ATTENDING WITHOUT SELECTING: VISUAL KEY COLLAPSE AND HALLUCINATION IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models often describe objects that are absent from the image, a failure usually attributed to insufficient attention to visual input. We find that hallucinated and faithful captions place the same share of attention on the image, but in hallucinated captions that attention depends less on the word being generated. The failure lies in selecting where to look rather than in how much to look. We trace it to a collapse of visual keys onto a few dominant directions, which leads every query to rank image regions by nearly the same criteria. The collapse appears across architectures with different visual interfaces, does not depend on the relative norms of visual and text keys, and is more severe for images that the model describes less faithfully. Building on this diagnosis, we introduce Whitened Attention Steering, a training-free method that suppresses the dominant directions in visual attention scores while keeping the total attention on the image fixed. Hallucination decreases steadily as the correction is strengthened and increases when it is reversed, whereas matched corrections of other directions or other heads have little effect. On LLaVA-1.5, the share of captions containing a hallucinated object falls from 54% to 30%, and across models the gain follows the severity of collapse. These findings indicate that hallucination in current vision-language models stems less from a shortage of visual attention than from a loss of selection within it.

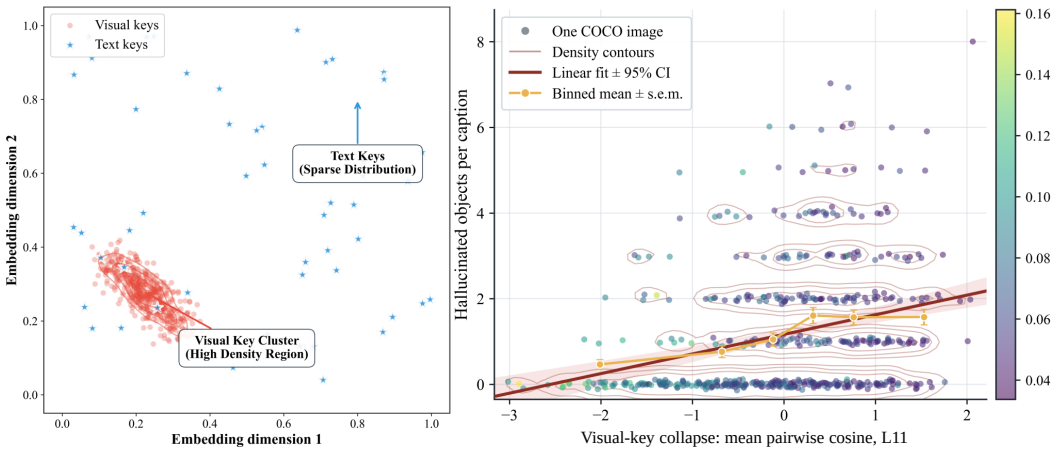
1 INTRODUCTION

Large vision-language models (LVLMs) couple a pretrained visual encoder with a large language model and perform strongly on captioning and visual question answering (Liu et al., 2023c;b; Dai et al., 2023; Bai et al., 2025), yet they often describe objects that are not in the image (Rohrbach et al., 2018; Li et al., 2023; Gunjal et al., 2024). This object hallucination is commonly attributed to an imbalance between modalities. The language prior is strong, the visual evidence is weak by comparison, and when the image receives too little attention the prior fills the gap (Leng et al., 2024; Zhou et al., 2023; Zhai et al., 2023). Most remedies act on this account, contrasting the output with a reference that leans more on the language prior (Leng et al., 2024; Favero et al., 2024; Chuang et al., 2024), raising the attention placed on image tokens (Liu et al., 2024; Yin et al., 2025), or strengthening grounding with curated data (Liu et al., 2023a; Yu et al., 2024b). These approaches rest on a common premise: a model hallucinates because it attends too little to the image.

This premise can be tested directly. If it holds, a model should attend less to the image while producing a hallucinated caption than while producing a faithful one. On 500 MSCOCO images (Lin et al., 2014), LLaVA-1.5 (Liu et al., 2023b) places virtually the same share of attention on the image in both cases, and this share predicts hallucination hardly better than chance (Section 3.2). What differs is how that attention moves. In hallucinated captions, the attended regions change less from one generated word to the next. The model does not look at the image less; it looks at it less selectively. The failure lies in *selection*, which regions receive attention, rather than in *allocation*, how much attention the image receives.

Selection depends on the keys of the image tokens. Attention scores each region by the inner product between its key and the query of the current token. Whatever the visual keys have in common raises or

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) (b)

Figure 1: **Visual key collapse in LLaVA-1.5-7B.** (a) t-SNE of the visual keys (circles) and text keys (stars) of one image at layer 11, head 23. (b) Hallucinated objects per caption against visual key collapse at layer 11 (standardized mean pairwise cosine) for 500 MSCOCO images. Color shows relative local density.

lowers all regions alike, so only the ways in which they differ decide where attention goes. Different queries can therefore favor different regions only if the visual keys differ along many directions. When they differ along only a few, every query ranks the regions largely by their coordinates along those directions, and the word being generated can do little more than reweight them. Remedies that give the image more attention leave this limit untouched.

