Beyond Visual Neglect: Context-Preference Activation Steering for Mitigating MLLM Hallucinations

Jingwen Wu **Xijun Zhang** **Ge Song***
Nanjing Normal University
{19230415,19230424,g.song}@njnu.edu.cn

Abstract

Object hallucination remains a primary obstacle to the reliable deployment of Multimodal Large Language Models (MLLMs). Current inference-time mitigation methods mainly assume that hallucinations stem from *visual neglect* and therefore steer models toward greater visual reliance. In contrast, our systematic interventions on multiple MLLMs show that increasing visual reliance may *exacerbate* hallucinations on some models, whereas reducing it may mitigate them. This result suggests that visual neglect alone provides an incomplete explanation of hallucination. We characterize the image as a context that competes with both the model's parametric knowledge and its textual context. Based on this characterization, we propose a training-free framework, **Context-Preference Activation Steering** (CAS). CAS extracts two semantically distinct **Context Preference Vectors** (CPVs) from two small sets of designed conflict samples and injects signed residuals into the MLP-sublayer outputs before residual addition at four mid-early layers, without requiring an additional forward pass per decoding step. Experiments show that CAS substantially mitigates object hallucinations with negligible decoding overhead while maintaining F1, lexical diversity, and repetition rates close to those of the base models. Code and data are available at https://github.com/wss7/CAS.

1 Introduction

Multimodal Large Language Models (MLLMs) have made rapid progress in visual understanding and open-ended generation, yet object hallucination, in which the model describes objects absent from the input image, remains a core obstacle to reliable deployment (Li et al., 2023b). The problem is particularly severe in long-form generation (e.g., detailed captioning): as the generation length grows, the model progressively drifts away from visual evidence and hallucinations accumulate (Zheng et al., 2025; Zhou et al., 2024). In the literature, given the prohibitive cost of retraining and alignment fine-tuning, training-free inference-time intervention has emerged as an attractive route.

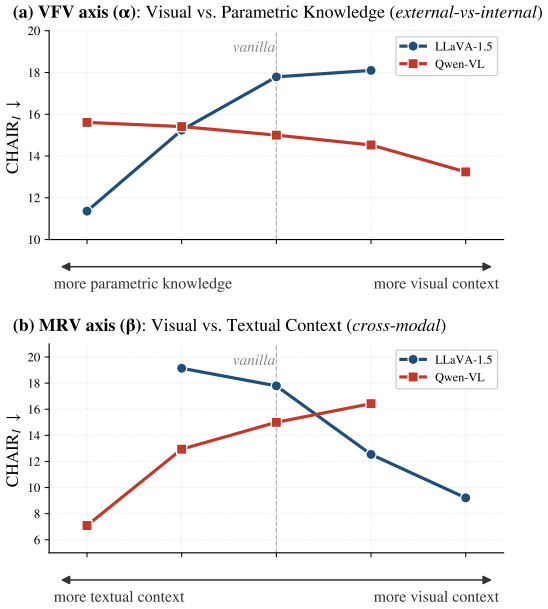


Figure 1: **CHAIR$_I$ of LLaVA-1.5 and Qwen-VL under VFV and MRV interventions.** This reveals three phenomena: (i) steering toward *more image context* on either CPV may instead exacerbate hallucinations; (ii) the response slopes across the two CPVs can take opposite signs within the same model; (iii) on the same CPV, different models exhibit divergent trends.

However, several representative inference-time mitigation methods (e.g., PAI (Liu et al., 2024b), AttnReal (Tu et al., 2025), and VCD (Leng et al., 2024)) build on a common visual-neglect hypothesis: hallucinations arise partly from insufficient reliance on visual evidence and excessive reliance on textual priors. They therefore encourage the

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 24454–24470
October 24-29, 2026 ©2026 Association for Computational Linguistics

model to *pay more attention to the image* through various interventions, regardless of the base model. Yet the gains of these methods are inconsistent across base models, suggesting that *visual neglect* alone is an incomplete characterization of the reason for MLLM hallucination.

This paper offers a systematic re-examination of this consensus. We argue that the image input, as a context, simultaneously competes against two semantic signals inside the model: the model's parametric knowledge from pretraining and the textual context formed jointly by the textual instruction and the generation history. However, prior work has not examined how these two competitions jointly shape hallucinations. Based on this characterization, we decouple visual preference along two distinct competing dimensions: the *external-vs-internal* competition, and the *cross-modal* competition between visual and textual context. We accordingly construct two conflict-based sample sets: *counterfactual images* that violate commonsense and *symmetric image-text conflict pairs*. By capturing the model's MLP-sublayer outputs under these conflicting inputs, we extract two semantically distinct **Context Preference Vectors (CPVs)**: the **Visual Fidelity Vector (VFV)**, which captures the *external-vs-internal* preference, and the **Modality Reliance Vector (MRV)**, which captures the *cross-modal* preference (see §3.1).

Figure 1 reports the results of using the two CPVs to perform visual-dependency interventions on LLaVA-1.5 and Qwen-VL. We observe that: (i) intervening toward *more visual context* does not always reduce hallucinations, as witnessed by the two ascending curves in the figure; (ii) on the same model, the optimal sign of intervention on VFV and MRV can be opposite, showing that their behavioral responses need not coincide; (iii) for the same CPV, the sign of the positive response differs across models. These phenomena reveal not only the incompleteness of the simplistic *more visual, less hallucination* assumption, but also the heterogeneity of different models with respect to different context preferences.

Based on the above observation, we find that each model has, on both VFV and MRV, a signed intervention setting that substantially reduces hallucinations. Therefore, we propose **Context-Preference Activation Steering (CAS)**, a training-free inference-time framework. CAS requires only a small set of task-irrelevant samples to extract VFV and MRV within minutes. In inference, it injects signed residuals into the MLP-sublayer outputs, before residual addition, at four mid-early layers, with lightweight model- and task-specific calibration of the signed coefficients. Experiments on mainstream base models and benchmarks show that CAS substantially mitigates hallucinations without requiring an additional forward pass per decoding step, with negligible decoding overhead while maintaining the reported captioning-quality and repetition metrics. We release the implementation, evaluation scripts, and 161 context-conflict samples to support reproducibility.

The main contributions are summarized as follows:

- **A decoupled characterization of MLLM visual preference.** We identify an incompleteness in prior work's characterization of visual preference, decouple it into two semantically distinct Context Preference Vectors (CPVs), namely VFV (*external-vs-internal*) and MRV (*cross-modal*), and experiments reveal the intrinsic heterogeneity of different models in visual-preference responses and the divergence between the two CPVs' behavioral responses.

- **The CAS framework.** We propose **Context-Preference Activation Steering (CAS)**, a training-free inference-time intervention framework. CAS first extracts the two CPVs from a small set of conflict-based samples by associating candidate-answer preferences with the model's MLP-sublayer outputs before residual addition. In inference, it injects signed CPV residuals at the same sites in intermediate MLLM layers. CAS requires no parameter updates and uses the same extraction procedure across models, followed by lightweight calibration of its signed coefficients.

- **Systematic empirical evaluation and analysis.** Extensive experiments on different base models and benchmarks verify that CAS efficiently mitigates MLLM hallucinations while avoiding the repetitive degeneration of some decoding-based intervention methods. We further find that different MLLM layer bands respond markedly differently to the two CPVs. This offers a new analytical perspective on how multimodal models process

visual-linguistic semantic conflicts.

2 Related Work

2.1 Training-based Hallucination Mitigation

A straightforward approach to mitigating MLLM hallucinations is to construct hallucination-related preference data and perform alignment training on the model. DPO-based methods form preferences between *faithful* and *hallucinated* responses to suppress hallucinations. Recent object-aware and perception-enhanced DPO variants further extend this line by incorporating hallucination-aware preference construction and multimodal perception priors (Compagnoni et al., 2025; Feng et al., 2026). For example, Re-Align (Xing et al., 2025) constructs dual text-visual preference pairs via image retrieval and fine-tunes VLMs with retrieval-augmented DPO; RLHF-V (Yu et al., 2024) brings RLHF ideas into MLLM hallucination mitigation by collecting segment-level fine-grained correctional feedback and performing behavioral alignment via Dense DPO. These methods yield stable gains but require large preference datasets and substantial training resources. Besides, aligned models sometimes underperform the original VLM on general tasks (Xing et al., 2025), and the coarse-grained preferences of standard RLHF are prone to *reward hacking* and *behavior degeneration* (Yu et al., 2024). Consequently, even with substantial human correction and computational resources, training-based methods struggle to suppress hallucinations without compromising general capability.

