63.7	21.1	75.9	70.9	73.3	80.01
SPIN	55.6	19.8	66.3	<u>72.6</u>	69.3	67.75
VHR	52.5	22.1	67.9	69.8	68.9	64.83
Ours	<u>53.8</u>	14.5	85.7	77.0	78.6	90.96
Model: Shikra-7B						
Vanilla	78.5	24.5	77.8	62.3	69.2	99.97
VCD	66.8	22.2	73.6	66.8	70.1	97.84
OPERA	69.0	23.2	76.0	67.0	<u>71.0</u>	101.20
PAI	71.5	22.7	<u>78.0</u>	64.3	70.5	98.14
MLIH	33.5	12.3	54.6	80.0	64.9	53.20
SPIN	57.0	17.8	72.9	69.4	71.1	92.61
VHR	<u>54.0</u>	19.8	66.2	70.9	68.4	75.20
Ours	60.1	<u>16.1</u>	78.9	<u>72.5</u>	75.5	98.89
Model: Qwen2.5-VL-7B-Instruct						
Vanilla	69.4	21.6	51.2	75.0	60.9	120.14
VCD	74.4	25.4	48.8	73.0	58.5	110.86
OPERA	55.6	25.5	33.8	71.8	46.0	119.75
PAI	56.3	24.4	38.1	74.7	50.4	88.82
MLIH	43.1	23.6	28.4	73.3	41.0	63.54
SPIN	<u>48.1</u>	23.0	28.4	76.1	41.4	62.86
VHR	52.5	<u>20.1</u>	40.4	<u>78.0</u>	53.2	105.43
Ours	60.0	14.8	54.5	81.7	65.4	120.43

uate on 500 COCO validation images per model that produce hallucinated objects under Vanilla inference.

Finding 1: Hallucination-focused attribution provides more informative pruning signals. We compare attributions computed from random images and hallucination-prone images under the same generalization pipeline and multi-round aggregation strategy. As shown in Table 2 for LLaVA-7B, hallucination-focused attribution generally produces stronger pruning gains, with the clearest improvements appearing in the merged group and on held-out evaluation images. This suggests that high-risk cases expose head-level patterns that are more relevant to hallucination control than randomly sampled images, while aggregation across layer ranges helps stabilize the resulting pruning behavior.

Finding 2: Hallucination-focused attribution enhances task-specific hallucination mitigation.

Table 4: Comparison of hallucination-mitigation methods on 1000 randomly selected multi-label COCO images. Best results are in **bold** and second best are underlined. [†] *Ours-fixed* uses one fixed configuration; sensitivity analysis is in Appendix D.4.

Method	$C_s \downarrow$	$C_i \downarrow$	Recall $\uparrow$	Precision $\uparrow$	F1 $\uparrow$	Len
Vanilla	66.7	18.0	69.1	77.3	73.0	104.79
VCD	72.0	17.4	<u>73.1</u>	76.9	74.9	115.14
OPERA	68.9	17.9	72.2	77.4	74.7	113.96
PAI	71.3	17.3	73.3	77.0	<u>75.1</u>	114.41
MLIH	51.9	<u>12.1</u>	60.3	<u>82.2</u>	69.6	98.69
SPIN	<u>49.4</u>	14.1	53.0	81.5	64.2	72.00
VHR	70.0	18.2	71.3	76.3	73.7	108.64
<i>Ours-fixed</i> [†]	58.1	12.7	70.1	83.9	74.5	103.84
Ours	43.4	8.6	70.8	89.0	77.0	103.00

We next examine whether the stronger attribution signals observed on hallucination-prone inputs translate into improved performance when used in the task-specific pruning pipeline on high-risk evaluation sets. As shown in Table 3, HACP obtains the lowest or near-lowest C_i across the three evaluated LLMs while achieving the highest recall and F1 within each model block. Together with caption lengths close to Vanilla, these results indicate that HACP improves instance-level grounding while retaining descriptive coverage. In comparison, some baselines reduce hallucination scores with shorter outputs or lower recall. These results support the use of hallucination-prone cases as an informative attribution source for task-specific head intervention.

4.4 Broad Evaluation Across Diverse Benchmarks

The previous sections examined attribution behavior and head selection under diagnostic settings. We now evaluate the task-specific pipeline in broader scenarios: a 1000-image multi-label COCO subset, full POPE, and MME, covering caption-level hallucination, object-level consistency, and broader multimodal capability.

Evaluation on Object-Dense COCO Images for Multi-Object Captioning. Table 4 reports results on an object-dense COCO subset of 1000 randomly selected images, where each image has at least five annotated object categories according to the COCO ground-truth labels. Compared to general captioning settings, this subset tends to induce longer descriptions because the images contain more annotated objects on average, making it a useful stress test for hallucination mitigation in

Table 5: Comparison of hallucination-mitigation methods on the full POPE benchmark. All metrics are higher-is-better. Best results are in **bold** and second best are underlined. [†] *Ours-fixed* uses one fixed configuration.

Method	Random				Popular				Adversarial			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
Vanilla	88.5	95.7	80.5	87.5	86.8	92.1	80.5	86.0	83.5	85.6	80.5	83.0
VCD	87.8	91.5	83.4	87.3	85.6	87.2	83.4	85.3	81.1	79.8	83.4	81.5
OPERA	89.2	92.5	85.3	88.8	86.8	88.0	85.3	86.6	81.2	78.8	85.3	81.9
PAI	88.3	85.6	<u>92.1</u>	88.7	84.9	80.6	<u>91.8</u>	85.8	76.5	70.3	<u>91.8</u>	79.6
MLIH	81.1	99.0	62.9	76.9	80.7	97.8	62.9	76.5	79.9	95.2	62.9	75.7
SPIN	89.6	94.3	84.3	89.0	87.1	89.4	84.2	86.7	83.2	82.3	84.7	83.5
VHR	88.9	86.3	92.4	89.2	85.4	81.0	92.4	86.3	77.2	70.9	92.4	80.2
<i>Ours-fixed</i> [†]	88.8	96.9	80.3	87.8	<u>87.6</u>	94.0	80.3	86.6	<u>85.1</u>	88.7	80.5	<u>84.4</u>
Ours	90.4	<u>98.6</u>	82.0	89.5	89.8	<u>97.4</u>	81.9	89.0	88.7	<u>95.0</u>	81.8	87.9

Table 6: Summary of MME scores. Complete task-level results are provided in Appendix C.1.

