HCA-EDIT: HYBRID CATEGORY-AWARE EDITING FOR VISUAL HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

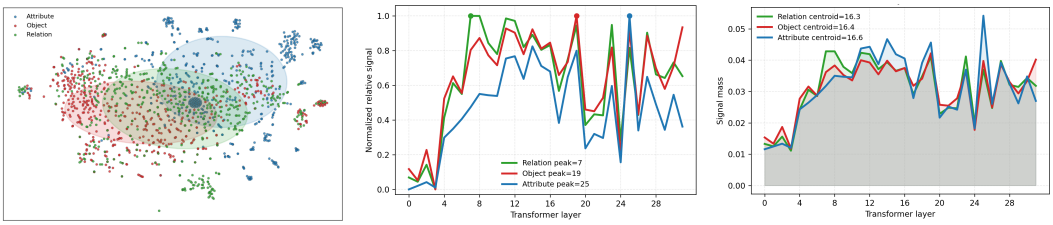
Activation editing offers an efficient, training-free strategy to mitigate visual hallucinations in large vision-language models (LVLMs). However, visual hallucinations span distinct semantic categories-specifically object, attribute, and relation errors. Estimating a single correction direction from mixed hallucination examples obscures category-specific information and favors categories with larger activation magnitudes. Furthermore, applying a uniform editing strength across layers overlooks the variation of these category signals across network depth. To address these limitations, we propose HCA-Edit (Hybrid Category-Aware Editing), a training-free framework that combines category-balanced direction construction with category-aware layer gating. Using paired calibration data constructed through targeted semantic perturbations, HCA-Edit estimates correction directions separately for object, attribute, and relation hallucinations, normalizes each direction within each layer, and combines them with equal weights. This balances category contributions independently of their raw direction magnitudes. To allocate editing strength across depth, we normalize category-specific activation-difference signals across layers and take their maximum across categories, retaining strong responses from any category. The resulting shared gate assigns continuous, layer-specific editing strengths while preserving the balanced category mixture. Extensive experiments across multiple LVLMs and benchmarks demonstrate that HCA-Edit effectively reduces visual hallucinations while maintaining general visual understanding capabilities.

1 INTRODUCTION

Large vision-language models (LVLMs) (Bai et al., 2025b; Liu et al., 2023; Ye et al., 2023; Liu et al., 2024b; Bai et al., 2023; 2025a; Dai et al., 2023; Zhu et al., 2024) have become increasingly capable of describing images, answering visual questions, and reasoning about multimodal inputs. Yet fluent responses do not always remain faithful to the image. A model may introduce an absent object, assign an incorrect attribute to a visible entity, or describe a relation unsupported by the image (Rohrbach et al., 2018; Wang et al., 2023; Rawte et al., 2023; Kalai et al., 2025). These errors span distinct hallucination categories, namely object, attribute, and relation hallucinations, and concern different aspects of visual grounding. Correcting what appears in an image is not necessarily the same as correcting what it looks like or how it relates to other objects.

Activation editing offers an efficient approach to hallucination mitigation without retraining (Liu et al., 2025; Yang et al., 2025; Dang et al., 2026). Its effectiveness depends on both the correction encoded in the intervention and the strength with which it is applied. These choices raise two questions. First, how should these different hallucination categories contribute to the correction? Estimating a single direction from pooled object, attribute, and relation errors makes their individual contributions difficult to control. Categories with larger activation differences may exert greater influence, although their magnitudes need not reflect their usefulness for correction. Second, how should the intervention be distributed across layers? A uniform editing strength does not account for variation in hallucination-related signals across depth. Recent work has addressed the choice of editing locations (Dang et al., 2026). We argue that category information should guide both direction construction and strength allocation: category contributions should be balanced before ensembling, while layer-specific signals should determine the strength of the resulting intervention.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Activation projection (b) Layer-wise signal (c) Signal distribution over depth

Figure 1: Category-conditioned activation differences on LURE-3C. (a) t-SNE projection of paired activation differences. (b) Layer-wise normalized signals. (c) Normalized signal distributions over depth. The categories exhibit overlapping distributions and similar overall depth trends, with differences in local responses. These observations motivate retaining category information during direction construction and layer gating.

To investigate these questions, we construct LURE-3C calibration pairs from LURE (Zhou et al., 2024) through targeted object, attribute, and relation perturbations of reference descriptions. For each category, we compare the activations elicited by the reference and perturbed descriptions under the same image. Figure 1 shows overlapping but non-identical distributions of the paired activation differences in the projection. Their layer-wise signals follow similar overall trends, but differ in local responses and peak locations. This combination of overlap and local variation motivates estimating correction directions separately by category before combining them into a shared intervention. It also suggests using signals from all three categories to determine how intervention strength should vary across layers.

Building on this analysis, we propose **HCA-Edit** (Hybrid Category-Aware Editing), a training-free framework that combines category-balanced direction construction with category-aware layer gating. We first estimate correction directions separately for object, attribute, and relation hallucinations using LURE-3C. Each direction is normalized within each layer before the three are combined with equal weights. This prevents raw direction magnitudes from determining the relative category weights, while retaining all three category directions in the ensemble. The resulting correction does not require a classifier to assign each input to a single hallucination category.

The layer gate determines how strongly this ensemble direction is applied across depth. We derive category-specific signals from paired language MLP outputs and normalize each category's signal across layers. Taking the maximum across categories then produces a shared depth profile. Unlike averaging, this operation retains a strong response from one category even when the others respond weakly at the same layer. We transform the profile into continuous editing coefficients for all layers. The gate thus adjusts the strength of the correction without changing its category mixture: direction construction controls category contribution, while layer gating controls depth-dependent intervention strength. Both the directions and coefficients are computed offline and remain fixed during inference.

Experiments across multiple LVLMs and benchmarks show that HCA-Edit reduces visual hallucinations while maintaining general visual understanding capabilities. Component ablations further support the joint use of category-balanced directions and category-aware layer gating.

Our main contributions are:

- We construct LURE-3C through targeted semantic perturbations of LURE reference descriptions and analyze category-conditioned activation differences, providing a calibration basis for category-aware editing.
- We propose **HCA-Edit**, which balances separately estimated category directions and uses category-aware layer gating to allocate editing strength across network depth without changing their mixture.
- We evaluate HCA-Edit across multiple LVLMs and benchmarks, showing competitive or improved performance, with ablation studies supporting the joint design of category-balanced direction construction and category-aware layer gating.

2 Related Work

2.1 Hallucination in vision-language models

Visual hallucination occurs when an LVLM generates claims that contradict or are not supported by the visual evidence in the input image. It includes references to absent objects, incorrect attributes of visible objects, and unsupported relations between them (Rohrbach et al., 2018; Wang et al., 2023; Yu et al., 2024). These errors can arise even when the response is fluent and plausible, making linguistic quality alone an unreliable indicator of visual grounding. Recent mitigation methods address the problem through changes to decoding, training, or internal model representations.

Decoding-based methods modify how the model selects tokens during generation. VCD contrasts predictions from the original image and a distorted version to reduce reliance on language priors (Leng et al., 2024). M3ID strengthens the influence of the image by favoring tokens with higher mutual information with the visual input (Favero et al., 2024). AGLA combines predictions informed by global image features and prompt-relevant local content to improve visual grounding (An et al., 2025). OPERA penalizes excessive attention to summary tokens and revisits earlier decoding decisions through a rollback mechanism (Huang et al., 2024).