2.2 Training-free Mitigation

Beyond alignment training, a complementary family of methods mitigates hallucinations primarily through inference-time procedures, although some approaches learn lightweight auxiliary components while keeping the base VLM fixed. We review three sub-classes below.

2.2.1 Detect-then-Correct Workflows

These methods construct detect-then-correct post-hoc workflows that perform multi-stage review and rewriting of generated outputs. LURE (Zhou et al., 2024) trains a hallucination revisor that takes the original caption, localizes hallucinated objects, and rewrites them into faithful descriptions. Woodpecker (Yin et al., 2024) further decomposes detection and correction into five stages, sequentially invoking external VLMs to complete sub-tasks such as key concept extraction, visual claim verification, and text rewriting. These post-hoc revisors can be attached to a frozen base VLM at inference time, but their sequential use of external VLMs and detectors introduces substantial latency and computational overhead. REVERSE instead combines hallucination-aware training with on-the-fly retrospective verification and resampling, allowing the model to revise unsupported generations during decoding (Wu et al., 2025); it therefore represents a train-then-verify approach rather than a purely plug-and-play post-hoc corrector.

2.2.2 Retrospection and Layer-Contrastive Decoding

These methods assume that the model already encodes implicit signals of *relative faithfulness*, which can be detected and exploited during decoding to suppress hallucinations. OPERA (Huang et al., 2024) detects an 'over-trust' attention pattern during beam search and backtracks to reselect tokens. DoLa (Chuang et al., 2024) contrasts vocabulary distributions from different layers, whereas DeCo (Wang et al., 2025) dynamically selects preceding layers and integrates their vocabulary-projected knowledge into the final-layer logits. Recent work further develops this line through learned positive and negative image projections in DCD and attention-steerable contrastive signals in ASCD (Chen et al., 2025; Wang et al., 2026). DCD requires learning lightweight auxiliary projections and is therefore not strictly training-free, although its contrastive correction is applied at decoding time. Complementary work uses gradient-aware token saliency and rejection sampling to detect and suppress impending hallucinations (Zhang et al., 2026). However, retrospection and layer-aware decoding can disrupt streaming output, incur additional inference overhead, or be sensitive to the choice of layers; strong layer contrast may also induce repetitive degeneration.

2.2.3 Methods under the Visual-Neglect Hypothesis

Within this family, methods build on the hypothesis that hallucinations stem from *visual neglect* and encourage models to *pay more attention to the image* by enhancing reliance on visual signals or suppressing textual priors. More broadly, inference-time activation steering builds on evidence that behavioral attributes can be identified and controlled through directions in internal rep-

resentation space (Li et al., 2023a; Turner et al., 2023; Zou et al., 2023). On the *decoding side*, VCD (Leng et al., 2024) contrasts the output logits of the original image against those of a Gaussian-noised image to suppress residual linguistic priors that persist after visual degradation. On the *forward-pass side*, PAI (Liu et al., 2024b) amplifies the attention weights of image tokens, complemented by contrastive logits from text-only input. AttnReal (Tu et al., 2025) redistributes the excessive attention accumulated on historical tokens back to visual tokens. DMAS (Yin et al., 2026) injects dynamically retrieved steering vectors into deep-layer attention heads; SSL (Hua et al., 2025) leverages a pretrained sparse autoencoder to apply predefined faithful and hallucination feature vectors to visual and output tokens; V-ITI (Sun et al., 2025) trains per-head linear classifiers to detect *visual neglect* and selectively inject visual signals. Other work extends this line with caption-sensitive attention intervention and adaptive visual reinforcement strategies (Li et al., 2025; Zhu et al., 2026).

Despite the success of these methods, there remain some limitations: (i) the assumption that *amplifying visual signals suffices to reduce hallucinations* overlooks the behavioral heterogeneity across different base models, resulting in limited hallucination reduction on some models; (ii) strong interventions may lead to repetitive degeneration, as shown in our experiments (§4.2); (iii) the extraction of steering signals in existing activation-steering methods often relies on external models or pretrained resources.

Unlike previous methods, CAS introduces a decoupled view of visual preference and injects two signed CPVs, allowing the intervention to follow different base models' empirical responses. The CPVs can be extracted within minutes and require only a small amount of context-conflict data. CAS operates within the model's standard autoregressive forward pass, without an additional forward pass per decoding step, and targets only the MLP-sublayer outputs before residual addition at four mid-early layers, achieving high efficiency while avoiding the repetitive degeneration observed for several baselines.

3 Methodology

We propose **Context-Preference Activation Steering (CAS)** (Figure 2), a training-free framework for mitigating MLLM hallucinations. CAS consists of two components: (i) **Extracting Context Preference Vectors** (ğ3.1); (ii) **Signed Activation Steering** (ğ3.2).

3.1 Extracting Context Preference Vectors

We decouple visual preference into two axes: the visual context with the parametric knowledge and with the textual context, i.e., *external-vs-internal* and *cross-modal*.

Setup I: Counterfactual Image Construction. To extract the **Visual Fidelity Vector (VFV)** that captures the *visual context vs. parametric knowledge* preference, we construct a set of counterfactual images that violate commonsense. Each sample takes the form $(I_{\text{cf}}, q, y^+, y^-)$, where I_{cf} is the counterfactual image, q is the question, y^+ corresponds to the image content, and y^- corresponds to commonsense knowledge. Thus, the model's choice over (y^+, y^-) directly reflects its relative preference between the visual context and parametric knowledge.

Setup II: Symmetric Image-Text Conflict Pairs. To extract the **Modality Reliance Vector (MRV)** that captures the *visual context vs. textual context* preference, we first select a pair of equally weighted but semantically distinct candidate answers (A, B) (e.g., *apple* vs. *orange*; *blue* vs. *green*). To eliminate the interference caused by any prior preference of the model toward one specific answer (i.e., the residual influence of parametric knowledge), we construct two symmetric image-text subsets:

$$\mathcal{S}_a = \{(I_A, q_B, y^+ = A, y^- = B)\}, \quad (1)$$
$$\mathcal{S}_b = \{(I_B, q_A, y^+ = B, y^- = A)\}, \quad (2)$$

where I_X denotes an image whose content is X, and q_X denotes a text prompt whose content is X. Since the answers favoring visual evidence and those favoring the textual context are flipped between $\mathcal{S}_a$ and $\mathcal{S}_b$, jointly fitting the two subsets helps reduce biases associated with answer-specific token frequencies. This symmetric design therefore encourages the fitted MRV to reflect cross-modal preference rather than vocabulary-frequency artifacts.

The detailed data construction of the two sample sets is provided in Appendix D.

Context Preference Vectors Extraction. The two setups share the same extraction procedure.