Method	Perception	Cognition	Total
Vanilla	1374.50	304.64	1679.14
Ours	1486.89	364.29	1851.18

object-dense, multi-object captioning.

Across methods, different empirical behaviors emerge. VCD, OPERA, PAI, and VHR produce noticeably longer captions in this multi-label setting, yet do not exhibit commensurate reductions in hallucination metrics. In contrast, MLIH and SPIN substantially reduce C_s and C_i , but also show clear drops in recall and output length, indicating reduced descriptive coverage. Interestingly, VHR behaves similarly to MLIH and SPIN on hallucination-prone evaluation sets (Table 3), but shifts toward the former group in the multi-label setting, suggesting sensitivity to object density.

Both *Ours-fixed* and *Ours* substantially lower hallucination metrics while maintaining recall and caption length close to Vanilla. Notably, *Ours* achieves the lowest C_s and C_i among all compared methods, together with the highest precision and F1 score, while keeping recall and caption length close to Vanilla. *Ours-fixed* already provides a strong trade-off without per-instance search, while *Ours* further improves hallucination suppression and semantic precision through adaptive search. Appendix C.2 and Appendix C.3 further examine the same 1000-image subset with ALOHa and GPT-4V-aided evaluation, providing complementary evidence on open-vocabulary object grounding and caption-level faithfulness.

Evaluation on POPE for object-level binary consistency. Table 5 presents results on the full

POPE benchmark covering the Random, Popular, and Adversarial splits. In all three settings, both *Ours-fixed* and *Ours* demonstrate stable improvements in object-level consistency while maintaining balanced precision and recall. *Ours-fixed* already improves Acc and F1 on the Popular and Adversarial splits without per-instance search, while *Ours* further achieves the best Acc and F1 across all three splits. Across random, popular, and adversarial object queries, these gains suggest that the head-level adjustment improves object-level consistency while keeping precision and recall comparably balanced.

Evaluation on MME across multimodal perception and cognitive tasks. MME evaluates multimodal abilities beyond object hallucination, reporting a total score together with perception and cognition sub-scores. As shown in Table 6, applying HACP to LLaVA-1.5-7B increases the total MME score from 1679.14 to 1851.18, with gains in both perception (1374.50 to 1486.89) and cognition (304.64 to 364.29). Together with the task-level results in Appendix C.1, this suggests that HACP’s hallucination-control mechanism remains compatible with broader multimodal evaluation on MME.

5 Conclusions

This paper introduced HACP, a training-free inference-time framework for mitigating instance-level object hallucinations in LVLMs through head-level attribution and counteractive pruning. HACP uses InfoSpectralScore to identify attention heads associated with grounded and hallucinatory generation, and modulates their contributions during decoding to steer generation toward more visually supported outputs.

Across multiple LVLMs and evaluation settings, HACP consistently reduces object hallucinations while retaining Vanilla-level descriptive coverage and caption quality. Random-head and reversed-direction controls show that these gains depend on both head selection and intervention direction. Results on POPE and MME further suggest that this intervention remains compatible with broader object-level consistency and multimodal evaluation. These results show that semantics-oriented head attribution provides an actionable signal for fine-grained inference-time control toward more faithful LVLM generation.

Limitations

This study focuses on instance-level object hallucinations, particularly unsupported object mentions in LVLM outputs. Architectural differences may affect which heads are selected and how well the intervention transfers beyond the evaluated LVLMs. Attribute and spatial hallucinations remain outside the current evaluation. Future work will extend HACP to broader model families, hallucination types, and open-vocabulary settings.

The full task-specific pipeline adds per-instance attribution and Bayesian optimization to inference, requiring repeated decoding during parameter search. Appendix B.2 reports the measured runtime and scaling behavior. *Ours-fixed* avoids per-instance search and provides lower-latency operation, but applies a shared pruning configuration across inputs. Future work will investigate faster attribution and more efficient search strategies for tighter latency budgets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62472162, 62225205, and 62532005).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.

Jing Bi, Junjia Guo, Yunlong Tang, Liangdong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. [Unveiling visual perception in language models: An attention head analysis approach](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4135–4144.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. [INSIDE: LLMs’ internal states retain the power of hallucination detection](#). In *The Twelfth International Conference on Learning Representations*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. [Shikra: Unleashing multimodal llm’s referential dialogue magic](#). *Preprint*, arXiv:2306.15195.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024b. [HALC: Object hallucination reduction via adaptive focal-contrast decoding](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7824–7846. PMLR.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. 2025. [Qwen look again: Guiding vision-language reasoning models to re-attention visual information](#). *Preprint*, arXiv:2505.23558.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven C. Hoi. 2023. [InstructBLIP: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. [Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding](#). *Preprint*, arXiv:2402.15300.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. [Toy models of superposition](#). *Preprint*, arXiv:2209.10652.

- Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. [Multi-modal hallucination control by visual information grounding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 162549–162567.
- Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. 2024. [Interpreting CLIP’s image representation via text-based decomposition](#). In *The Twelfth International Conference on Learning Representations*.
- Jorge García-Carrasco, Alejandro Maté, and Juan Trujillo. 2025. [Extracting interpretable task-specific circuits from large language models for faster inference](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16772–16780.
- Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. [Does object grounding really reduce hallucination of large vision-language models?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2728–2742, Miami, Florida, USA. Association for Computational Linguistics.
- Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. 2023. [HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention](#). In *International Conference on Learning Representations*.
- Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2025. [Cracking the code of hallucination in LVLMs with vision-aware head divergence](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3488–3501, Vienna, Austria. Association for Computational Linguistics.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. [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 (CVPR)*, pages 13418–13427.
- Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. [Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25004–25014.
- 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.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. [Otter: A multi-modal model with in-context instruction tuning](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557.
- Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. 2025b. [Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding](#). *Preprint*, arXiv:2501.01926.
- 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*, 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 J. 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*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. [Mitigating hallucination in large multi-modal models via robust instruction tuning](#). In *International Conference on Learning Representations*.
- Hanchao Liu, Wenyan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. [A survey on hallucination in large vision-language models](#). *Preprint*, arXiv:2402.00253.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

- Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. [Paying more attention to image: A training-free method for alleviating hallucination in LVLMs](#). In *Computer Vision – ECCV 2024*, volume 15141 of *Lecture Notes in Computer Science*, pages 125–140. Springer.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. [Progress measures for grokking via mechanistic interpretability](#). In *International Conference on Learning Representations*.
- Suzie Petryk, David Chan, Anish Kachinthaya, H. Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. 2024. [Aloha: A new measure for hallucination in captioning models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 342–357. Association for Computational Linguistics.
- Xiaoye Qu, Qiyuan Chen, Wei Wei, Jiashuo Sun, Daizong Liu, and Jianfeng Dong. 2025. [Alleviating hallucination in large vision-language models with active retrieval augmentation](#). *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(9):258:1–258:22.
- 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.
- Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. [Mitigating hallucinations in vision-language models through image-guided head suppression](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12481–12500, Suzhou, China. Association for Computational Linguistics.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: A circuit for indirect object identification in GPT-2 small](#). In *International Conference on Learning Representations*.
- Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. [Logical closed loop: Uncovering object hallucinations in large vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6944–6962, Bangkok, Thailand. Association for Computational Linguistics.
- Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. [V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273, Miami, Florida, USA. Association for Computational Linguistics.
- Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. [Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention](#). In *International Conference on Learning Representations*, volume 2025, pages 51546–51568.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024a. A survey on multimodal large language models. *National Science Review*, 11(12).
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024b. [Woodpecker: hallucination correction for multimodal large language models](#). *Science China Information Sciences*, 67(12).
- Chung-En (Johnny) Yu, Brian Jalaian, and Nathaniel D. Bastian. 2024. [Mitigating large vision-language model hallucination at post-hoc via multi-agent system](#). *Proceedings of the AAAI Symposium Series*, 4(1):110–113.
- Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024. [Reflective instruction tuning: Mitigating hallucinations in large vision-language models](#). In *Computer Vision – ECCV 2024*, volume 15126 of *Lecture Notes in Computer Science*, pages 196–213. Springer.
- Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2025. [Mitigating object hallucination in large vision-language models via image-grounded guidance](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 77461–77486. PMLR.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. [MiniGPT-4: Enhancing vision-language understanding with advanced large language models](#). In *International Conference on Learning Representations*.

A Experimental Setup and Reproducibility Details

A.1 Attribution and Evaluation Protocol Details

Head-selection thresholds and aggregation.

We apply quantile-based thresholds to the InfoSpectralScore distribution. In aggregated attribution, hallucination-prone candidates are selected from the outer tails, faithful candidates from a central interval around the median, and only heads selected consistently across multiple attribution subsets and rounds are retained. In the per-input task-specific pipeline, candidates are selected directly from the input-specific score distribution using a fixed rule: hallucination-prone candidates lie below the 0.1 quantile or above the 0.9 quantile, whereas faithful candidates lie in the 0.4–0.6 median-centered interval.

Attribution-set sizes. For the aggregate attribution experiments in Section 4.2, we use fixed attribution-set sizes of $|\mathcal{D}| = 160$ for each LLaVA-7B/13B layer group and $|\mathcal{D}| = 320$ for each Shikra-7B layer group. These values were selected during preliminary calibration to balance attribution stability and computational cost. Specifically, we chose the smallest operating points that produced stable faithful/hallucinatory head partitions across repeated aggregated-attribution rounds for the corresponding model family. The resulting sizes were fixed before downstream pruning evaluation and were not selected using CHAIR or F1 outcomes.

Controlled evaluation splits. For each model and layer group, the Hallucination Set and Generalization Set are sampled to match the size of the corresponding Validation Set. For merged groups, each split is formed by the de-duplicated union of the corresponding primary groups.

A.2 Method-Specific Configuration Notes

All experiments are conducted with fixed random seeds.

Vanilla. Vanilla denotes the original LVLm inference without any hallucination-control intervention. Unless otherwise specified, we keep the default model parameters unchanged and use greedy decoding with $\text{max_new_tokens}=256$ for long-form captioning experiments and $\text{max_new_tokens}=16$ for POPE.

OPERA. We evaluate OPERA using the official implementation and the hyperparameter settings recommended by the authors. Concretely, for LLaVA-1.5 on the CHAIR benchmark, we use $\text{beam}=5$, $\text{scale_factor}=50$, $\text{threshold}=15$, $\text{num_attn_candidates}=5$, and $\text{penalty_weights}=1$, with all remaining parameters kept at their default values in the released codebase. The same configuration is consistently applied to Shikra and to OPERA evaluations on POPE.

SPIN. For LLaVA-1.5 on CHAIR evaluation, the configuration reported in the SPIN paper ($\text{start_layer}=0$, $\text{end_layer}=32$) remains stable and we use it without modification. However, applying the same configuration to POPE causes severe degradation of generation quality. The original paper provides no task-specific guidance, and we therefore explored several alternative layer ranges. Among these tested configurations, we report results using $\text{start_layer}=5$ and $\text{end_layer}=32$, which consistently yielded the most stable and competitive performance on POPE. A similar phenomenon is observed for Shikra, where the 0–32 intervention led to unusable outputs. Following controlled exploration, we adopted the configuration used in our final results: $\text{start_layer}=5$, $\text{end_layer}=32$, $\text{routed_heads}=0.45$, $\text{small_num_mask}=0.001$.

PAI. For LLaVA-1.5 on the CHAIR benchmark, we follow the configuration described in the original paper, using interventions on layers 2–32 together with $\alpha=0.5$, $\gamma=1.1$, and classifier-free attention. On POPE, however, applying the same layer span leads to noticeable instability. Because the paper does not provide a recommended setting for POPE, we carried out a moderate search over several candidate layer ranges and chose layers 6–20, which showed the most reliable performance among the tested options. All other hyperparameters remain aligned with the CHAIR configuration.

VCD. We strictly follow the hyperparameter settings reported in the original VCD paper and its official implementation. Specifically, we use $\gamma = 0.1$, $\alpha = 1.0$, and $\beta = 0.1$ in all experiments. For the CHAIR benchmark, the number of noise steps is set to $T = 500$. For the POPE benchmark, we use $T = 999$. All remaining parameters are kept consistent with the released codebase.

VHR. All experiments with VHR strictly follow the repository-provided scripts and default hyperparameters for both POPE and CHAIR, without modification.

MLIH. We reproduce MLIH using the hyperparameters provided in the paper: for LLaVA-1.5, we use guide-range=5–18 and alpha=0.5; for Shikra, we use guide-range=3–13 and alpha=0.55. These intervention ranges naturally align with those used in our own method, ensuring fair and architecture-consistent comparison.