Training-based methods improve visual grounding through instruction tuning and preference alignment. LRV-Instruction and HalluciDoctor address the quality and composition of visual instruction data (Liu et al., 2024a; Yu et al., 2024), while LLaVA-RLHF and HA-DPO use human feedback or preference optimization to encourage faithful responses (Sun et al., 2024; Zhao et al., 2025).

Editing-based methods intervene in internal activations or model weights rather than relying solely on token selection or additional instruction tuning (Liu et al., 2025; Yang et al., 2025; Dang et al., 2026). This line of work includes activation steering, hallucination-related subspace removal, and adaptive choices of editing locations. HCA-Edit follows this approach, focusing on how correction directions from different hallucination categories are combined and how their signals inform editing strength across network depth.

2.2 Activation steering and knowledge editing

Activation steering derives directions associated with a target behavior and applies them to internal activations during inference (Li et al., 2023a; Rimsky et al., 2024). Knowledge editing methods such as ROME instead modify model parameters to update factual associations (Meng et al., 2022). In LVLMs, VTI derives visual and textual intervention directions from image-mask contrasts and paired descriptions (Liu et al., 2025), while Nullu identifies a hallucination-related subspace and projects MLP weights onto its orthogonal complement (Yang et al., 2025). AFTER further introduces factual semantic guidance and a query-aware offset estimator to adapt activation editing to individual queries (Wang et al., 2026). These methods illustrate different ways to construct and apply corrections within the model.

The location and strength of an intervention are also important. DoLa exploits differences between decoder layers through contrastive decoding (Chuang et al., 2024), while LTS-FS uses causal attribution to identify layers for feature steering (Dang et al., 2026). HCA-Edit considers how hallucination categories should contribute to both the correction direction and its allocation across layers. Unlike AFTER's query-specific offsets, our correction combines separately estimated and normalized object, attribute, and relation directions without selecting a category at inference time. For layer gating, we use category-conditioned activation magnitudes rather than causal attribution, taking the maximum of the depth-normalized category signals to retain a strong response from any category. Our focus is therefore on balancing category directions and distributing editing strength across depth.

3 Method

HCA-Edit is a training-free framework that combines category-balanced direction construction with category-aware layer gating. The framework has two stages. Offline calibration constructs category-structured pairs, estimates correction directions, and computes layer-specific editing coefficients. During inference, these directions and coefficients remain fixed and guide intermediate activation editing. Figure 2 summarizes the framework.

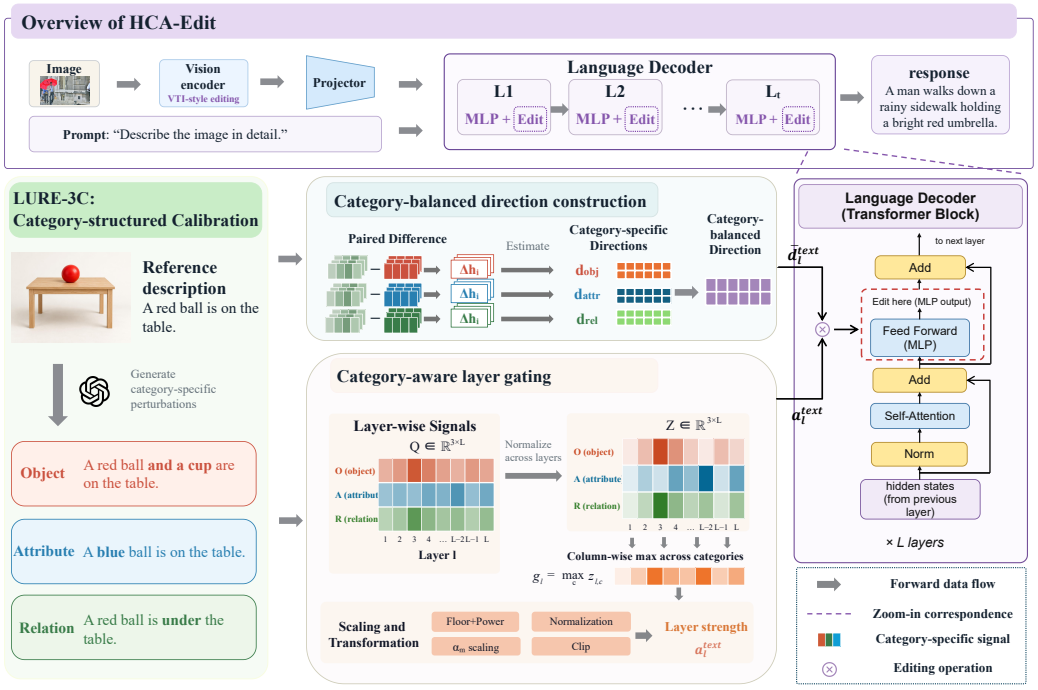


Figure 2: Overview of **HCA-Edit**. LURE-3C provides paired reference and perturbed descriptions for object, attribute, and relation hallucinations. Category-balanced direction construction estimates a direction for each category, then normalizes and combines the directions with equal weights at each layer. Category-aware layer gating takes the maximum of the depth-normalized category signals and converts the shared profile into layer-specific editing strengths. The expanded decoder block illustrates how the resulting direction and strength guide MLP-output editing before residual addition. The vision encoder retains VTI-style editing.

3.1 LURE-3C: Category-Structured Calibration

To systematically obtain category-specific activation signals, we establish LURE-3C, a dedicated calibration dataset constructed by introducing targeted semantic counterfactuals into the multimodal contexts of LURE (Zhou et al., 2024). Specifically, given a set of N seed image-description pairs $\{(x_i, y_i)\}_{i=1}^{N}$, we generate a category-specific perturbed description $\widetilde{y}_{i,c}$ for each error type $c \in \mathcal{C} = \{\text{obj}, \text{attr}, \text{rel}\}$. Object perturbations introduce an unsupported entity, attribute perturbations modify a property of a visible entity, and relation perturbations alter a spatial or action relation between entities. Each perturbation targets a single semantic dimension while aiming to preserve the remaining sentence content.

By pairing each perturbed description with its corresponding reference description and unchanged source image, we form three aligned calibration subsets:

$$\mathcal{D}_c = \{(x_i, y_i, \widetilde{y}_{i,c})\}_{i=1}^{N}, \quad \forall c \in \mathcal{C}, \tag{1}$$

where N denotes the number of shared calibration contexts used for each category. Sharing the same N visual and textual anchors helps control for scene-level and reference-description variations when comparing category-specific activation differences. These subsets are used to estimate both category-specific correction directions and layer-wise activation signals. Detailed dataset construction procedures are provided in Appendix A.2.

3.2 CATEGORY-BALANCED DIRECTION CONSTRUCTION

For each category, we estimate a correction direction from paired hidden-state differences in $\mathcal{D}_c$. The directions are then normalized and combined at each language layer, as illustrated in Figure 2(a).