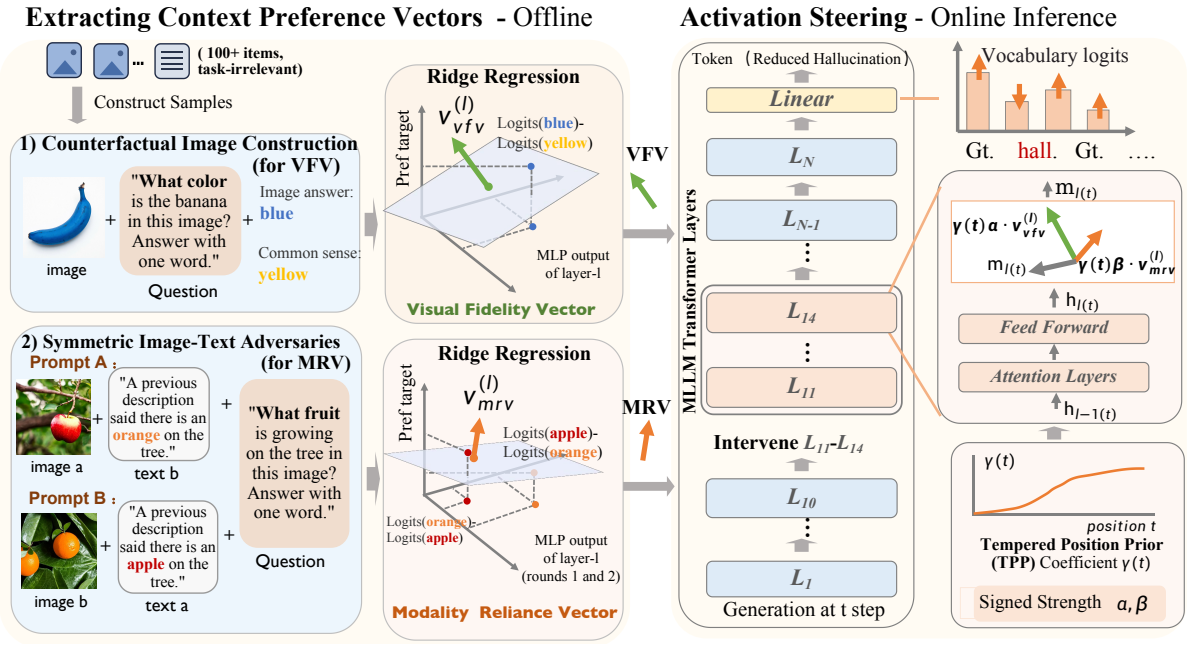


Figure 2: **Overview of the CAS framework. Left: Extracting Context Preference Vectors (offline).** Construct two types of context-conflict setups, read the layer-l MLP-sublayer output before residual addition together with the first-token logit difference pref, and obtain $v_{\text{vfv}}^{(l)}$ (VFV) and $v_{\text{mrv}}^{(l)}$ (MRV) via layer-wise ridge regression. **Right: Activation Steering (online inference).** At inference, we apply an additive residual $\gamma(t) \cdot (\alpha\, v_{\text{vfv}}^{(l)} + \beta\, v_{\text{mrv}}^{(l)})$ only to the MLP-sublayer outputs, before residual addition, of the mid-early layers L11–14, where α, β are signed-strength coefficients and $\gamma(t)$ is the position-wise gate given by the Tempered Position Prior.

We choose the MLP-sublayer outputs as the extraction site, motivated by prior evidence that Transformer MLP sublayers represent and mediate factual associations (Geva et al., 2021; Meng et al., 2022). For each sample, we let the MLLM autoregressively generate the answer, locate the earliest occurrence of y^+ or y^- within the generated token sequence, and record the position of its first token as $t^\star$. At the preceding position $t^\star - 1$, we read the next-token logits $\ell \in \mathbb{R}^{|V|}$ that predict the token at $t^\star$, and define the continuous preference signal.

$$\text{pref} = \max_{i \in \mathcal{T}^+} \ell_i - \max_{j \in \mathcal{T}^-} \ell_j, \tag{3}$$

where $\mathcal{T}^+$ is the set of first-token candidates for the answer y^+ aligned with the visual context, and $\mathcal{T}^-$ is the set of first-token candidates for the answer y^- aligned with the opposing context (parametric knowledge or textual context), with case and leading-space variants included. At the same position $t^\star - 1$, we record the output $m_l \in \mathbb{R}^d$ of the MLP sublayer before it is added to the residual stream at every layer l, and perform a per-layer ℓ_2-regularized ridge regression

$$\min_{w_l, b_l} \sum_s \left(\text{pref}_s - w_l^\top m_l^{(s)} - b_l\right)^2 + \lambda \|w_l\|^2, \tag{4}$$

where samples whose behavioral signal is weak ($|\text{pref}| < \epsilon$) are uniformly removed before fitting. The weight vector $w_l \in \mathbb{R}^d$ obtained independently at each layer yields the corresponding Context Preference Vector: the weight fitted on Setup I is denoted $v_{\text{vfv}}^{(l)}$ (VFV); the weight jointly fitted over the two halves $\mathcal{S}_a \cup \mathcal{S}_b$ of Setup II is denoted $v_{\text{mrv}}^{(l)}$ (MRV).

Layer localization. We use zero-based layer indexing throughout the paper. We fix both CPV extraction and residual injection to the mid-early MLP layers L11–14; this band is determined by the layer-wise attribution experiments reported in ğ4.5.

3.2 Signed Activation Steering

At inference, CAS formalizes the intervention as a **signed residual injection within the standard autoregressive forward pass**, requiring no additional forward pass at each decoding step. For a target layer $l \in \{11, 12, 13, 14\}$ and a generation step t,

$$\hat{m}_l^{(t)} = m_l^{(t)} + \gamma(t) \cdot \left(\alpha\, v_{\text{vfv}}^{(l)} + \beta\, v_{\text{mrv}}^{(l)}\right), \tag{5}$$

where α and β are the **signed** strength coefficients of VFV and MRV, respectively, and $\gamma(t)$ is the

position-wise temporal gate. The modified MLP-sublayer output $\hat{m}_l^{(t)}$ is subsequently added to the residual stream by the original Transformer layer. During prefill, the intervention is applied to the last prompt token. During decoding, it is triggered once per newly generated token.

Tempered Position Prior (TPP) In long-form generation, the hallucination rate varies markedly with decoding time, i.e., later positions are increasingly prone to drifting away from the visual evidence (Zheng et al., 2025; Zhou et al., 2024). To trade off intervention strength against caption-level F1 or coverage, we optionally couple the intervention strength to a position-level hallucination probability, assigning stronger interventions to high-risk positions. Specifically, we estimate the prior hallucination probability of each position on a separate calibration set by calculating the fraction of hallucinated object tokens generated by the base model at each position. Because the per-token statistic exhibits high variance and adjacent positions are intrinsically correlated, we apply variable-length position bucketing (details in Appendix K) to aggregate neighboring positions and suppress random noise, yielding a per-bucket hallucination fraction P_b. Then, we apply a power transform with temperature T to obtain the bucket coefficient:

$$c_b = \left(P_b \,/\, \max_{b'} P_{b'}\right)^{1/T}, \qquad b \in \mathcal{B}. \quad (6)$$

$T > 1$ raises the low-bucket coefficients (making buckets more uniform); $T = 1$ recovers the original distribution. Generative tasks may take $\gamma(t) = c_{b(t)}$ as the dynamic temporal gate, while discriminative tasks, which output only a single token without autoregressive drift, set $\gamma(t) \equiv 1$. To isolate the independent impacts of the steering directions, we set $\gamma(t) \equiv 1$ in the per-CPV and layer ablations. The position dependence of c_b is fit on the calibration set. The full calibration protocol, TPP ablation, bucket boundaries, and other details are provided in Appendices F, G, and K.

4 Experiments

4.1 Experimental Setup

Models. We evaluate CAS on four open-source MLLMs: LLaVA-1.5 (Liu et al., 2024a), Shikra (Chen et al., 2023), Qwen-VL (Bai et al., 2023), and InstructBLIP (Dai et al., 2023). These models differ in cross-modal fusion design, allowing us to assess the cross-architecture applicability of CAS.

Benchmarks. We cover both generative and discriminative hallucination evaluations. Generative: (1) **CHAIR** (Rohrbach et al., 2018), on 500 MS COCO (Lin et al., 2014) images for detailed captioning, reporting CHAIR_S, CHAIR_I, and F1; (2) **AMBER** (Wang et al., 2024) generative split (1,004 images), reporting Hal, CHAIR, Cog, and Cover. Discriminative: (3) **POPE** (Li et al., 2023b), with 3,000 questions in each of the random / popular / adversarial subsets, reporting Acc and F1.

Baselines. We compare vanilla greedy decoding against eight training-free methods, ordered by publication date: DoLa (Chuang et al., 2024), VCD (Leng et al., 2024), OPERA (Huang et al., 2024), PAI (Liu et al., 2024b), CODE (Kim et al., 2024), DeCo (Wang et al., 2025), AttnReal (Tu et al., 2025), and SSL (Hua et al., 2025). All methods follow their original recommended configurations.

Text-quality metrics. Some baselines exhibit repetitive generation (representative cases in Appendix L). Such repetition is not penalized by CHAIR / Hal / F1; worse, because CHAIR_I is counted at the token-instance level without deduplication, repeated GT-grounded objects deflate its denominator. We therefore additionally report **Rep** (n-gram repetition rate) for generative tasks. Definitions are provided in Appendix K.