A.3 Adjustments for Qwen2.5-VL-Instruct

Compared with LLaVA and Shikra, Qwen2.5-VL-Instruct is more expressive in its unmodified form but also exhibits less stable decoding under loosely constrained prompts or layer-level interventions. To ensure fair comparison and generate usable captions across all methods, we introduce the following minimal and model-specific adjustments.

Prompt adaptation. Most models (e.g., LLaVA and Shikra) are evaluated using the default prompt: “Please describe the image in detail.” However, Qwen2.5-VL-Instruct exhibits noticeably more unstable generations under this loose instruction, often producing overly long, drifting, or structurally irregular text. To stabilize decoding and maintain comparability across methods, we adopt the following controlled prompt:

“Describe the image in a single natural, fluent paragraph. Use complete sentences with a smooth narrative flow, as a human would. Avoid list-like structures, numbered descriptions, rigid templates, bullet points, or headings. Include a reasonable amount of detail, similar to a typical detailed caption.”

This prompt provides mild structural guidance without enforcing rigid templates or explicit content enumeration, ensuring stable and comparable generations across all evaluated methods.

Architectural alignment and layer mapping.

Several competing methods specify intervention layers only for LLaVA (e.g., layers 2–31). To preserve comparable relative depth, we map these ranges to Qwen2.5 by aligning proportional layer indices (e.g., 2–27). For methods that operate on middle layers (ours and MLIH), we follow prior analyses of Qwen architectures (Qwen-LA (Chu et al., 2025)) and adopt:

TARGET_LAYERS = [12, 13, 14, 15, 16,
18, 19, 20, 22, 27],

which covers layers 12–16, 18–22, and 27. Layer 21 is excluded from modification due to instability but retained as a probing layer for token extraction.

Other adjustments. All remaining hyperparameters (decoding strategy, max tokens, safety checks, and batch configuration) are kept identical to those used for LLaVA and Shikra, unless a method fails to run under the common configuration. In such rare cases, we introduce the smallest possible fix required for successful execution without changing the method’s effective behavior.

B Implementation and Pipeline Engineering

B.1 Key Engineering Improvements for the Task-Specific Pruning Pipeline

Accelerated Head-Level Attribution via Cached Semantic Reconstruction

A direct implementation of Algorithm 1 would require running the LVLM once for every head ablation in every target layer, resulting in hundreds of full forward passes per attribution set. This is prohibitively expensive for modern LVLMs. To address this, we adopt a fast approximation that performs *only one* unmodified-model generation pass and caches, for each target layer:

- the residual stream (final token) before attention,
- the concatenated multi-head output (final token) *before the output projection* W_o ,
- the post-attention semantic embedding of the final token.

Because the post-attention update for each head is linear before the post-attention layer normalization, the effect of ablating a head can be reconstructed offline. For a cached layer l , let $\mathbf{h}_j^{(l)} \in \mathbb{R}^{d_h}$ be the output of head j , and let $W_o^{(l)}$ be the output projection. Under mean-replacement, the reconstructed semantic embedding for head j is

$$\tilde{\mathbf{x}}_{\text{sem}}^{(l)} = \mathbf{x}_{\text{resid}}^{(l)} + W_o^{(l)} (\text{concat}(\mathbf{h}_1^{(l)}, \dots, \bar{\mathbf{h}}^{(l)}, \dots, \mathbf{h}_H^{(l)})), \quad (13)$$

where $\bar{\mathbf{h}}^{(l)} = \frac{1}{H} \sum_{k=1}^H \mathbf{h}_k^{(l)}$ is the per-sample mean head output at layer l , followed by the same post-attention layer normalization used in the model.

This enables the computation of InfoSpectralScore for all heads via Eqs. (4)–(6) without

invoking additional forward passes or generating captions.

Under this approximation, the semantic matrix Z used for InfoSpectralScore is constructed by stacking the reconstructed post-attention semantic vectors $\tilde{\mathbf{x}}_{\text{sem}}^{(l)}$ of the final token across the K attribution samples, rather than the fully propagated final-token embeddings. Formally, for each layer l and ablated head j , we form $Z^{(l,j)} = [\tilde{\mathbf{x}}_{\text{sem}}^{(l)}(1), \dots, \tilde{\mathbf{x}}_{\text{sem}}^{(l)}(K)]^\top \in \mathbb{R}^{K \times d}$.

In practice, cached reconstruction reduces the attribution overhead to a small constant per instance (e.g., a few seconds on CHAIR and sub-second on POPE), enabling exhaustive head-by-head analysis without additional full forward passes.

Why this approximation is valid. The key observation is that the final-token semantic vector before the MLP block is a linear function of the concatenated head outputs:

$$\mathbf{x}_{\text{sem}}^{(l)} = \text{LN}\left(\mathbf{x}_{\text{resid}}^{(l)} + W_o^{(l)} \text{concat}\left(\mathbf{h}_1^{(l)}, \dots, \mathbf{h}_H^{(l)}\right)\right). \quad (14)$$

Thus, modifying only one head corresponds to a low-rank perturbation in the semantic space. Since InfoSpectralScore depends solely on the covariance structure of these semantic embeddings, we can compute the change in spectral statistics directly from reconstructed embeddings.

Empirically, we observed that this fast approximation yields head-importance rankings that closely track those produced by the full forward-regeneration procedure. More importantly, when incorporated into the broader analyses of attribution behavior and hallucination scenarios presented in Sections 4.3 and 4.4, the approximation continues to support effective hallucination mitigation. These empirical trends suggest that the cached-semantic representation preserves the discriminative signals required for reliable head attribution, while making instance-level adaptation computationally feasible in practice.

Stabilized and Efficient Bayesian Optimization for Per-Instance Pruning

The task-specific pruning pipeline (Algorithm 3) uses a lightweight Bayesian optimization (BO) loop to determine the pruning parameters (μ, λ) for each input instance. Although the search space is only two-dimensional, each evaluation requires full LVLM decoding, making runtime the primary

computational bottleneck. To ensure that BO remains both stable and efficient in this setting, we introduce several practical refinements.