For each calibration pair, we pass the reference and perturbed descriptions through the frozen model with the same image. Let L denote the number of language decoder layers. At layer l, we denote the last-token hidden states of the reference and category-c perturbed descriptions by $h^+_{i,l}$ and $h^-_{i,l,c}$, respectively. Their paired difference is

$$\Delta h_{i,l,c} = h^+_{i,l} - h^-_{i,l,c}. \tag{2}$$

We estimate a direction separately for each category:

$$d^{\text{text}}_{l,c} = \left[\mathcal{E}\left(\{\Delta h_{i,c}\}_{i=1}^{N}\right)\right]_l, \tag{3}$$

where $\Delta h_{i,c}$ concatenates the paired differences across the hidden-state blocks used for estimation, $\mathcal{E}$ is the direction estimator, and $[\cdot]_l$ selects the block corresponding to decoder layer l. We implement $\mathcal{E}$ using rank-one PCA, adding the leading principal component of the centered differences to their mean.

To prevent raw direction magnitudes from determining category weights, we normalize each layer-specific direction before equal-weight ensembling:

$$\widehat{d}^{\text{text}}_{l,c} = \mathcal{N}\left(d^{\text{text}}_{l,c}\right), \qquad \bar{d}^{\text{text}}_{l} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \widehat{d}^{\text{text}}_{l,c}, \tag{4}$$

where $\mathcal{N}(v) = v/\max(\|v\|_2, \epsilon)$, $\|\cdot\|_2$ denotes the Euclidean norm, and $\epsilon > 0$ ensures numerical stability. The resulting $\bar{d}^{\text{text}}_l$ is the category-balanced correction direction applied at language layer l.

3.3 CATEGORY-AWARE LAYER GATING

We next compute a scalar coefficient for applying $\bar{d}^{\text{text}}_l$ at each layer. As illustrated in Figure 2(b), the gate is derived from paired language MLP outputs using the same calibration subsets.

Let $u^+_{i,l}$ and $u^-_{i,l,c}$ denote the last-token MLP outputs under the reference and category-c perturbed descriptions. We measure the mean magnitude of their paired differences:

$$q_{l,c} = \frac{1}{N}\sum_{i=1}^{N}\left\|u^+_{i,l} - u^-_{i,l,c}\right\|_2, \tag{5}$$

where $q_{l,c}$ measures the activation change associated with category c at layer l. We normalize each category's signal across layers so that aggregation reflects its relative depth profile rather than its overall magnitude:

$$z_{l,c} = \frac{q_{l,c} - \min_k q_{k,c}}{\max_k q_{k,c} - \min_k q_{k,c}}, \qquad g_l = \max_{c\in\mathcal{C}} z_{l,c}. \tag{6}$$

Here and below, k ranges from 1 to L. The normalized signal $z_{l,c}$ indicates the relative response of category c at layer l, and g_l is the shared layer signal. Taking the maximum retains a strong response from any category. This signal controls the shared editing strength without changing the category weights in the correction direction.

We rescale the shared signal and transform it into a relative editing weight:

$$\widehat{g}_l = \frac{g_l - \min_k g_k}{\max_k g_k - \min_k g_k}, \qquad w_l = \frac{f + (1-f)\widehat{g}_l^{\,p}}{\frac{1}{L}\sum_{k=1}^{L}\left[f + (1-f)\widehat{g}_k^{\,p}\right]}. \tag{7}$$

Here, $\widehat{g}_l$ is the rescaled shared signal, and w_l is the relative editing weight at layer l. The floor f retains a baseline weight before mean normalization, while the exponent p controls the profile shape. The denominator gives the layer weights a mean of one, separating relative depth allocation from the overall editing scale. Constant profiles are assigned zero during min–max normalization.

Table 1: CHAIR results. C_s and C_i are hallucination rates at sentence and instance level (lower is better); recall and average length are in percent and words. Lowest C_s and C_i per model are in bold.

	LLaVA-1.5-7B				LLaVA-1.5-13B				Qwen2.5-VL-7B			
Method	$C_s \downarrow$	$C_i \downarrow$	Recall	Len	$C_s \downarrow$	$C_i \downarrow$	Recall	Len	$C_s \downarrow$	$C_i \downarrow$	Recall	Len
Regular	53.0	13.90	77.20	98.0	40.8	9.50	77.20	111.8	27.0	7.40	61.60	120.6
VCD (Leng et al., 2024)	55.2	16.70	77.50	89.2	39.2	9.20	79.10	108.2	26.2	7.60	61.20	120.3
AGLA (An et al., 2025)	50.8	16.10	75.20	88.1	38.4	9.10	78.70	109.3	25.2	7.10	59.50	118.6
Nullu (Yang et al., 2025)	50.2	13.70	76.90	93.3	38.0	9.40	74.50	105.8	27.4	7.70	60.70	121.6
VTI (Liu et al., 2025)	47.4	13.90	76.20	88.9	36.3	9.20	75.90	94.4	25.5	7.10	61.60	121.3
LTS-FS (Nullu) (Dang et al., 2026)	46.8	13.50	76.60	93.2	35.7	8.90	76.10	109.8	23.8	**6.00**	60.80	120.6
LTS-FS (VTI) (Dang et al., 2026)	35.8	11.90	75.40	82.2	32.0	8.80	74.20	83.6	24.8	6.60	62.50	120.0
HCA-Edit	**27.8**	**7.66**	77.37	84.9	**27.4**	**8.23**	78.33	81.9	**20.0**	6.02	64.89	103.0

Finally, we obtain the language editing coefficient:

$$a_l^{\text{text}} = \text{clip}(\alpha_{\text{text}} w_l,\, 0,\, a_{\max})\,, \tag{8}$$

where α_{text} is the base editing strength and $a_{\max}$ is the coefficient cap. The clipping operation, $\text{clip}(x, 0, a_{\max}) = \min(\max(x, 0), a_{\max})$, restricts the coefficient to $[0, a_{\max}]$. The resulting coefficients assign editing strengths across all language layers.

3.4 Inference intervention

At inference time, the direction $\bar{d}_l^{\text{text}}$ and coefficient a_l^{text} jointly determine the correction applied to the language MLP output. Following the norm-preserving intervention in VTI (Liu et al., 2025), we apply

$$\widetilde{u}_l^{\text{text}} = \left\|u_l^{\text{text}}\right\|_2 \mathcal{N}\left(\mathcal{N}\left(u_l^{\text{text}}\right) + \eta a_l^{\text{text}} \bar{d}_l^{\text{text}}\right), \tag{9}$$

where u_l^{text} and $\widetilde{u}_l^{\text{text}}$ are the original and edited MLP-output vectors, respectively, $\mathcal{N}$ is the normalization mentioned in 3.2 , and $\eta = 0.1$ is a fixed implementation scale. The operation modifies the normalized MLP output and restores its original magnitude before residual addition. The visual branch retains VTI-style image-mask contrast editing, with editing strengths obtained by linearly interpolating the shared layer-weight profile across relative depth using a separate visual base strength.

All directions and coefficients are computed offline, so generation requires neither input-dependent category selection nor additional contrastive decoding passes.