CAS configuration. CAS injects signed residuals into the MLP-sublayer outputs before residual addition at layers L11–14 within each standard autoregressive forward pass, requiring no additional forward pass per decoding step. Full hyperparameter configurations are listed in Appendix E.

Decoding & hardware. All methods use greedy decoding to eliminate sampling variance, with `max_new_tokens=512` for CHAIR / AMBER and 64 for POPE. All experiments are conducted on a single NVIDIA RTX 3090. Prompt templates for each benchmark are listed in Appendix K.

4.2 Generative Hallucination

CHAIR results are shown in Table 1. CAS attains the lowest or tied-lowest CHAIR_S on all base models, reducing CHAIR_S relative to vanilla by 37.8% / 27.9% / 25.2% / 31.6% on LLaVA-1.5 / Shikra / Qwen-VL / InstructBLIP, respectively. DoLa, OPERA, PAI, AttnReal, and SSL exhibit

Method	LLaVA-1.5			Shikra			Qwen-VL			InstructBLIP		
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Rep	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Rep	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Rep	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Rep
Vanilla	57.6	17.79	1.7	61.0	18.58	0.4	50.8	15.00	0.6	52.6	16.83	4.4
DoLa	74.4	27.23	9.4	50.6	16.78	285.7	53.6	16.71	0.2	73.0	33.54	38.3
VCD	57.6	18.82	2.0	64.0	19.49	0.2	51.0	14.99	0.5	54.8	17.44	4.8
OPERA	70.4	22.21	66.3	57.4	21.37	17.7	52.0	21.62	0.0	59.6	20.29	0.2
PAI	37.4	**10.13**	27.2	49.6	14.84	1.7	–	–	–	–	–	–
CODE	51.8	16.90	1.1	62.2	19.24	0.9	53.6	16.84	0.6	52.2	17.41	3.6
DeCo	46.2	14.10	5.0	54.0	16.06	0.6	45.4	13.16	1.1	45.4	14.49	9.8
AttnReal	45.4	13.97	8.4	60.4	17.34	1.1	49.0	13.30	1.6	47.8	14.93	19.2
SSL	44.8	15.53	7.6	–	–	–	–	–	–	48.0	15.74	5.6
CAS	**35.8**	12.21	0.8	**44.0**	**13.56**	0.6	**38.0**	**12.19**	1.4	**36.0**	**13.01**	3.8

Table 1: CHAIR results. $CHAIR_S$ / $CHAIR_I$ lower is better. Rep ($\times 10^{-3}$); Rep $\geq 3\times$ Vanilla highlighted in red. Best $CHAIR_S$ / $CHAIR_I$ per model in **bold**, second best underlined. "–" indicates the method is not available for that base model. Full results including F1 and TTR are reported in Appendix I (Table 15).

Model	Method	Hal↓	CHAIR↓	Cog↓	Cover↑	Rep
LLaVA-1.5	Vanilla	31.1	7.7	3.6	49.4	0.5
	VCD	36.2	8.7	3.9	50.4	0.5
	PAI	23.2	**4.4**	**1.6**	50.5	29.0
	AttnReal	26.0	5.3	2.4	50.2	4.9
	SSL	31.6	8.3	3.4	44.2	0.4
	CAS	**17.2**	**4.4**	**1.6**	46.7	0.1
Shikra	Vanilla	48.5	10.8	5.5	51.4	0.3
	VCD	50.0	10.7	4.9	52.3	0.2
	PAI	**21.4**	8.2	**1.9**	33.5	1.4
	AttnReal	43.3	8.7	4.3	51.7	0.5
	CAS	34.2	**7.6**	2.9	50.5	0.1
Qwen-VL	Vanilla	29.6	6.3	2.6	53.3	0.2
	VCD	37.5	8.8	3.4	50.8	0.2
	AttnReal	25.9	**5.3**	2.1	54.2	0.7
	CAS	**23.5**	**5.3**	**1.8**	50.2	0.5
InstructBLIP	Vanilla	35.3	8.1	3.7	53.8	2.3
	VCD	39.1	8.3	3.7	54.2	2.1
	AttnReal	30.5	**5.7**	2.9	54.3	27.2
	CAS	**28.8**	6.1	**2.8**	51.9	2.6

Table 2: AMBER generative results. Hal / CHAIR / Cog lower is better (↓); Cover higher is better (↑). Rep ($\times 10^{-3}$); Rep values $\geq 3\times$ Vanilla and abnormal Cover collapses are highlighted in red. Cover / Rep are reported as reference and are otherwise not emphasized; **bold** marks best per column among the three hallucination metrics (Hal / CHAIR / Cog), second best underlined.

red-flagged Rep values on multiple base models, indicating visible repetitive degeneration, whereas CAS's Rep stays on the order of vanilla. Full results (including F1 and TTR) are reported in Appendix I, where CAS's F1 is no lower than vanilla on any base model; length-aligned CHAIR results are reported in Appendix C.

AMBER results are shown in Table 2. CAS attains the lowest Hal on LLaVA, Qwen, and InstructBLIP, with a modest decrease in Cover; relative to vanilla, CAS reduces Hal by 44.7% / 20.6% / 18.4% on the three models. On Shikra, PAI attains a lower Hal but its Cover collapses by 34.8% (51.4 → 33.5, red); CAS instead reduces Hal by 29.5% while preserving Cover.

Table 3: POPE results (averaged over random / popular / adversarial subsets). Best per column within each model in **bold**, second best underlined. "–" indicates the method is not available for that base model. Per-subset breakdown in Appendix J (Table 16).

Method	LLaVA-1.5		Shikra	
	Acc↑	F1↑	Acc↑	F1↑
Greedy	85.17	83.49	77.21	80.23
DoLa	77.78	75.12	63.30	70.28
VCD	84.37	83.22	74.13	78.46
OPERA	85.26	83.63	76.98	79.42
PAI	85.17	83.83	78.37	75.26
CODE	85.04	83.29	78.50	80.82
DeCo	85.17	83.49	70.44	76.18
SSL	84.94	83.11	–	–
CAS (ours)	**86.58**	**86.11**	**78.57**	**81.20**

Method	Qwen-VL		InstructBLIP	
	Acc↑	F1↑	Acc↑	F1↑
Greedy	87.67	87.10	76.42	79.51
DoLa	81.83	84.20	56.16	60.87
VCD	87.33	87.20	75.08	78.14
OPERA	86.25	87.13	69.23	72.90
CODE	87.31	87.12	77.06	79.72
DeCo	87.65	87.12	76.80	79.69
SSL	–	–	78.77	81.02
CAS (ours)	**87.84**	**87.53**	**84.69**	**84.74**

4.3 Discriminative Hallucination

POPE results are shown in Table 3 (averaged over the three subsets; per-subset breakdown in Appendix Table 16). CAS attains the top average Acc and F1 on all four models, improving F1 on InstructBLIP by 6.6% relative to vanilla. In the per-subset results, PAI's F1 on Shikra drops by 6.5% / 4.4% relative to vanilla on the popular / adversarial subsets, whereas CAS is no lower than vanilla on any (base model, subset) pair.

4.4 Per-model response to VFV and MRV

The intensity coefficients α and β of CAS are two signed scalars. Figure 3 reports VFV-only and

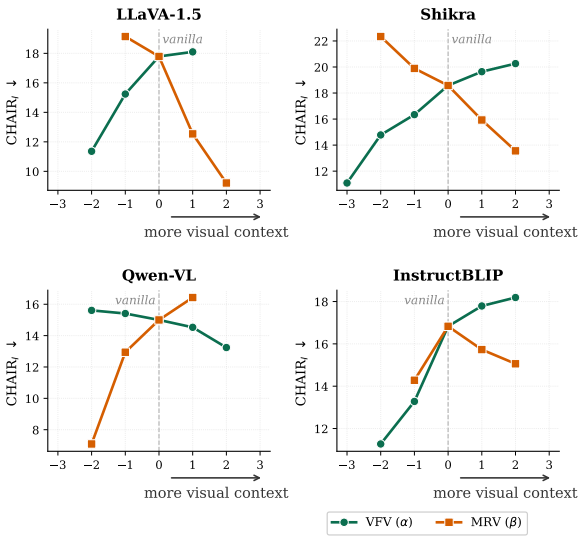


Figure 3: **CHAIR$_I$ under VFV-only and MRV-only intervention.** Each subplot corresponds to one base model and reports CHAIR$_I$ across signed intensities α (VFV) and β (MRV); the dashed line marks that model's vanilla CHAIR$_I$. Full numbers are in Appendix Tables 7 and 8.