Warm-start initialization combined with random exploration. To accelerate convergence while preserving exploration, we warm-start Bayesian optimization with a small set of empirically strong configurations identified from preliminary development runs. Specifically, we initialize BO with four fixed seed points in the (μ, λ) space: $(0.65, 0.99)$, $(0.75, 0.99)$, $(0.90, 0.99)$, and $(0.99, 0.99)$. These seeds are shared across all models and experiments and are only used to initialize the optimizer. In addition, we include a few randomly sampled points to maintain exploratory diversity. This initialization strategy places the optimizer near effective regions of the pruning space while still allowing it to explore previously unvisited areas.

Constrained search region. We restrict the optimization domain to

$$\mu \in [0.62, 0.99], \quad \lambda \in [0.80, 0.99],$$

which substantially reduces the variance of early BO iterations while covering the practically relevant operating range.

Robust handling of irregular LVLM generations.

Under our dynamic pruning method, LVLM decoding may occasionally become unstable and produce malformed outputs, such as truncated sentences, incoherent token sequences, or repeated prefix fragments. If these outputs were treated as valid optimization samples, they could distort the BO surrogate model and impede stable convergence. To avoid this, we apply lightweight validity checks that examine output length, character-level anomalies, and prefix-overlap patterns. Outputs that fail these checks receive a soft penalty rather than being removed, which keeps the surrogate model reliable while maintaining smooth optimization progress for each instance.

Early stopping based on difficult-to-improve hallucination gains.

Extensive empirical analysis shows that some inputs yield consistently non-positive hallucination improvements across a broad range of (μ, λ) settings. This behavior typically arises when the pruning adjustment causes unstable decoding, producing borderline or partially degraded text, or when the objective surface becomes effectively flat with respect to the pruning

parameters and no meaningful improvement can be reached. In these situations, continued exploration brings little benefit. Therefore, the BO loop is terminated early once several consecutive iterations show no improvement or fall below a small gain threshold, preventing excessive and unproductive evaluations on instances that are intrinsically difficult to optimize.

Fallback retry for atypical attribution distributions. If the primary BO phase fails to identify an improving configuration, we activate a short retry stage. This stage recomputes hallucination heads using a robust interquartile-range (IQR) rule and runs a small BO pass with freshly initialized priors. The fallback mechanism compensates for unusually flat attribution distributions and ensures that the pruning pipeline remains reliable even on challenging inputs.

Given that head-level pruning inherently incurs additional decoding passes, this runtime remains practical for per-instance hallucination mitigation.

B.2 Computational cost

Hardware and runtime environment. We measure runtime on NVIDIA RTX 5090 GPUs with 32 GB memory and NVIDIA L20 GPUs with 48 GB memory. Unless otherwise noted, each individual run fits on a single RTX 5090 GPU. The only exception is OPERA with beam size 5, which is evaluated on a single L20 GPU because its beam-search decoding has a larger memory footprint. The wall-clock runtimes reported in Tables 7 and 8 are measured under this hardware environment.

Tables 7 and 8 summarize the end-to-end runtime of all competing methods on CHAIR (1000 COCO multi-label images, max_new_tokens=256) and POPE (9000 binary questions, max_new_tokens=16). Our method introduces two additional computations per instance: (i) head-level attribution, and (ii) the Bayesian search over (μ, λ) during counteractive pruning.

Attribution cost. As defined in Algorithm 1, attribution requires evaluating each attention head on an attribution set of size K . With the cached semantic reconstruction mechanism (Appendix B.1), this step becomes lightweight, incurring approximately:

- **5 seconds per instance** on CHAIR, due to long-form caption generation;

- **0.5 seconds per instance** on POPE, where responses are short.

Overall runtime and scalability. Both attribution and counteractive pruning operate on a per-instance basis and do not require any dataset-level preprocessing or accumulation. As a result, the per-instance overhead remains constant, and the overall runtime grows linearly with the number of evaluated samples.

On CHAIR, long-form caption generation (256 tokens) increases latency, as counteractive pruning requires multiple forward passes under different (μ, λ) configurations. As a result, the full configuration of our method still incurs additional overhead on this benchmark, but its updated runtime is lower than OPERA (17.34 vs. 20.12 s/img) and substantially faster than the previous measurement. The fixed variant disables per-instance hyperparameter search and remains a lighter deployment-oriented operating point with reduced computational overhead. In contrast, the full configuration retains instance-level adaptation through Bayesian optimization, incurring additional decoding cost. Overall, the observed runtime remains within a practically acceptable range for scenarios that require instance-level hallucination mitigation while maintaining semantic richness.

On POPE, which involves short binary answers (16 tokens), our method remains within the same order of magnitude as all compared approaches. Even with the additional attribution step (about 0.5 s per instance), the overall end-to-end runtime stays moderate and does not introduce prohibitive overhead. These results indicate that the method is suitable for practical applications requiring adaptive, instance-level hallucination control.

C Additional Analyses and Supporting Results

C.1 Complete MME Results

Tables 9 and 10 provide the task-level breakdown of the MME results. On perception tasks, the gains are mainly observed in visually grounded categories such as existence, count, position, posters, celebrity, scene, landmark, artwork, and OCR, while color remains unchanged. On cognition tasks, the method improves common-sense reasoning, numerical calculation, and code reasoning, whereas text translation remains unchanged. These trends are consistent with the main-paper observation that head-level counteractive pruning primarily benefits visu-

Table 7: Runtime on CHAIR evaluation with 1000 COCO multi-label images. All time entries are in **hh:mm:ss**.

Method	Vanilla	SPIN	PAI	OPERA	VCD	VHR	MLIH	Ours	Ours-fixed
Total Time	00:33:00	00:35:30	00:50:40	05:35:20	00:33:47	00:33:40	00:35:40	04:49:00	01:54:44
Avg Time (s/img)	1.98	2.13	3.04	20.12	2.03	2.02	2.14	17.34	6.88

Table 8: Runtime on POPE evaluation. The upper block reports runtime on the three POPE subsets (random / popular / adversarial). The lower block reports the total runtime and the average time per question (total time divided by 9000). All time entries are in **hh:mm:ss**.

Method	Vanilla	SPIN	PAI	OPERA	VCD	VHR	MLIH	Ours	Ours-fixed
Subset Runtime									
Random	00:13:17	00:06:05	00:22:19	01:04:57	00:08:17	00:09:42	00:10:57	01:07:04	00:08:17
Popular	00:13:21	00:06:55	00:21:48	01:06:30	00:08:16	00:09:40	00:10:55	01:08:52	00:08:33
Adversarial	00:14:09	00:06:17	00:21:27	01:06:10	00:08:17	00:09:41	00:11:21	01:13:21	00:09:07
Overall Runtime									
Total Time	00:40:48	00:19:17	01:05:34	03:17:37	00:24:50	00:29:03	00:33:13	03:29:17	00:25:57
Avg Time (s/question)	0.272	0.128	0.437	1.317	0.166	0.194	0.221	1.395	0.173

ally grounded and hallucination-related behaviors, while producing more limited changes on tasks less directly tied to object-level visual grounding.