4 Experiments

4.1 Experimental setup

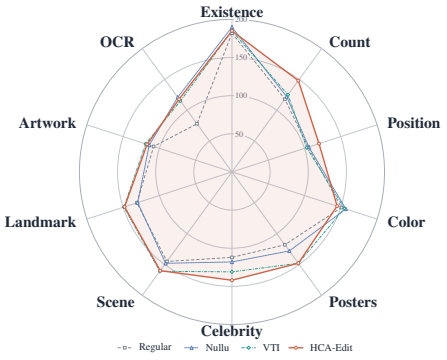
We conduct experiments on LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025b). We assess hallucination mitigation using CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b), and evaluate broader visual understanding and response quality using MME (Fu et al., 2025) and LLaVA-Bench (Liu et al., 2023). POPE results are reported for the two LLaVA models, while MME and LLaVA-Bench evaluations use LLaVA-1.5-7B. We compare HCA-Edit with unedited models, decoding-based methods including VCD and AGLA (Leng et al., 2024; An et al., 2025), and editing methods including VTI, Nullu, and LTS-FS (Liu et al., 2025; Yang et al., 2025; Dang et al., 2026). Benchmark descriptions, evaluation protocols, and implementation details are provided in Appendices A.1 and A.3.

4.2 Results on CHAIR

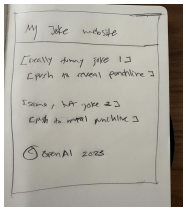
We evaluate hallucinations in open-ended responses using CHAIR, which measures hallucinations at both the sentence level (C_s) and instance level (C_i). As shown in Table 1, HCA-Edit achieves the lowest C_s across all three models and the lowest C_i on both LLaVA models. On Qwen2.5-VL-7B, its C_i is close to the best result obtained by LTS-FS (Nullu), while its recall is the highest among the compared methods. Moreover, the recall is maintained or improved in every case. These results suggest that the reduction in hallucinations is not achieved simply by sacrificing the coverage of correctly described content.

Table 2: POPE results across all three negative-sampling settings. All metrics are percentages; higher is better. The best F1 within each split and model is in bold.

		LLaVA-1.5-7B				LLaVA-1.5-13B			
Split	Method	Acc.↑	Prec.↑	Rec.↑	F1↑	Acc.↑	Prec.↑	Rec.↑	F1↑
Random	Regular	85.37	80.77	93.22	86.47	81.91	75.84	93.82	83.85
	VCD	86.55	84.02	90.69	87.16	82.27	75.97	92.68	83.76
	AGLA	85.32	83.56	91.34	86.77	82.64	76.19	93.16	83.58
	Nullu	86.35	84.36	91.09	86.28	83.24	77.93	92.89	84.73
	VTI	84.84	80.02	93.36	86.08	84.08	76.29	93.04	83.82
	LTS-FS (Nullu)	87.13	84.69	91.02	87.64	83.96	78.89	93.85	85.56
	LTS-FS (VTI)	86.77	84.13	91.00	87.32	86.59	82.35	93.47	87.48
	HCA-Edit	89.67	90.04	89.20	**89.62**	88.00	85.63	91.33	**88.39**
Popular	Regular	77.52	71.45	93.22	80.71	78.40	71.78	93.76	81.30
	VCD	79.09	73.21	92.17	81.23	79.38	72.24	92.47	82.01
	AGLA	78.67	75.39	89.02	81.47	80.11	72.88	92.16	82.32
	Nullu	79.42	74.45	91.04	81.67	80.88	74.91	93.02	82.97
	VTI	77.03	70.90	93.36	80.40	79.22	73.26	93.04	81.83
	LTS-FS (Nullu)	80.09	75.28	91.07	82.20	81.46	75.62	92.93	83.42
	LTS-FS (VTI)	79.96	75.25	91.14	82.25	81.77	75.58	93.47	83.58
	HCA-Edit	84.70	80.85	90.93	**85.60**	82.33	77.84	90.40	**83.65**
Adversarial	Regular	70.13	64.14	93.22	75.85	71.07	64.60	93.83	76.47
	VCD	71.48	66.28	89.62	76.04	71.73	63.61	94.23	75.76
	AGLA	71.63	66.59	90.13	75.89	72.27	64.14	93.56	75.48
	Nullu	71.57	66.06	90.53	76.17	72.43	66.08	92.44	77.04
	VTI	69.40	63.46	93.36	75.42	71.77	65.58	93.01	76.80
	LTS-FS (Nullu)	72.62	65.99	90.62	76.22	73.06	67.01	92.96	78.36
	LTS-FS (VTI)	73.04	67.37	91.24	77.32	73.78	67.51	93.47	79.91
	HCA-Edit	77.90	72.93	88.73	**80.06**	77.13	70.77	92.47	**80.17**



(a) MME perception results on LLaVA-1.5-7B.



Query: Describe this sketch in detail.

[Regular] The image features a handwritten list of jokes on a piece of paper. The jokes are written in a comical manner, with some of them being quite funny. The list includes a total of 13 jokes, each with varying levels of humor. The jokes are spread across the paper, with some jokes being closer to the top, while others are positioned towards the bottom. The handwriting style suggests that the jokes were written by someone who enjoys humor and has a creative approach to writing.

[HCA-Edit] The image features a handwritten note on a piece of paper, likely a sketch or a doodle. The note consists of several lines, with one line at the top reading "my joke website." The rest of the lines are filled with various words and phrases, possibly related to the joke website or the content of the note. The handwriting appears to be casual and informal, suggesting a personal or creative expression. The paper is placed on a wooden surface, which could be a table or a desk.

(b) Qualitative comparison on a LLaVA-Bench example using LLaVA-1.5-7B. Hallucinated content is highlighted in red.

Figure 3: MME perception results and a qualitative LLaVA-Bench comparison on LLaVA-1.5-7B.

4.3 Results on POPE

Table 2 presents the POPE results on LLaVA-1.5-7B and LLaVA-1.5-13B. HCA-Edit achieves the highest accuracy and F1 among the compared methods under all three negative-sampling settings. The improvements over the unedited models are accompanied by higher precision, with some reduction in recall, suggesting that HCA-Edit effectively suppresses false-positive predictions associated with hallucinations. Overall, these results demonstrate the effectiveness of HCA-Edit in mitigating object hallucinations across different models and negative-sampling settings.

4.4 RESULTS ON MME

Figure 3a presents the MME results on LLaVA-1.5-7B. HCA-Edit achieves the highest aggregate score across the perception subtasks, outperforming the unedited model, VTI, and Nullu. It improves nine subtasks over the unedited model, with gains spanning counting, spatial localization, recognition, and OCR. Compared with VTI, the largest improvements occur in count, position, and celebrity recognition. These results complement the hallucination evaluations, showing that HCA-Edit also improves overall visual understanding on MME. Detailed subtask scores are provided in Appendix B.6.

4.5 RESULTS ON LLAVA-BENCH

As shown in Table 3, HCA-Edit improves average accuracy and detailedness over the unedited model, together with lower object and relation hallucination severity. Attribute severity shows a small increase, reflecting a category-specific trade-off rather than a uniform improvement across all dimensions. The results show that HCA-Edit produces more accurate and informative responses while reducing hallucination severity in two of the three evaluated categories. Details of the evaluation protocol are provided in Appendix B.5.

Qualitative analysis. Figure 3b compares the responses to a handwritten website sketch. The unedited model describes the sketch as a list of jokes and introduces an unsupported count of thirteen, followed by further claims about their arrangement and humour. HCA-Edit instead identifies the visible heading, "my joke website," and describes the handwritten layout without repeating the unsupported count or object. This example illustrates how HCA-Edit can reduce hallucination while retaining details grounded in the image. Additional examples covering counting and scene explanation are provided in Appendix B.8.