MRV-only sweeps on the four base models (without positional priors). Models respond to VFV and MRV in markedly different ways: the four models already exhibit three distinct response patterns. More visual emphasis does not automatically lower hallucination; nevertheless, for every model both VFV and MRV admit signed settings that bring CHAIR$_I$ below vanilla, and some models are particularly sensitive to one of the two. CAS exploits this per-model response to substantially reduce hallucination (full numbers in Appendix Tables 7 and 8).

VFV and MRV are geometrically non-redundant. Since both VFV and MRV concern vision, we further investigate whether they are redundant estimates of the same underlying preference. We compute $\cos(v^{(l)}_{\text{vfv}}, v^{(l)}_{\text{mrv}})$ on the intervention layers (L11–14) of four base models. All 16 values fall within $[-0.06, +0.11]$, with mean $|\cos| \approx 0.033$, matching the order of magnitude expected of two random vectors in $\mathbb{R}^{4096}$. Thus, in the examined layers, the fitted VFV and MRV directions are nearly orthogonal and are not simple geometric duplicates.

Random and shuffled-label controls. The learned CPVs outperform norm-matched random and shuffled-label controls (Table 4), showing that the gains do not arise from residual injection alone.

Direction	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Rep
Vanilla	57.6	17.79	1.7
VFV	**48.2**	**15.24**	2.2
Random VFV	54.6	17.13	3.2
Shuffled-label VFV	59.4	19.05	4.9
MRV	**36.6**	**12.54**	1.0
Random MRV	48.8	15.45	0.8
Shuffled-label MRV	55.8	15.82	12.1

Table 4: Direction controls for LLaVA-1.5 on CHAIR. Random vectors are norm-matched; shuffled-label vectors use permuted preference targets. All interventions use L11–14 without TPP, with $\alpha{=}-1$ for VFV and $\beta{=}+1$ for MRV. Rep is reported in units of 10^{-3}.

4.5 Layer Analysis

To verify the choice of intervention layers, we place the same CPVs into different four-layer windows for LLaVA-1.5 on CHAIR. To amplify the differences between windows, we use stronger fixed intensities (MRV $\beta{=}+2$, VFV $\alpha{=}-2$). As shown in Table 5, the two CPVs follow a consistent pattern across windows: within the windows we sweep, the mid-early band L11–14 is the only band that is both effective and stable; from L16 onward the effect returns to the vanilla level and is essentially null; at the shallow layers L1–4 and L6–9, both MRV and VFV trigger severe degeneration and diverge into runaway repetition. We therefore fix the intervention at L11–14 in all experiments.

5 Conclusion

This paper challenges the prevailing assumption that MLLM object hallucinations stem solely from visual neglect. Our intervention results show that increasing visual reliance can unexpectedly exacerbate hallucinations and are consistent with a model-dependent balance among image context, parametric knowledge, and textual context. By decoupling visual preference into two distinct dimensions, represented by the Visual Fidelity Vector (VFV) and the Modality Reliance Vector (MRV), we introduce Context-Preference Activation Steering (CAS), a lightweight training-free framework for adjusting information reliance during inference. By injecting signed residuals of these vectors into the MLP-sublayer outputs before residual addition at four mid-early layers within the standard autoregressive forward pass, CAS adapts to each model's empirical context-preference re-

Window	CPV	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
L1–4	MRV	*degenerate, Rep* $\approx 575\times$	
	VFV	*degenerate, Rep* $\approx 587\times$	
L6–9	MRV	*degenerate, Rep* $\approx 278\times$	
	VFV	*degenerate, Rep* $\approx 280\times$	
L11–14 (selected)	MRV	**18.8**	**9.21**
	VFV	**35.4**	**11.36**
L16–19	MRV	58.8	18.35
	VFV	56.4	18.04
L21–24	MRV	56.8	17.33
	VFV	55.8	17.75
L26–29	MRV	55.8	18.19
	VFV	57.2	18.23
vanilla		57.6	17.79

Table 5: **Layer ablation.** Effect of MRV (β=+2) and VFV (α=−2) interventions across different four-layer windows on LLaVA-CHAIR. $CHAIR_S$ / $CHAIR_I$ lower is better. Degenerate rows are marked in the CHAIR columns as "Rep $N\times$", where N is the Rep multiplier relative to vanilla. The selected window **L11–14** is in bold.

sponse without requiring an additional forward pass per decoding step. Evaluations across four 7B-scale MLLMs and three benchmarks show that CAS substantially suppresses hallucinations with negligible decoding overhead while maintaining the reported F1, lexical-diversity, and repetition metrics. A preliminary experiment on LLaVA-1.5-13B further supports transfer beyond the primary 7B setting. We hope that treating visual preference as a per-model control signal rather than a uniform *more-visual* intervention provides a useful perspective for inference-time MLLM mitigation.

Limitations

Although Context-Preference Activation Steering (CAS) demonstrates empirical effectiveness in mitigating hallucinations in Multimodal Large Language Models (MLLMs), several limitations remain. First, our evaluation primarily focuses on object hallucination; effectiveness on attribute errors, spatial and logical relation errors, and more complex multimodal reasoning hallucinations has not been systematically verified. Second, the main experiments use four 7B-scale models. We additionally report a preliminary transfer experiment on LLaVA-1.5-13B, but this single result is insufficient to establish broad generalization to larger or more recent architectures. The optimal intervention window may also shift in substantially deeper models. Third, CAS requires lightweight model- and task-specific calibration of the signed steering coefficients, and some generative settings additionally use a calibrated position prior; it is therefore not a zero-calibration method. Furthermore, limitations in the fidelity and diversity of the generated conflict samples may introduce visual or textual biases into the extracted vectors. Finally, our interventions establish empirical directional effects but do not constitute a complete mechanistic or causal explanation of hallucination or multimodal feature fusion.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grants 62576171, 62476136, and 62276138.

References

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

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. 2025. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2504.08809*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Alberto Compagnoni, Davide Caffagni, Nicholas Moratelli, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2025. Mitigating hallucinations in multimodal llms via object-aware preference optimization. *arXiv preprint arXiv:2508.20181*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiawei Feng, Jiancan Wu, Xingyu Zhu, Junkang Wu, Xiang Wang, and Xiangnan He. 2026. Pea-dpo: Perception-enhanced alignment direct preference optimization for mllms alignment. *arXiv preprint arXiv:2608.19598*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhenglin Hua, Jinghan He, Zijun Yao, Tianxu Han, Haiyun Guo, Yuheng Jia, and Junfeng Fang. 2025. Steering LVLMs via sparse autoencoder for hallucination mitigation. *arXiv preprint arXiv:2505.16146*.

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)*.

Junho Kim, Hyun J Kim, Yeon J Kim, and Yong M Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599.

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)*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*. Spotlight.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Libo Qin, Ruihan Chen, Baohang Li, Kui Jiang, Yaowei Wang, Ting Liu, and Bing Qin. 2025. Cai: Caption-sensitive attention intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2506.23590*.

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 (EMNLP)*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Highlight.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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 (EMNLP)*, Brussels, Belgium. Association for Computational Linguistics.

Nan Sun, Zhenyu Zhang, Xixun Lin, Kun Wang, Yanmin Shang, Naibin Gu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang, and Yanan Cao. 2025. V-iti: Mitigating hallucinations in multimodal large language models via visual inference-time intervention. *arXiv preprint arXiv:2512.03542*.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of MLLMs. *arXiv preprint arXiv:2503.08342*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Yujun Wang, Aniri, Jinhe Bi, Soeren Pirk, and Yunpu Ma. 2026. ASCD: Attention-steerable contrastive decoding for reducing hallucination in MLLM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10306–10314.

Tsung-Han Patrick Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M.