C.2 Caption-Level ALOHa Evaluation

We employ ALOHa (Petryk et al., 2024) to further evaluate caption-level object grounding beyond the fixed COCO object vocabulary. ALOHa extracts groundable object mentions from generated captions and assesses whether these mentions are supported by reference objects and localized visual evidence, making it suitable for open-vocabulary hallucination evaluation at the caption level. A higher ALOHa score indicates that generated object mentions are better supported by visual evidence and contain fewer unsupported object mentions.

We conduct this evaluation on the LLaVA-1.5-7B outputs from the same 1000-image multi-label COCO subset used in Table 4. The evaluation compares the same 1000 responses produced by Vanilla inference with the same 1000 responses produced by HACP. We use the DETR-mode setting of ALOHa, which incorporates detector-based localized object evidence for matching object mentions.

As shown in Table 11, HACP improves the ALOHa score from 42.25 to 44.23, yielding an absolute gain of +1.98 and a relative improvement of approximately +4.7%. This result focuses on object-mention grounding in an open-vocabulary setting: beyond the fixed COCO vocabulary used by CHAIR, object mentions generated by HACP are more often supported by localized visual evidence. The ALOHa result therefore provides complementary evidence that HACP improves the visual grounding of generated object mentions and

extends the evaluation beyond hallucination statistics computed under a fixed object vocabulary.

C.3 GPT-4V-aided Evaluation

We conduct a GPT-4V-aided evaluation to complement automatic hallucination metrics with a pairwise assessment of generated captions. Following the GPT-4V-aided evaluation protocol used in recent hallucination studies (Yin et al., 2024b), GPT-4V is asked to compare two assistant responses for the same image and instruction. The judge rates each response along two axes: Accuracy, which measures how faithfully the response describes the image content, and Detailedness, which measures the richness of necessary details. Both criteria are scored on a 1–10 scale, where higher is better, and hallucinated content is penalized rather than counted as useful detail.

We perform this evaluation on the same LLaVA-1.5-7B outputs from the 1000-image multi-label COCO subset used in Table 4. Specifically, the evaluation compares the 1000 responses produced by Vanilla inference with the 1000 responses produced by HACP. For each sample, GPT-4V is provided with the image, the original instruction, the Vanilla response, and the HACP response, and returns the two scores according to the prompt template in Table 13.

As shown in Table 12, GPT-4V assigns HACP a higher Accuracy score than Vanilla, improving from 5.38 to 5.83 with an absolute gain of +0.45. Detailedness also increases slightly from 5.39 to 5.48, yielding a gain of +0.09. This judge-based comparison complements the automatic metrics by directly assessing caption faithfulness and necessary detail. The higher Accuracy together with the

Table 9: Results on all MME perception-related tasks. Best results are in **bold**.

Method	Existence	Count	Position	Color	Posters	Celebrity	Scene	Landmark	Artwork	OCR	Perception Total
Vanilla	185.00	138.33	108.33	155.00	152.04	130.29	163.00	143.00	102.00	97.50	1374.50
Ours	190.00	145.00	120.00	155.00	184.69	147.94	179.00	146.75	116.00	102.50	1486.89

Table 10: Results on all MME cognition-related tasks. Best results are in **bold**.

Method	Common Sense Reasoning	Numerical Calculation	Text Translation	Code Reasoning	Cognition Total
Vanilla	102.14	60.00	87.50	55.00	304.64
Ours	114.29	90.00	87.50	72.50	364.29

Table 11: Caption-level ALOHa evaluation on the same 1000-image multi-label COCO subset used in Table 4. Higher scores indicate better visual support for generated object mentions.

Method	ALOHa $\uparrow$
Vanilla	42.25
Ours	44.23

Table 12: GPT-4V-aided caption-level evaluation on the same 1000-image multi-label COCO subset used in Table 4. Scores are on a 1–10 scale, and higher is better.

Method	Accuracy $\uparrow$	Detailedness $\uparrow$
Vanilla	5.38	5.39
Ours	5.83	5.48

slight Detailedness gain indicates that HACP reduces unsupported visual claims while retaining informative content and necessary visual detail.

D InfoSpectralScore Validation and Hyperparameter Studies

D.1 Motivation and Intervention-Based Validation

InfoSpectralScore is designed as a semantics-oriented attribution signal for changes induced by head-level perturbations. EigenScore captures the dominant semantic variance of the perturbed representations, SpectralVar penalizes overly anisotropic spectra, and SpectralEntropy discourages variance from concentrating in only a few directions. Intuitively, heads that strongly affect semantic structure under perturbation can be informative for hallucination control, whereas anisotropic or overly concentrated spectra indicate less robust semantic behavior under intervention.

We further evaluate this attribution signal through targeted interventions in the 1000-image multi-label COCO setting used in Table 4. In addition to Vanilla and HACP, we compare a random-

head control, which applies intervention without InfoSpectralScore-based selection, and a reversed-direction control, which reverses the prescribed suppression of hallucination-prone heads and reinforcement of faithful heads.

Relative to Vanilla, HACP reduces C_s/C_i from 66.7/18.0 to 43.4/8.6 and improves F1 from 73.0 to 77.0, whereas neither control reproduces this joint improvement. The random-head result shows that intervention on arbitrary heads is insufficient to explain the gain, while the reversed-direction result shows that the attributed heads do not remain effective when their prescribed suppression/reinforcement directions are reversed. Together, these controls provide intervention-based evidence that HACP’s gains depend on combining InfoSpectralScore-based head selection with the intended counteractive intervention. They complement the weight-sensitivity analysis below, where $\alpha = \gamma = 0$ reduces InfoSpectralScore to EigenScore alone and all tested settings yield positive hallucination-reduction gains.

D.2 Sensitivity of InfoSpectralScore Regularization Weights

Due to the absence of ground-truth annotations for hallucination-inducing attention heads, we perform hyperparameter sensitivity analysis using a proxy evaluation framework based on CHAIR metrics. Our goal is to examine whether attention heads selected under different InfoSpectralScore regularization weights consistently suppress hallucinations when pruned.