Table 3: LLaVA-Bench results on LLaVA-1.5-7B under our evaluation protocol. Accuracy and detailedness are mean scores on a 1–10 scale (higher is better). Hallucination severity is scored on a 0–2 scale (lower is better); Avg. sev. is the arithmetic mean across the three categories.

Method	Accuracy↑	Detailedness↑	Obj. sev.↓	Attr. sev.↓	Rel. sev.↓	Avg. sev.↓
Regular	5.95	5.12	0.82	**0.53**	0.75	0.70
HCA-Edit	**6.40**	**5.25**	**0.62**	0.57	**0.67**	**0.62**

4.6 ABLATION STUDIES

Direction construction and layer gating. Table 4 compares variants within the same three-category calibration and intervention framework. The pooled baseline estimates a single direction from the combined activation differences of all three calibration categories, while uniform gating assigns the same editing coefficient to all layers within each branch. Category-balanced directions combined with category-max gating achieve the lowest C_s and C_i, indicating that the two components are most effective when used together. The results highlight the benefit of pairing balanced category contributions with depth-dependent strength allocation. Replacing max aggregation with the mean or shuffling the layer profile increases both hallucination rates, supporting the use of category-max aggregation and the calibrated assignment of editing strengths to layers.

Source of gating signals. Table 5 compares three signal sources for layer gating: hidden states, attention residuals, and MLP outputs. Among them, MLP-output signals achieve the lowest hallucination rates together with the highest recall and BLEU. The three choices differ in how closely they reflect the representation that is directly modified by HCA-Edit. Hidden-state signals capture the combined effect of preceding transformations and residual pathways, while attention-residual signals characterize the contribution of the attention sublayer. In contrast, MLP-output signals are computed from the same MLP output that is directly modified during inference, providing a more localized measure of the activation change at the intervention site. The results suggest that signals measured at the intervention site provide more effective guidance for allocating editing strength

Table 4: Direction construction and layer gating within the HCA-Edit framework on LLaVA-1.5-7B evaluated with CHAIR. Both direction variants use three-category calibration data.

Direction	Layer gate	$C_s \downarrow$	$C_i \downarrow$	Recall
Pooled	Uniform	29.4	8.49	76.44
Category-balanced	Uniform	31.0	8.70	76.54
Pooled	Category max	32.2	8.70	77.70
Category-balanced	Category max	**27.8**	**7.66**	77.37
Category-balanced	Category mean	30.8	8.54	77.10
Category-balanced	Shuffled max	29.4	8.82	76.91

Table 5: Effect of the gating signal source on LLaVA-1.5-7B evaluated with CHAIR.

Signal source	$C_s \downarrow$	$C_i \downarrow$	Recall	Avg. len.	BLEU
Hidden states	28.4	8.09	75.37	84.82	9.93
Attention residual	28.0	8.03	76.27	85.61	9.91
MLP output (ours)	**27.8**	**7.66**	77.37	84.90	10.27

across layers than signals from upstream or parallel components. We therefore use MLP-output signals for category-aware layer gating.

Discussion: category-specific gating. An alternative to shared gating is to use each category's signal to scale its corresponding direction independently. Specifically, instead of aggregating the three category-specific signals into a shared layer profile, we retain a separate layer-wise signal for each category and use it to determine the editing strength of its corresponding correction direction. The three category-specific directions are therefore modulated independently and then combined at each layer. As shown in Table 6, this variant offers no improvement over the shared gate, which obtains slightly lower hallucination rates and higher recall. Independent gating thus changes both the editing strength and the relative contribution of different categories across layers, whereas shared gating uses the aggregated signal to adjust the overall layer strength while preserving the balanced category mixture. We therefore retain shared gating in HCA-Edit.

Additional ablations, hyperparameter analyses, and evaluation controls are provided in Appendix B.

Table 6: Comparison of shared and category-specific gating on LLaVA-1.5-7B evaluated with CHAIR. Shared gating scales the category ensemble as a whole, while category-specific gating scales each category direction independently.

Layer gating	$C_s \downarrow$	$C_i \downarrow$	Recall	Avg. len.	BLEU
Shared (ours)	**27.80**	**7.66**	77.37	84.90	10.27
Category-specific	28.20	7.74	77.01	84.92	10.26

5 Conclusion

We presented **HCA-Edit**, a training-free framework for mitigating visual hallucinations through category-balanced direction construction and category-aware layer gating. Using object, attribute, and relation calibration pairs, HCA-Edit estimates correction directions separately and normalizes them before equal-weight ensembling. A shared gate aggregates their depth-normalized signals to allocate editing strength across layers while preserving the balanced category mixture. Experiments across multiple LVLMs show reduced hallucinations alongside competitive or improved performance on complementary capability evaluations. Ablation studies highlight the importance of combining category balancing with depth-dependent strength allocation.

AI USE STATEMENT

In this work, we used generative AI tools to assist in generating the synthetic calibration data used to construct the LURE-3C dataset, including category-specific semantic perturbations for object, attribute, and relation hallucinations.

We also used generative AI tools to assist with research ideation and execution, including providing feedback on the methodology and experimental design and assisting with modifications to research code. AI-assisted code was reviewed, tested, and verified by the authors against the actual implementation and experimental results.

Additionally, we used generative AI tools to edit and improve the manuscript for grammar, clarity, and readability. All AI-assisted text and other outputs were reviewed and revised by the authors, and the methodological details, claims, experimental settings, and reported results were independently checked against the actual implementation and experiments.

We take responsibility for the final content of this work, including all text, data, code, claims, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The proposed HCA-Edit framework is described in detail in Section 3, covering LURE-3C calibration data construction, category-balanced direction construction, category-aware layer gating, and inference-time intervention. Section 4 presents the experimental setup, including the evaluated models, benchmarks, baselines, and evaluation protocols. The appendix provides implementation settings, hyperparameters, additional ablation studies, and further details on calibration data construction and processing. We will release the source code and experiment scripts at the anonymous repository `https://anonymous.4open.science/r/HCA-Edit-28F3` to facilitate reproduction of our results.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL `https://arxiv.org/abs/2502.13923`.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Tiantian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. Locate-then-sparsify: Attribution guided sparse strategy for visual hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18251–18260, June 2026.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312. IEEE, 2024.

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. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems*, 2025. URL https://mlanthology.org/neurips/2025/fu2025neurips-mme/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 57689–57733, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fed9263d7f1086e735904ff690527a53-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024b.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and

Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. After: Mitigating the object hallucination of lvlm via adaptive factual-guided activation editing. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 93458–93484, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/96ee35c170cb8720033b16259c305da9-Paper-Conference.pdf.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645. IEEE, 2025.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12944–12953. IEEE, 2024.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond multimodal hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209377.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

A Evaluation and implementation details

A.1 Benchmarks and metrics

The main comparisons use CHAIR on all three models, POPE on the two LLaVA models, and MME and LLaVA-Bench on LLaVA-1.5-7B. The unedited model (Regular) provides the reference for each model family. Decoding baselines include VCD and AGLA, while editing baselines include VTI, Nullu, and the Nullu- and VTI-based LTS-FS variants. The available comparisons for each benchmark are shown in its results table.