Chan. 2025. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. In *Advances in Neural Information Processing Systems*, volume 38, pages 65749–65777.

Shuo Xing, Peiran Li, Yuping Wang, Ruizheng Bai, Yueqi Wang, Chan-Wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. 2025. Re-Align: Aligning vision language models via retrieval-augmented direct preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2379–2397, Suzhou, China. Association for Computational Linguistics.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. 2026. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. 2026. Hallucination begins where saliency drops. *arXiv preprint arXiv:2601.20279*.

Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibei Yang. 2025. Why LVLMs are more prone to hallucinations in longer responses: The role of context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, and Hanwang Zhang. 2026. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. *arXiv preprint arXiv:2602.24041*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Inference Latency

We measure the per-token decoding latency of each method on LLaVA-1.5; results are reported in Table 6. CAS performs only an additive residual on the MLP-sublayer outputs, before residual addition, at four mid-early layers within the standard autoregressive forward pass, with no additional forward pass per decoding step and no contrastive decoding. Its per-token latency is therefore close to vanilla and substantially lower than double-forward methods (PAI / VCD / CODE) and multi-step search methods (OPERA).

Method	Per-token Latency (ms)	std
Vanilla	29.65	±2.76
CAS (ours)	**28.09**	±1.05
SSL	28.49	±1.05
DoLa	33.68	±1.40
AttnReal	35.08	±2.56
DeCo	40.97	±1.22
PAI	56.35	±1.33
VCD	56.04	±1.10
CODE	113.05	±12.33
OPERA	464.03	±385.38

Table 6: Per-token decoding latency (ms) on the LLaVA-1.5 setting. The only extra cost of CAS comes from one additive residual at each selected mid-early MLP-sublayer output before residual addition, yielding latency of the same order as vanilla.

B Per-CPV Sweeps: MRV-only and VFV-only on Four Models

For each of the four base models, we sweep MRV-only (α=0, varying β) and VFV-only (β=0, varying α). Full results are reported in Tables 7 and 8. Three observations follow. (i) For each base model, both VFV and MRV admit signed settings that bring CHAIR_S below vanilla. (ii) The beneficial sign combination (α, β) varies across base models; at least three distinct sign patterns therefore coexist across the four base models. (iii) Extreme intensities can cause repetitive degeneration; nevertheless, for every base model there exists an intensity at which the CPV substantially reduces hallucination without such degeneration, and CAS operates within this regime.

Model	β	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Rep
LLaVA-1.5	−2	*degenerate*		828.2
	−1	67.8	19.14	12.7
	0	57.6	17.79	1.7
	+1	36.6	12.54	1.0
	+2	18.8	9.21	2.6
Shikra	−2	68.9	22.34	1.6
	−1	65.8	19.89	0.6
	0	61.0	18.58	0.4
	+1	55.2	15.93	0.3
	+2	44.0	13.56	0.7
Qwen-VL	−2	21.0	7.09	2.5
	−1	41.0	12.94	1.1
	0	50.8	15.00	0.6
	+1	53.8	16.43	0.7
	+2	*degenerate*		29.8
InstructBLIP	−2	*degenerate*		110.1
	−1	38.4	14.28	12.6
	0	52.6	16.83	4.4
	+1	49.4	15.73	3.8
	+2	50.0	15.06	5.8

Table 7: **MRV-only sweep** (α=0, CHAIR, no position prior). Rep is reported in units of 10^{-3}; values $\geq 3\times$ the vanilla row (β=0) of each model are highlighted in red.

Model	α	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Rep
LLaVA-1.5	+2	*degenerate*		121.3
	+1	53.0	18.10	1.5
	0	57.6	17.79	1.7
	−1	48.2	15.24	2.2
	−2	35.4	11.36	5.3
Shikra	+2	68.2	20.26	0.5
	+1	64.0	19.64	0.7
	0	61.0	18.58	0.4
	−1	53.6	16.34	0.7
	−2	44.6	14.78	0.2
	−3	36.8	11.09	0.4
Qwen-VL	+3	*degenerate*		10.6
	+2	41.8	13.24	2.7
	+1	48.2	14.53	1.2
	0	50.8	15.00	0.6
	−1	50.6	15.41	0.6
	−2	49.8	15.61	0.8
InstructBLIP	+2	59.6	18.19	8.5
	+1	58.6	17.79	5.9
	0	52.6	16.83	4.4
	−1	40.0	13.28	3.5
	−2	32.6	11.27	5.7

Table 8: **VFV-only sweep** (β=0, CHAIR, no position prior). Rep is reported in units of 10^{-3}; values $\geq 3\times$ the vanilla row (α=0) of each model are highlighted in red.

C Length-aligned CHAIR

We truncate both CAS and vanilla captions to the first 64 tokens and recompute CHAIR. Table 9 shows that under this equal-budget setting, CAS's CHAIR$_S$ is substantially lower than vanilla on all four base models, while F1 is no lower than vanilla; this confirms that CAS's hallucination reduction also holds under shorter, length-aligned captions.

D CPV Sample Construction

CAS extracts both CPVs from on the order of one hundred samples (Table 10), far below the thousands-to-tens-of-thousands of prompts typically required by contrastive decoding or alignment fine-tuning. This data efficiency stems from the paradigm design rather than from any released dataset: context-conflict samples provide a high-SNR differential signal in the latent space (each sample carries its own (y^+, y^-) candidate pair, and the symmetric image–text conflict pairs further cancel token-frequency bias via symmetric flipping), so ridge regression converges stably with on the order of one hundred samples. We do not claim that these samples constitute a high-quality dataset; on the contrary, as discussed in Limitations, the images of the symmetric image–text conflict pairs are generated by a 4 GB-scale SD 1.5 model and are of limited quality. The core contribution is the paradigm of context-conflict samples combined with cross-symmetric subtraction, not the released samples themselves.

Prompts and candidate answers. All (q, y^+, y^-) triples (i.e., a question template paired with two candidate answers) are first drafted in a single LLM call and then verified manually. They cover more than ten commonsense concepts including color, material, shape, and counting; different samples within the same concept share the same question template with the specific object swapped, and the concept serves as the cross-validation grouping unit (samples of the same concept do not cross folds).

Counterfactual image generation. The counterfactual-image samples require images that violate commonsense (e.g., *a blue banana*), which places a relatively high demand on the T2I model's text faithfulness: smaller T2I models tend to *self-correct* the color back to the commonsense value. We therefore use the commercial-grade Qwen-Image-2.0 to generate the 51 images.

Symmetric image–text conflict pair generation. The symmetric image–text conflict pairs only re-

Model	Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	F1 ↑
LLaVA-1.5	Vanilla	43.4	13.51	74.76
	CAS	**30.8**	**10.72**	**75.09**
Shikra	Vanilla	41.8	14.17	70.94
	CAS	**32.0**	**10.43**	**74.08**
Qwen-VL	Vanilla	30.4	**9.80**	73.84
	CAS	**28.4**	9.95	**74.78**
InstructBLIP	Vanilla	38.6	12.76	73.33
	CAS	**28.2**	**9.97**	73.31

Table 9: **Length-aligned CHAIR.** CHAIR_S / CHAIR_I (lower is better) and F1 (higher is better) recomputed after truncating all captions to the first 64 tokens. The best value within each model is in **bold**.

Table 10: CPV sample counts. VFV is fitted alone on the counterfactual-image samples; MRV is jointly fitted on the two halves $\mathcal{S}_a \cup \mathcal{S}_b$ of the symmetric image–text conflict pairs, where the opposite-sign pref of the two halves automatically realizes cross-symmetric subtraction. Each sample carries a pair of first-token candidates (y^+, y^-), which together define a continuous pref signal on that sample.

Setup	Form	Count
Counterfactual image	$(I_{\text{cf}}, q, y^+, y^-)$	51
Symmetric conflict pair $\mathcal{S}_a$	$(I_A, q_B, y^+{=}A, y^-{=}B)$	55
Symmetric conflict pair $\mathcal{S}_b$	$(I_B, q_A, y^+{=}B, y^-{=}A)$	55
Total		161

quire images of plain and unambiguous objects (e.g., *pencil* vs. *pen*), and the demand on T2I text faithfulness is low. We therefore use the open-source 4 GB-scale Stable Diffusion v1.5 to generate the 110 images, trading off resource cost. As discussed in Limitations, MRV may therefore inherit some visual bias from SD 1.5. Re-extracting MRV with a stronger T2I model is a direct improvement target for future work.