We conduct this analysis on the LLaVA-13B model, focusing on attention layers 5 through 7. We first perform inference on 300 samples using the original model and apply an automatic CHAIR-based evaluator to identify hallucinated outputs. From these, we randomly select 32 hallucinated samples as the evaluation subset. For

Table 13: Prompt template for GPT-4V-aided evaluation. {question} is the original instruction; {answer 1} is the Vanilla response; and {answer 2} is the response generated by HACP.

Prompt template for GPT-4V-aided evaluation

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image.

Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:

1. **Accuracy**: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.

2. **Detailedness**: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.

Please output a single line for each criterion, containing only two values indicating the scores for Assistant 1 and Assistant 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Question: {question}

Assistant 1: {answer 1}

Assistant 2: {answer 2}

Output format:

Accuracy:

Scores of the two answers:

Reason:

Detailedness:

Scores of the two answers:

Reason:

each hyperparameter configuration, we compute InfoSpectralScore-based consistency scores for all attention heads in the selected layers, aggregating over 15 stochastic generations per head. Hallucination-inducing heads are then selected from the extreme ends of the consistency score distribution using a quantile-based strategy and pruned in each layer (e.g., 20 heads per layer out of 40 total). Model performance before and after pruning is evaluated on the hallucinated samples using the CHAIR-i and CHAIR-s metrics.

We quantify hallucination suppression using

$$\Delta_h = (bh_i - ph_i) + (bh_s - ph_s), \quad (15)$$

where bh_i, bh_s denote pre-pruning CHAIR-i/s scores on hallucinated samples and ph_i, ph_s denote the corresponding scores after pruning. A larger Δ_h

Table 14: Random-head and reversed-direction controls for InfoSpectralScore-based attribution. All methods are evaluated in the same 1000-image multi-label COCO setting as Table 4.

Intervention	$C_s \downarrow$	$C_i \downarrow$	Recall $\uparrow$	Precision $\uparrow$	F1 $\uparrow$	Len
Vanilla	66.7	18.0	69.1	77.3	73.0	104.79
Ours	43.4	8.6	70.8	89.0	77.0	103.00
Random heads	67.7	17.7	68.5	77.0	72.5	102.60
Reversed direction	69.5	18.0	68.6	76.5	72.4	103.04

Table 15: Grid search of Δ_h over α and γ . Top-3 values are in **bold**.

$\alpha \backslash \gamma$	0.0	0.5	1.0	1.5	2.0
0.0	0.0470	0.0332	0.0332	0.0332	0.0332
0.1	0.0320	0.0279	0.0373	0.0332	0.0332
0.2	0.0320	0.0390	0.0279	0.0373	0.0373
0.3	0.0320	0.0257	0.0670	0.0279	0.0259
0.4	0.0320	0.0257	0.0390	0.0279	0.0279
0.5	0.0320	0.0724	0.0705	0.0390	0.0279

indicates more effective suppression of hallucinations induced by the pruned attention heads.

The grid search results in Table 15 show that Δ_h remains positive across all tested hyperparameter configurations, ranging from approximately 0.026 to 0.072. In this proxy evaluation setting, this trend indicates that InfoSpectralScore provides a stable signal for selecting hallucination-associated heads and is not highly sensitive to exact hyperparameter choices within the tested grid. Moreover, the stronger results are concentrated in a small region of the (α, γ) space. In particular, the configurations (0.3, 1.0), (0.5, 0.5), and (0.5, 1.0) achieve the largest Δ_h values within the tested grid.

Finally, we observe that larger α values are often associated with higher Δ_h in this setting, while γ exhibits a comparatively milder effect over the tested range.

D.3 Hyperparameter Validation for Dynamic Counterbalance

To validate the robustness of our dynamic counterbalance mechanism, we conduct an extensive hyperparameter search over the modulation strength μ and smoothing factor λ , which govern the magnitude and temporal stability of per-head activation adjustment.

Experimental Setup. All experiments are conducted on LLaVA-7B, with dynamic counteractive pruning applied to layers 5–18 of the unmodified model. The sets of hallucination-inducing and faithful heads are identified in advance using attribution on a dedicated 160-image subset and kept fixed

Table 16: Line search over μ with fixed $\lambda = 0.95$. Values are relative improvements (%) on CHAIR-s, CHAIR-i, and F1.

μ	λ	CHAIR-s (%)	CHAIR-i (%)	F1 (%)
0.20	0.95	-2.70	-0.84	0.49
0.25	0.95	0.00	-1.30	0.74
0.30	0.95	-4.05	-6.94	1.74
0.35	0.95	1.35	-5.20	0.93
0.40	0.95	2.70	-3.63	-0.41
0.45	0.95	0.00	-8.50	0.90
0.50	0.95	-2.70	-9.22	0.52
0.55	0.95	0.00	-7.53	-0.03
0.60	0.95	0.00	-9.25	1.13
0.65	0.95	0.00	-8.15	0.36

across all runs. Both counteractive pruning and evaluation are then carried out on this same subset, providing a consistent basis for hyperparameter validation.

We first fix $\lambda = 0.95$ and perform a line search over $\mu \in [0.2, 0.65]$ with a step size of 0.05, as shown in Table 16. Subsequently, we conduct a two-dimensional grid search with $\mu \in \{0.3, 0.5, 0.55, 0.60, 0.65\}$ and $\lambda \in \{0.93, 0.95, 0.97, 0.99, 1.0\}$ (Table 17). Each configuration is evaluated using the hook-based intervention (Section 3.3). CHAIR-s, CHAIR-i, and F1 are reported as relative changes (%) with respect to the unmodified model.

Results. Across the evaluated configurations, different (μ, λ) settings exhibit distinct trade-offs between hallucination reduction and output quality. Within the tested grid, $(\mu = 0.3, \lambda = 0.97)$ corresponds to the largest reduction in CHAIR-s, while $(\mu = 0.60, \lambda = 0.95)$ yields the largest reduction in CHAIR-i and also improves F1.

Selection Across All Model Groups. For completeness, the optimal settings of (μ, λ) for each model, group, and evaluation split are provided in Table 18. These values are determined based on the grid search results above and subsequent validation experiments across the full model suite.