CHAIR. We use CHAIR to evaluate hallucinations in open-ended image descriptions on 500 sampled images, with COCO annotations and reference descriptions providing the grounding evidence (Rohrbach et al., 2018). We report sentence-level C_s, the fraction of captions containing at least one hallucination, and instance-level C_i, the fraction of generated object mentions that are unsupported. Lower values indicate fewer hallucinations. We additionally report recall and average response length to characterize descriptive coverage. For an image with a nonempty annotated category set G_i and generated category set P_i, recall is computed as $|P_i \cap G_i|/|G_i|$ and averaged across images. Average response length is measured in whitespace-separated words in the raw generated captions.

POPE. We use POPE to evaluate hallucinations in a discriminative visual question-answering setting (Li et al., 2023b). Unlike CHAIR, which assesses open-ended descriptions, POPE requires a yes-or-no response to a targeted question about image content. Its random, popular, and adversarial settings construct negative questions using randomly selected, frequently occurring, and commonly co-occurring objects, respectively. We report accuracy, precision, recall, and F1 separately for each setting, treating affirmative answers as the positive class. Accuracy measures the proportion of correct answers, precision measures the correctness of affirmative predictions, recall measures the proportion of positive questions correctly answered, and F1 is the harmonic mean of precision and recall.

MME. We use MME (Fu et al., 2025) to assess visual understanding across its ten perception subtasks, covering existence, count, position, color, recognition, and OCR. We report the benchmark's score for each subtask and their sum as the perception total. The detailed results are provided in Appendix B.6.

LLaVA-Bench. We use the benchmark images and prompts for judged response evaluation (Liu et al., 2023). The reported dimensions are accuracy, detailedness and the severity of object, attribute and relation errors. Judge scores complement automatic metrics but depend on the rubric and evaluator. The evaluator and scoring rubric are described in Appendix B.5.

A.2 Calibration and direction construction

LURE-3C is constructed from LURE image-description pairs with language-model-generated object, attribute and relation perturbations. The current collection contains 500 image contexts and a perturbed description for each category, yielding 1,500 paired calibration instances. These are our derived calibration data, not three original annotation categories provided by LURE. Figure 4 shows one concrete example.

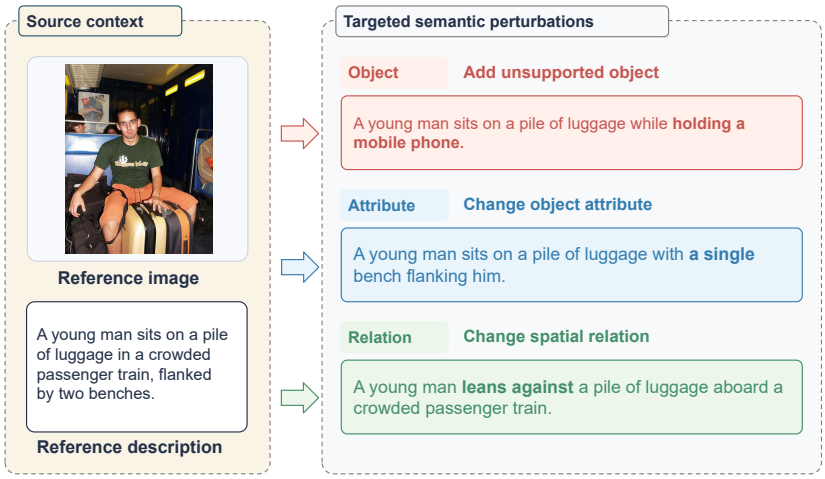


Figure 4: Construction of LURE-3C. For each image-description pair, we perturb the reference description to introduce object, attribute, or relation errors while keeping the image unchanged. Each perturbed description is paired with the reference to form a category-specific calibration example.

Construction prompt. We design the construction prompt to organize hallucination calibration around explicit semantic categories. As shown in Figure 5, the input includes the image identifier, reference caption, original hallucinated caption, candidate false objects, and uncertain objects. Rather than producing a single description with mixed errors, the prompt requests three minimally modified descriptions targeting object, attribute, and relation errors separately. Preserving the reference style and length helps focus the comparison on the targeted semantic change. The output is a structured JSON record containing the reference, the three perturbed descriptions, and an attribute-type label.

The three perturbations share the same image and reference description, providing a common anchor for category-wise activation comparisons. This construction reduces variation in scene content across categories and allows each correction direction to be estimated from its corresponding perturbations before aggregation. LURE-3C therefore provides the category-structured evidence used by both direction construction and layer gating: the paired differences inform what to correct, while their layer-wise magnitudes inform how to distribute editing strength.

Implementation of the direction estimator. We implement $\mathcal{E}$ using rank-one PCA. For each calibration pair, we extract the last-token hidden states and compute their differences between the reference and perturbed descriptions, taking the reference minus the perturbed states. These differences are concatenated across layers to form one vector per example. Within each category, we compute the mean vector and apply singular value decomposition to the centered vectors. The estimator returns the mean plus the first principal component, combining the average shift with the dominant direction of variation across examples. The resulting vector is split into layer-specific blocks, and the embedding-output block is discarded. The remaining blocks are normalized and ensembled as described in Section 3.2.

Visual directions. The visual estimates retain the image-mask contrast used by VTI. Cached visual directions are normalized and equally ensembled, but the cache groups do not constitute category-specific visual supervision. The object, attribute, and relation labels organize the textual calibration pairs rather than defining three distinct types of image masking.

```
You are an expert visual caption editor.

Given a clean image caption and an original hallucinated caption, construct three minimally modified hallucinated captions:
1. Object hallucination: add or replace a visible object/entity/count with a plausible but false object.
2. Attribute hallucination: change a visual attribute such as color, material, state, size, identity, or text while keeping objects and relations mostly unchanged.
3. Relation hallucination: change an action, spatial relation, interaction, or pose while keeping objects and attributes mostly unchanged.

Requirements:
- Preserve the sentence style and length of the clean caption as much as possible.
- Modify only one hallucination type at a time.
- The object hallucination should mainly affect object/entity/count.
- The attribute hallucination should mainly affect color/material/state/identity/text/count attribute.
- The relation hallucination should mainly affect action/spatial/interaction relation.
- Do not add explanations.
- Return JSON only.

Input:
image_id: {image_id}
clean_caption: {value}
original_hallucinated_caption: {h_value}
candidate_false_objects: {co_objects}
uncertain_objects: {uncertain_objects}

Output schema:
{
  "value": "...",
  "h_value_obj": "...",
  "h_value_attr": "...",
  "h_value_rel": "...",
  "attr_type": "Color|Material|State|Size|Identity|Text|Count|Other"
}
```

Figure 5: Prompt template for constructing LURE-3C. Each source record produces three targeted perturbations anchored to the same reference description, providing category-structured calibration pairs for direction estimation and layer gating. Braced fields are replaced with the corresponding source-record values.

Table 7: Generation settings for the reported HCA-Edit evaluations. Sampling is disabled.