Sample construction and CPV fitting time. Once the samples are ready, for each base model we run a forward pass over the 161 samples, record the MLP-sublayer output before residual addition at the focus token, and perform per-layer ridge regression. The entire pipeline finishes within a few minutes on a single NVIDIA RTX 3090.

E Hyperparameter Configurations

Table 11 lists the hyperparameters per (base model, task type) combination: VFV intensity α, MRV intensity β, and the position-prior temperature T. For generative tasks, the two values of T, separated by a slash, correspond to CHAIR and AMBER, respectively. A "–" indicates that TPP is disabled. We find that the best (α, β) often differ between generative and discriminative tasks within the same base model; a plausible explanation is that the two task types place different demands on visual reliance: discriminative prompts explicitly name the queried object, giving the model a clear visual focus to locate, whereas generative captioning is free-form and has no such focus. The intervention layers are fixed at L11–14 across all base models.

Model	Task type	α	β	T
LLaVA-1.5	Generative	-2	$+1$	2 / 5
	Discriminative	$+2$	$+1$	–
Shikra	All	0	$+2$	–
Qwen-VL	Generative	$+1$	-1	– / 7
	Discriminative	$+1$	$+1$	–
InstructBLIP	Generative	-2	0	1 / –
	Discriminative	-2	0	–

Table 11: CAS hyperparameters.

F Calibration Protocol

The 161 context-conflict samples in Appendix D are used only to fit the CPVs. On a small held-out calibration set disjoint from the evaluation data, we scan each direction over the integer grid $\{-2, -1, +1, +2\}$, discard settings that cause repetitive degeneration, and retain settings that reduce hallucination without a material decrease in F1. The intervention window L11–14, ridge coefficient $\lambda = 1.0$, weak-signal threshold $\epsilon = 0.1$, and five-fold group-aware cross-validation are fixed across models.

TPP uses a single position prior estimated from LLaVA-1.5 captions generated on a separate calibration set of 500 MS COCO images, with no overlap with the 500-image CHAIR evaluation set. The resulting prior is shared across all models and tasks. Only the temperature varies across the generative settings in Table 11; discriminative tasks use $\gamma(t) = 1$.

G Effect of the Tempered Position Prior

TPP is an optional trade-off mechanism rather than a necessary component of the signed CPV intervention. "Off" sets $\gamma(t) = 1$ at all decoding positions; "On" applies the calibrated bucket coefficient in Eq. 6. Tables 12 and 13 report the CHAIR and AMBER ablations, respectively. In each case,

all settings other than TPP are held fixed between the on and off rows. Turning TPP off often yields a larger reduction in hallucination, whereas enabling it recovers part of the caption-level F1 or coverage. For settings in which this trade-off was not beneficial, TPP remains disabled, as shown in Table 11.

Model	TPP	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1↑
LLaVA-1.5	– (Vanilla)	57.60	17.79	75.20
	Off	22.60	10.25	71.89
	On	35.80	12.21	76.00
InstructBLIP	– (Vanilla)	52.60	16.83	72.90
	Off	32.60	11.27	73.83
	On	36.00	13.01	74.65

Table 12: TPP ablation on CHAIR. Only TPP is varied within each model.

Model	TPP	Hal↓	CHAIR↓	Cog↓	Cover↑
LLaVA-1.5	– (Vanilla)	31.10	7.70	3.60	49.40
	Off	16.50	4.50	1.00	45.00
	On	17.20	4.40	1.20	46.70
Qwen-VL	– (Vanilla)	29.60	6.30	2.60	53.30
	Off	23.40	6.00	1.60	48.60
	On	23.50	5.30	1.80	50.20

Table 13: TPP ablation on AMBER. Only TPP is varied within each model.

H Transfer to the 13B Scale

We additionally evaluate LLaVA-1.5-13B on CHAIR by directly reusing the configuration selected for LLaVA-1.5-7B, without any tuning on the 13B model. As shown in Table 14, CAS remains effective at the larger model scale while maintaining generation quality.

Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1↑	Rep
Vanilla	50.20	15.65	76.92	1.7
CAS	**42.80**	**13.94**	**77.15**	1.7

Table 14: LLaVA-1.5-13B results on CHAIR. CAS directly reuses the 7B configuration without additional tuning. Rep is reported in units of 10^{-3}.

I CHAIR Full Results

Table 15 reports the full CHAIR results including F1 and TTR (the main paper Table 1 only reports $CHAIR_S$ / $CHAIR_I$ / Rep). CAS's F1 is no lower than vanilla across all four base models, indicating that the CHAIR reduction does not come from precision / recall trade-off; TTR remains close to vanilla, confirming that lexical diversity is preserved.

J POPE Per-Subset Breakdown

Table 16 reports the full random / popular / adversarial breakdown of the POPE results that appear averaged in the main paper (Table 3).

K Implementation Details

CPV fitting.

- Ridge regression with $\lambda = 1.0$, solved in closed form independently per layer.
- 5-fold group-aware cross-validation, grouped by the commonsense concept each sample belongs to so that samples sharing a concept do not cross folds.
- Samples with $|\text{pref}| < 0.1$ are removed (signal too weak to constrain the regression).

Tempered Position Prior.

- 11 buckets in total: one per 5 tokens in $[0, 20)$, one per 10 tokens in $[20, 80)$, and a single tail bucket for ≥ 80.
- Bucket coefficient $c_b = (P_b / \max_{b'} P_{b'})^{1/T}$, with T from Table 11.
- A single prior is estimated from LLaVA-1.5 captions generated on a separate calibration set of 500 MS COCO images, with no overlap with the 500-image CHAIR evaluation set, and reused across models and tasks. We locate the token spans of hallucinated object words and compute the per-bucket fraction P_b.

Evaluation Prompts. We adopt standard prompt templates tailored to each evaluation benchmark. For **CHAIR**, LLaVA-1.5, Shikra, and InstructBLIP use the standard captioning prompt: `"Please describe this image in detail."`. However, we observe that under this default instruction, Qwen-VL produces noticeably shorter captions than the other three base models. To prevent caption length from becoming a confounding factor in our CHAIR comparison, we apply a slightly shorter variant for Qwen-VL: `"Describe this image in detail."` , which successfully aligns its average generation length with the other models. For **AMBER generative**,

Method	LLaVA-1.5					Shikra				
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1↑	Rep	TTR↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1↑	Rep	TTR↑
Vanilla	57.6	17.79	75.2	1.7	0.703	61.0	18.58	71.8	0.4	0.732
DoLa	74.4	27.23	68.3	9.4	0.747	50.6	16.78	68.6	285.7	0.489
VCD	57.6	18.82	74.0	2.0	0.713	64.0	19.49	72.2	0.2	0.748
OPERA	70.4	22.21	71.8	66.3	0.700	57.4	21.37	54.1	17.7	0.781
PAI	37.4	**10.13**	75.2	27.2	0.631	49.6	14.84	73.0	1.7	0.697
CODE	51.8	16.90	74.9	1.1	0.711	62.2	19.24	72.4	0.9	0.752
DeCo	46.2	14.10	75.5	5.0	0.713	54.0	16.06	74.5	0.6	0.737
AttnReal	45.4	13.97	**76.3**	8.4	0.658	60.4	17.34	74.0	1.1	0.719
SSL	44.8	15.53	74.1	7.6	0.693	–	–	–	–	–
CAS (ours)	**35.8**	12.21	76.0	0.8	0.703	**44.0**	**13.56**	**75.4**	0.6	0.733

Method	Qwen-VL					InstructBLIP				
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1↑	Rep	TTR↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1↑	Rep	TTR↑
Vanilla	50.8	15.00	75.4	0.6	0.728	52.6	16.83	72.9	4.4	0.687
DoLa	53.6	16.71	72.5	0.2	0.746	73.0	33.54	64.8	38.3	0.680
VCD	51.0	14.99	74.3	0.5	0.763	54.8	17.44	73.0	4.8	0.691
OPERA	52.0	21.62	55.9	0.0	0.836	59.6	20.29	54.6	0.2	0.777
CODE	53.6	16.84	73.4	0.6	0.735	52.2	17.41	72.3	3.6	0.701
DeCo	45.4	13.16	75.6	1.1	0.727	45.4	14.49	71.3	9.8	0.693
AttnReal	49.0	13.30	**76.1**	1.6	0.696	47.8	14.93	73.9	19.2	0.643
SSL	–	–	–	–	–	48.0	15.74	73.1	5.6	0.672
CAS (ours)	**38.0**	**12.19**	**76.1**	1.4	0.717	**36.0**	**13.01**	**74.7**	3.8	0.691

Table 15: CHAIR full results. $CHAIR_S$ / $CHAIR_I$ lower is better; F1 / TTR higher is better. Rep ($\times 10^{-3}$); Rep $\geq 3\times$ Vanilla highlighted in red; Rep / TTR are reported as reference only and not bolded. Best per column within each model in **bold**, second best underlined.

we strictly adhere to the official evaluation protocol using `"Describe this image."`. For **POPE**, all models are evaluated under the standard binary classification template: `"Is there a [Object] in the image? Please answer yes or no."`, where `[Object]` is dynamically replaced per test item with the specific target object queried.