D.4 Fixed-Configuration Sensitivity for *Ours-fixed*

We report additional results for *Ours-fixed* under fixed hyperparameter configurations on the same 1000-image multi-label COCO subset used in Table 4. Each configuration applies the identical task-specific pruning pipeline as *Ours*, while keeping (μ, λ) fixed across all images and disabling per-instance hyperparameter search.

Table 17: Grid search over μ and λ . Values are relative improvements (%) on CHAIR-s, CHAIR-i, and F1.

μ	λ	CHAIR-s (%)	CHAIR-i (%)	F1 (%)
0.30	0.93	-1.35	-1.16	1.38
0.30	0.95	-4.05	-5.66	1.42
0.30	0.97	-5.41	-4.00	1.34
0.30	0.99	-2.70	-2.48	1.21
0.30	1.00	-1.35	-1.22	1.23
0.50	0.93	-2.70	-9.34	0.60
0.50	0.95	-2.70	-9.98	0.51
0.50	0.97	-2.70	-9.51	0.67
0.50	0.99	-1.35	-7.87	0.09
0.50	1.00	-1.35	-7.55	0.15
0.55	0.93	0.00	-8.62	-0.19
0.55	0.95	0.00	-7.57	0.03
0.55	0.97	0.00	-8.52	0.40
0.55	0.99	-1.35	-9.99	0.11
0.55	1.00	-1.35	-8.54	-0.16
0.60	0.93	0.00	-9.34	0.28
0.60	0.95	0.00	-11.03	0.89
0.60	0.97	0.00	-11.37	0.83
0.60	0.99	0.00	-11.61	0.61
0.60	1.00	0.00	-11.13	0.98
0.65	0.93	0.00	-9.87	1.13
0.65	0.95	0.00	-8.20	0.63
0.65	0.97	-1.35	-12.87	1.13
0.65	0.99	-2.70	-12.83	1.43
0.65	1.00	0.00	-11.02	1.00

Table 19 summarizes several representative fixed configurations within the tested range. The configuration $(\mu, \lambda) = (0.95, 0.99)$ is reported as *Ours-fixed* in Table 4.

E Qualitative Case Study

To provide a qualitative comparison among HACP, VHR, and PAI, we present captioning examples generated by LLaVA-1.5-7B. Consistent with Section 4.3, the examples are drawn from a curated set of 500 COCO validation images whose Vanilla captions contain hallucinated object mentions ($C_s = 1$).

Figure 3 shows three representative cases, comparing captions from Vanilla, VHR, PAI, and HACP. Real object mentions are highlighted in blue, and hallucinated objects in red. All captions are generated with the prompt: “Please describe the image in detail.”

In the first two examples, VHR produces much shorter captions than the other methods and omits several visible details. A similar pattern appears in the third, more complex scene, where the VHR caption remains brief and less informative. These cases illustrate the conservative behavior of VHR, which can reduce hallucinations at the cost of descriptive coverage.

Table 18: Best hyperparameter settings (μ, λ) for each model group and evaluation set.

Model: LLaVA-7B				Model: LLaVA-7B-hallu			
Group	Evaluation Set	μ	λ	Group	Evaluation Set	μ	λ
[5,18]	Validation	0.60	0.99	[5,18]	Validation	0.60	0.95
[5,18]	Hallucination	0.65	0.95	[5,18]	Hallucination	0.60	0.95
[5,18]	Generalization	0.65	0.95	[5,18]	Generalization	0.65	0.90
[19,26]	Validation	0.65	0.90	[19,26]	Validation	0.65	0.99
[19,26]	Hallucination	0.65	0.99	[19,26]	Hallucination	0.60	0.99
[19,26]	Generalization	0.65	0.97	[19,26]	Generalization	0.35	0.97
Merged	Validation	0.62 / 0.65	0.80 / 0.97	Merged	Validation	0.65 / 0.60	0.90 / 0.99
Merged	Hallucination	0.60 / 0.70	0.99 / 0.90	Merged	Hallucination	0.65 / 0.65	0.95 / 0.97
Merged	Generalization	0.62 / 0.65	0.97 / 0.97	Merged	Generalization	0.60 / 0.60	0.90 / 0.99

Model: LLaVA-13B				Model: Shikra-7B			
Group	Evaluation Set	μ	λ	Group	Evaluation Set	μ	λ
[5,18]	Validation	0.65	0.97	[3,13]	Validation	0.30	0.95
[5,18]	Hallucination	0.60	0.85	[3,13]	Hallucination	0.75	0.90
[5,18]	Generalization	0.55	0.97	[3,13]	Generalization	0.67	0.90
[19,26]	Validation	0.50	0.99	[14,28]	Validation	0.60	0.80
[19,26]	Hallucination	0.65	0.97	[14,28]	Hallucination	0.75	0.99
[19,26]	Generalization	0.55	0.80	[14,28]	Generalization	0.60	0.90
Merged	Validation	0.50 / 0.99	0.90 / 0.97	Merged	Validation	0.30 / 0.75	0.95 / 0.97
Merged	Hallucination	0.60 / 0.65	0.85 / 0.80	Merged	Hallucination	0.67 / 0.60	0.90 / 0.80
Merged	Generalization	0.60 / 0.65	0.80 / 0.80	Merged	Generalization	0.75 / 0.60	0.90 / 0.90

Table 19: Fixed-configuration sensitivity of *Ours-fixed* on the 1000-image multi-label COCO subset. The selected configuration is in **bold**.

μ	λ	$C_s \downarrow$	$C_i \downarrow$	Recall $\uparrow$	Precision $\uparrow$	F1 $\uparrow$	Len
0.95	0.99	58.1	12.7	70.1	83.9	74.5	103.84
0.99	0.95	58.2	12.8	69.9	83.8	74.4	103.72
0.99	0.99	58.6	12.9	70.1	83.6	74.3	103.84
0.90	0.90	57.9	12.9	70.1	83.8	74.4	103.98

PAI shows a different failure mode in Examples 1 and 3: its captions contain repeated phrasing and similar sentence structures. This repetition makes the captions less informative, even when some hallucinated objects are removed.

HACP reduces the highlighted hallucinated object mentions while retaining more correct object details in these examples. Its captions remain comparatively descriptive, suggesting a better balance between hallucination mitigation and caption informativeness in the illustrated cases.



Figure 3: Qualitative results of hallucination mitigation on LLaVA-1.5-7B. Real and hallucinated object words are marked in blue and red, respectively. The prompt is “Please describe the image in detail.”