Benchmark	Maximum new tokens	Beams	Temperature	Seed
CHAIR	128	1	0	0
POPE	16	1	0	0
MME	16	3	0	0
LLaVA-Bench	256	1	0	0

A.3 Generation settings and editing parameters

Generation and editing settings. Table 7 lists the decoding settings, while Table 8 specifies the editing parameters for each model and benchmark. Sampling is disabled in all HCA-Edit evaluations. For each LLaVA model, CHAIR and POPE use identical editing parameters; only the maximum generation length changes. Qwen2.5-VL-7B uses the same numerical editing settings for CHAIR. The MME results use a separate configuration with $\alpha_{\text{text}} = \alpha_{\text{image}} = 0.25$, $f = 0.10$, and $p = 1.0$. Across these configurations, the method retains category-balanced directions, a shared category-max gate, and editing at all language and visual layers. The implementation scale is fixed at $\eta = 0.1$. All directions and layer coefficients are computed offline and remain fixed during generation.

Table 8: Editing parameters for the reported HCA-Edit results. These parameters are distinct from the generation settings in Table 7.

Model	Benchmark	α_{text}	α_{image}	f	p	$a_{\max}$
LLaVA-1.5-7B	CHAIR / POPE	0.60	0.60	0.25	1.5	1.2
LLaVA-1.5-13B	CHAIR / POPE	0.60	0.60	0.25	1.5	1.2
Qwen2.5-VL-7B	CHAIR	0.60	0.60	0.25	1.5	1.2
LLaVA-1.5-7B	MME	0.25	0.25	0.10	1.0	1.2
LLaVA-1.5-7B	LLaVA-Bench	0.60	0.60	0.25	1.5	1.2

Visual-layer coefficients. The language-derived weights w_l also determine the depth profile of visual editing. We first multiply these weights by the visual base strength α_{image} and clip the resulting coefficients to $[0, a_{\max}]$. Since the two branches can have different numbers of layers, we represent their layer positions on a common relative-depth interval from zero to one. Linear interpolation then assigns each visual layer a coefficient from the language-derived profile at the corresponding relative depth. This transfers only the editing-strength profile; the visual correction directions are estimated separately from image-mask contrasts. Appendix B.4 compares this allocation with fixed visual strength.

B Additional analyses

The main text compares direction construction, layer gating, signal sources, and category-specific gating. This section reports complementary experiments on intervention branches, calibration categories, hyperparameters, visual strength allocation, and inference cost, together with detailed results. Unless otherwise stated, the additional ablations use LLaVA-1.5-7B and CHAIR.

B.1 Intervention branches

Table 9 compares language-only, vision-only, and dual-branch editing. Combining both branches yields the lowest C_s and C_i, outperforming either branch alone. Language-only editing is more effective than vision-only editing, and adding visual intervention further reduces hallucinations. The two branches draw on different calibration contrasts: image-mask pairs provide visual correction directions, while category-structured description pairs provide language correction directions. The results support combining these complementary sources of correction in HCA-Edit.

Table 9: Intervention branches on LLaVA-1.5-7B evaluated with CHAIR. The lowest C_s and C_i are in bold.

Intervention	$C_s \downarrow$	$C_i \downarrow$	Recall	Len
Text only	37.2	10.02	78.51	88.95
Vision only	40.0	10.52	78.39	81.13
Both (ours)	**27.8**	**7.66**	77.37	84.90

B.2 Hyperparameter sensitivity

Table 10 examines the editing strength α, gate exponent p, and floor f on LLaVA-1.5-7B. In the strength sweep, we set $\alpha_{\text{text}} = \alpha_{\text{image}} = \alpha$, jointly varying the base editing strength of both branches. The main results use the preset configuration $\alpha = 0.6$, $p = 1.5$, and $f = 0.25$, rather than a configuration selected to optimize the scores in this sensitivity study.

Editing strength. Increasing α from 0.2 to 0.8 progressively reduces both C_s and C_i, showing that stronger intervention improves hallucination suppression over the tested range. The default $\alpha = 0.6$ achieves the highest BLEU, while $\alpha = 0.8$ further lowers hallucination rates. These results demonstrate that the preset strength already provides an effective correction, with additional scope to reduce hallucinations through stronger editing.

Gate exponent. Increasing p also lowers both hallucination rates, with the best CHAIR results obtained at $p = 2.0$ and the highest BLEU at $p = 1.5$. A larger exponent suppresses intermediate normalized signals relative to their peaks before floor adjustment and mean normalization. The results suggest that giving greater emphasis to strong-signal layers can further improve the intervention beyond the default depth profile.

Gate floor. Among the tested floor values, $f = 0.25$ achieves the lowest hallucination rates and the highest BLEU. The floor controls how much intervention is retained at weak-signal layers: increasing it makes the schedule more uniform, while decreasing it allows larger differences between strong- and weak-signal layers. The intermediate value performs best in this comparison, favoring a depth-dependent schedule that retains a baseline contribution from weaker-signal layers.

Overall, the main configuration performs well without selecting the best hyperparameters from these sweeps. The stronger results obtained with alternative values of α and p indicate further potential for hallucination reduction through hyperparameter calibration.

Table 10: Hyperparameter sensitivity on LLaVA-1.5-7B evaluated with CHAIR. The lowest C_s and C_i within each parameter block are in bold.

Parameter	Value	$C_s \downarrow$	$C_i \downarrow$	Recall	Avg. len.	BLEU
α	0.20	40.0	10.81	78.86	84.75	10.18
	0.40	33.0	8.75	76.30	85.53	9.91
	0.60	27.8	7.66	77.37	84.90	10.27
	0.80	**26.8**	**6.95**	73.25	86.73	9.50
p	0.50	29.2	7.97	74.75	86.57	9.73
	1.00	28.6	7.74	75.03	85.84	9.91
	1.50	27.8	7.66	77.37	84.90	10.27
	2.00	**27.4**	**7.46**	75.46	85.07	9.97
f	0.10	28.2	7.82	76.10	85.35	9.90
	0.25	**27.8**	**7.66**	77.37	84.90	10.27
	0.50	28.4	8.14	76.77	85.35	9.93

B.3 CALIBRATION CATEGORIES

Table 11 compares single-category directions with the three-category ensemble. Relation-only calibration achieves the lowest C_i, suggesting that relational perturbations provide correction information beyond an isolated relation change. Relations describe how entities interact or are spatially arranged, and their representations can depend on both the participating entities and their properties. A relation-derived direction may therefore capture information shared with object and attribute corrections. This interpretation is consistent with Figure 1, where the category-conditioned activation differences overlap in the projection and follow similar overall depth trends. At the same time, their distinct local responses suggest that the categories are not interchangeable. Combining all three directions yields the lowest C_s and the highest recall, indicating that category-specific information remains useful alongside the shared structure captured by relation-only calibration.

Table 11: Category ablation on LLaVA-1.5-7B with CHAIR.

Direction	$C_s \downarrow$	$C_i \downarrow$	Recall	Len	BLEU
Object only	30.2	7.93	75.50	84.56	9.97
Attribute only	30.8	8.40	76.38	82.50	10.06
Relation only	28.8	**7.13**	76.81	85.12	10.39
Three categories (ours)	**27.8**	7.66	77.37	84.90	10.27

B.4 VISUAL STRENGTH ALLOCATION

We compare the transferred depth profile with a constant visual coefficient, $a_j^{\text{vision}} = \alpha_{\text{image}}$, keeping language-side editing, visual directions, and the intervention operator unchanged. Table 12 shows that depth-profile transfer lowers both C_s and C_i on LLaVA-1.5-13B and Qwen2.5-VL-7B, while fixed strength gives lower rates on LLaVA-1.5-7B. By transferring the category-conditioned depth profile, HCA-Edit uses the calibration signals to distribute visual editing strength instead of assigning the same coefficient to every layer. The gains on LLaVA-1.5-13B and Qwen2.5-VL-7B support the usefulness of this cross-branch information, and we adopt depth-profile transfer as a unified allocation rule across models.