Text-Quality Metrics. To evaluate caption-level quality and detect potential repetitive degeneration, we report the **repetition rate (Rep)** and **Moving-Average Type-Token Ratio (MATTR)**, denoted as TTR in the result tables for compactness. Repetition is quantified via seq-rep-6, defined as $1 - |G_6^{\text{unique}}|/|G_6|$, where G_6 represents the multiset of all 6-grams in the generated caption. Lexical diversity is measured using MATTR with a fixed 50-token window slid one token at a time, which mitigates the length-induced dilution effect inherent in standard TTR.

L Case Studies: Repetition Degeneration and $CHAIR_I$ Deflation

We present two representative cases (Figure 4 and Figure 5) showing how strong attention-based interventions (PAI and AttnReal) push the base MLLM into severe repetitive degeneration. The caption contains very few hallucinated objects, but is paired with a multiplied repetition of GT-grounded objects. Because $CHAIR_I$ is counted at the token-instance level without deduplication, the inflated denominator from repetition combined with a sparse hallucination numerator produces a mechanical imbalance: $CHAIR_I$ reads artificially low and masks the underlying generation failure. On the same image, the CAS output (rightmost panel of each figure) maintains vanilla-level fluency without introducing repetition.

	LLaVA-1.5						Shikra					
	Random		Popular		Adversarial		Random		Popular		Adversarial	
Method	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
Greedy	86.57	84.80	85.40	83.69	**83.53**	81.98	80.73	82.65	77.97	80.85	72.93	77.18
DoLa	82.10	78.32	75.90	73.74	75.33	73.29	65.43	71.57	62.97	70.35	61.50	68.91
VCD	86.47	85.11	84.93	83.69	81.70	80.86	76.23	80.00	74.20	78.63	71.97	76.75
OPERA	86.77	85.04	85.47	83.80	**83.53**	82.04	79.73	81.36	78.27	80.30	72.93	76.60
PAI	87.07	85.58	85.70	84.29	82.73	81.63	79.70	76.39	78.73	75.57	**76.67**	73.81
CODE	86.43	84.59	85.30	83.51	83.40	81.77	81.87	83.37	**78.83**	81.00	74.80	78.10
DeCo	86.57	84.80	85.40	83.69	**83.53**	81.98	73.27	77.88	72.57	77.49	65.47	73.17
SSL	–	–	–	–	–	–	–	–	–	–	–	–
CAS (ours)	**89.13**	**88.40**	**87.37**	**86.76**	83.23	**83.16**	**82.43**	**84.11**	78.67	**81.30**	74.60	**78.19**

	Qwen-VL						InstructBLIP					
	Random		Popular		Adversarial		Random		Popular		Adversarial	
Method	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
Greedy	90.03	89.27	88.10	87.45	**84.87**	84.57	78.23	80.75	78.00	80.58	73.03	77.19
DoLa	88.27	89.05	82.23	84.30	75.00	79.24	51.57	58.43	58.23	61.97	58.67	62.22
VCD	90.37	89.91	87.90	87.64	83.73	84.06	77.50	79.81	75.90	78.67	71.83	75.95
OPERA	87.43	89.30	87.23	87.48	84.10	**84.60**	73.53	75.72	68.03	72.08	66.13	70.90
PAI	–	–	–	–	–	–	–	–	–	–	–	–
CODE	90.27	89.76	88.03	87.70	83.63	83.91	77.77	80.19	79.37	81.35	74.03	77.61
DeCo	90.07	89.33	88.17	87.55	84.70	84.47	79.97	81.91	77.40	80.06	73.03	77.09
SSL	–	–	–	–	–	–	82.47	83.73	79.40	81.41	74.43	77.92
CAS (ours)	**90.63**	**90.06**	**88.40**	**87.98**	84.50	84.56	**87.57**	**87.19**	**85.30**	**85.20**	**81.20**	**81.82**

Table 16: POPE per-subset results. Best per column within each model in **bold**, second best underlined. "–" indicates the method is not available for that base model.



Input Image

LLaVA-1.5 :
The image features a pair of red **scissors** placed on a white surface. The **scissors** are positioned close to each other, with one of them being slightly closer to the edge of the surface. The **scissors** are of a similar size and shape, with the handle of the **scissors** being visible. The **scissors** are likely used for cutting or crafting purposes.

LLaVA-1.5 + PAI:
The image features a pair of red **scissors** with a white background. The **scissors** are red and have a black handle. The **scissors** are placed on a white surface, which is the background of the image. The **scissors** are placed on a table, and the handle is visible.The **scissors** are placed on a table, and the handle is visible.The **scissors** are placed on a table, and the handle is visible.The **scissors** are placed on a table, and the handle is visible.The **scissors** are placed on a table, and the handle is visible.The **scissors** are placed on a table, and the handle is visible.The **scissors** are placed on a table, and the handle is visible. ● ● ●
(loop x29) ⟶ *CHIR$_I$ deflation*

LLaVA-1.5 + CAS :
The image features a pair of red **scissors** with a white background. The **scissors** are open, and they are placed on a table. The **scissors** are positioned in the center of the image, making them the main focus of the scene.

Figure 4: **Repetitive degeneration of PAI on LLaVA-1.5.** A single hallucinated attribute ("black handle") is repeated about 29 times, artificially inflating the mention count of GT-grounded objects; the token-level CHAIR$_I$ is therefore diluted by an inflated denominator and reads artificially low even though the caption has already degenerated. CAS produces a clean, repetition-free caption on the same image.



Input Image

LLaVA-1.5 :

The image features a **street sign** pole with multiple street signs attached to it. The pole is situated in front of a building with a green roof, likely a market or store. The signs are visible from different angles, indicating the presence of a public market. There are several **street sign**s on the pole, including one at the top, indicating the intersection of Pike and Pine Streets. The signs are positioned in a way that they can be easily read by people passing by.

LLaVA-1.5 + AttnReal

The image features a **street sign** with a **street sign** on top of it, indicating the intersection of Pine Street and Pike Street.The **street sign** is located on a pole, and the street sign below it is green.The **street sign** is positioned on a pole, and the **street sign** below it is green.The **street sign** is positioned on a pole, and the **street sign** below it is green.The **street sign** is positioned on a pole, and the **street sign** below it is green.The street sign is positioned on a pole, and the **street sign** below it is green.The street sign is positioned on a pole, and the **street sign** below it is green. The **street sign** is positioned on a pole, and the **street sign** below it is green.The **street sign** is positioned on a pole, and the **street sign** below it is green. ● ● ● **(loop x24)** ⟶ ***CHIR$_I$ deflation***

LLaVA-1.5 + CAS :

The image features a **street sign** with two **street sign**s attached to it. The **street sign**s are for Pine Street and Public Market. The **street sign**s are located near a building, possibly a market or a store. The scene is set in a cloudy sky, giving it a somewhat gloomy atmosphere.

Figure 5: **Repetitive degeneration of AttnReal.** The same CHAIR$_I$ deflation pattern as Figure 4: repetition inflates the denominator and depresses the metric reading, while the caption itself has failed. CAS produces a normal generation on the same image.