Table 12: Transferred versus fixed visual strength on CHAIR. Language-side editing, visual directions, and the intervention operator are unchanged. The lowest C_s and C_i within each model are in bold.

Model	Visual strength	$C_s \downarrow$	$C_i \downarrow$
LLaVA-1.5-7B	Transferred profile	27.8	7.66
	Fixed	**25.0**	**6.95**
LLaVA-1.5-13B	Transferred profile	**27.4**	**8.23**
	Fixed	31.6	8.51
Qwen2.5-VL-7B	Transferred profile	**20.0**	**6.02**
	Fixed	24.4	7.23

B.5 LLAVA-BENCH EVALUATION PROTOCOL.

We use GPT-5.6 to evaluate 60 pairs of responses from Regular and HCA-Edit on LLaVA-Bench. Each pair corresponds to the same image and question, with responses generated using temperature zero, one beam, and a maximum of 256 new tokens. Figure 6 presents the evaluation prompt, which assesses visual accuracy, useful detailedness, and hallucination severity for objects, attributes, and relations. The evaluator scores each response independently, with instructions to disregard presentation order and reward visually grounded details rather than unsupported elaboration. For creative or explanatory questions, non-visual extensions are assessed only when explicitly presented as claims about the image. Scores are returned in a structured JSON format and averaged over the 60 questions.

You are an impartial AI evaluator scoring two AI assistants that describe the same image.
Your primary focus is visual accuracy and useful detailedness. Assess accuracy by identifying hallucinations: any statement inconsistent with the image content. Assess detailedness by counting necessary, visually grounded details only; never reward unsupported verbosity. Evaluate each assistant independently and do not let its presentation order influence the score. For creative, travel, recipe, or explanatory questions, ignore non-visual extensions unless they are explicitly grounded in the image.

For each answer output numeric labels:
acc: visual accuracy, 1-10 (higher is better).
det: useful visually grounded detailedness, 1-10 (higher is better).
obj: object/entity/count hallucination severity, 0 none, 1 minor or uncertain, 2 clear.
attr: attribute/color/property hallucination severity, 0 none, 1 minor or uncertain, 2 clear.
rel: relation/action/spatial hallucination severity, 0 none, 1 minor or uncertain, 2 clear.

Return JSON only. Do not provide an explanation, markdown, or any text outside the JSON.
Schema:
{"a1":{"acc":1,"det":1,"obj":0,"attr":0,"rel":0},"a2":{"acc":1,"det":1,"obj":0,"attr":0,"rel":0}}

Figure 6: Evaluation prompt for the paired LLaVA-Bench responses. The evaluator returns accuracy and grounded detailedness scores, together with object, attribute, and relation hallucination severity scores, in JSON format.

B.6 MME

Table 13 reports the detailed MME results corresponding to Figure 3a. HCA-Edit achieves the highest total score across the ten evaluated perception subtasks and improves nine of them over Regular. The gains span counting, spatial localization, recognition, and OCR. Compared with VTI, the largest improvements occur in count, position, and celebrity recognition. These results demonstrate that HCA-Edit improves visual understanding alongside hallucination mitigation.

Table 13: MME results on LLaVA-1.5-7B across ten perception subtasks. Scores are rounded to the nearest integer for presentation. Totals are rounded independently and may differ from the sums of the displayed subtask scores.

Subtask	Regular	Nullu	VTI	HCA-Edit
Existence	182	190	185	185
Count	118	122	125	148
Position	105	106	103	120
Color	151	157	155	145
Posters	118	128	148	148
Celebrity	112	118	131	142
Scene	145	148	161	160
Landmark	131	130	149	148
Artwork	108	114	119	117
OCR	78	121	115	118
Perception total	1248	1334	1390	1430

B.7 Runtime analysis

Table 14 compares the average generation time per image on 100 CHAIR images using LLaVA-1.5-7B, excluding model loading, warm-up, and offline calibration. HCA-Edit computes its correction directions and layer coefficients offline. During generation, it applies vector operations to MLP outputs without additional model forward passes for contrastive decoding. HCA-Edit takes 4.33 seconds per image, compared with 3.33 seconds for Regular and 3.95 seconds for VTI. Its category-balanced directions and layer-dependent strengths therefore add 0.38 seconds per image over VTI in this evaluation. Compared with the recorded VCD configuration, HCA-Edit reduces generation time by approximately 25%.

Table 14: Average generation time on 100 CHAIR images with LLaVA-1.5-7B. Model loading, warm-up, and offline calibration are excluded.

Method	Seconds per image	Relative to Regular
Regular	3.33	1.00
VTI	3.95	1.19
HCA-Edit	4.33	1.30
VCD	5.80	1.74

B.8 Additional qualitative examples

Figures 7 and 8 provide two further comparisons between Regular and HCA-Edit on LLaVA-Bench using LLaVA-1.5-7B. They illustrate improvements in a targeted counting question and an open-ended explanation of a traffic scene.

Counting uncut fruits. In Figure 7, the image contains three uncut fruits alongside a cut fruit. Regular answers "four," whereas HCA-Edit correctly answers "three." The edited response matches the requested count of uncut fruits, rather than the total number of fruits shown.



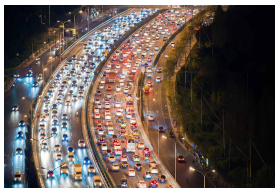
Query: How many uncut fruits are in the image?

[Regular] There are four uncut fruits in the image.

[HCA-Edit] There are three uncut fruits in the image.

Figure 7: Qualitative comparison on a LLaVA-Bench example using LLaVA-1.5-7B. Hallucinated content is highlighted in red.

Explaining traffic conditions. In Figure 8, Regular introduces multiple traffic lights and elaborates on their locations and role in controlling traffic, although these details are not supported by the image. HCA-Edit instead bases its explanation on the visible vehicles and their distribution along the highway. This example shows how avoiding an unsupported entity can also prevent further claims built around it.



Query: Explain all the cues that indicate the current traffic conditions.

[Regular] The image depicts a busy highway with heavy traffic, as evidenced by the numerous cars and trucks traveling in both directions. The traffic is congested, with vehicles occupying all lanes, and some cars are even driving on the shoulder. The high volume of vehicles on the road, combined with the presence of multiple traffic lights, indicates that the traffic conditions are challenging and require extra caution from drivers. The traffic lights are scattered throughout the scene, with some closer to the foreground and others further away, suggesting that they are controlling the flow of traffic at various points along the highway.

[HCA-Edit] The image depicts a busy highway with heavy traffic, including numerous cars and trucks. The traffic is moving in both directions, with some vehicles closer to the foreground and others further away. The highway is filled with cars and trucks, indicating a high volume of traffic at the time the photo was taken.

Figure 8: Qualitative comparison on a LLaVA-Bench example using LLaVA-1.5-7B. Hallucinated content is highlighted in red.