In the models we examine, the visual keys vary along very few directions. Within a single attention head of LLaVA-1.5, the visual keys of an image crowd together while the text keys spread widely (Figure 1a). Averaged over heads and images, the visual keys in the intermediate layers have an effective rank of only six to nine out of 128 dimensions, compared with 28 to 42 for the text keys (Section 3.2). This concentration is directional rather than a matter of magnitude. It appears whether visual keys are larger in norm than text keys, as in LLaVA-1.5, or smaller, as in InstructBLIP (Dai et al., 2023) and Qwen3-VL (Bai et al., 2025). Anisotropy is well documented in language models (Ethayarajh, 2019; Gao et al., 2019), but here it affects one modality within the same attention computation. We call this phenomenon *visual key collapse*.

Consistent with this mechanism, the severity of collapse varies across images and is associated with hallucination. Captions contain more hallucinated objects on average when the visual keys of the image are more collapsed (Figure 1b), an association that persists after controlling for scene complexity and caption length (Section 3.3). If collapse is what limits selection, counteracting it should reduce hallucination without giving the image any more attention.

We introduce *Whitened Attention Steering* (WAS), a training-free method built on this idea. WAS suppresses the dominant directions of the visual keys when image regions are scored, then restores the mean and scale of the original scores so that the image keeps the same share of attention. It applies the correction only in heads whose attention over the image has become nearly uniform, and it changes neither the keys nor the model weights. Because the total attention on the image is held fixed, any reduction in hallucination under WAS can be attributed to where attention is directed rather than to how much of it the image receives.

On LLaVA-1.5, WAS leaves the total attention on the image unchanged, makes that attention depend more on the generated word, and lowers the share of captions containing a hallucinated object from 54% to 30% (Appendix P). The reduction grows steadily with the strength of the correction, and reversing the correction, which deepens the collapse, increases hallucination beyond that of the unmodified model. Applying the correction to other directions or other heads, or merely sharpening the scores, achieves at most two fifths of the reduction. Precision rises while recall barely changes, so the model names fewer absent objects rather than fewer objects, and counting of repeated objects shows no measurable decline. Across three LVLMs and three hallucination benchmarks, WAS compares favorably with training-free alternatives, and its gain follows the severity of collapse. These

results indicate that hallucination in current LVLMs stems less from a shortage of attention to the image than from a loss of selection within it.

Our contributions are threefold.

- We show that faithful and hallucinated captions place the same attention on the image but differ in how that attention depends on the generated word, locating the failure in selection rather than allocation.
- We identify visual key collapse, explain how it limits selection, and show that it occurs across three architectures regardless of key norms and is associated with hallucination beyond scene complexity and caption length.
- We introduce WAS, which suppresses the dominant key directions under a fixed attention budget and reduces hallucination in a graded, direction-specific way, with gains that follow the severity of collapse.

2 Related Work

Hallucination mitigation in LVLMs. Approaches to reducing object hallucination broadly divide into three families. *Data-centric* methods construct negative or counterfactual training data and refine generation through contrastive tuning or preference optimisation Liu et al. (2023a); Yu et al. (2024a); Sarkar et al. (2024); Jiang et al. (2024); Chen et al. (2025); Shang et al. (2024); Lyu et al. (2024); Yang et al. (2025b); Yoon et al. (2025); Liu et al. (2025); these can be effective but require curated data and additional training. *Contrastive decoding* methods operate at inference time by contrasting output distributions under standard and distorted inputs Leng et al. (2024); Wang et al. (2024b); Favero et al. (2024); Chuang et al. (2024); Huo et al. (2024); Chen et al. (2024); Wang et al. (2024a); Li et al. (2025); Suo et al. (2025); Zhang et al. (2025). VCD Leng et al. (2024) removes visual tokens to obtain a text-only reference; DoLa Chuang et al. (2024) contrasts across layers within the language backbone. While principled, these methods typically double inference cost. *Attention intervention* methods modify the attention mechanism directly Huang et al. (2024); An et al. (2025); Huang et al. (2025); Wan et al. (2025); Liu et al. (2024); Woo et al. (2025); Tang et al. (2025a); Jiang et al. (2025); Tang et al. (2025b); Yang et al. (2025a); Yin et al. (2025); Qian et al. (2025); Zhuang et al. (2025); Dong et al. (2025); Duan et al. (2025); Che et al. (2025). OPERA Huang et al. (2024) penalises over-trusted tokens via beam-search retrospection; SHIELD Huang et al. (2025) defends against encoder bias through a secondary scoring pass. These methods share a common premise that visual tokens receive insufficient attention due to modality bias in logit magnitudes. Our work tests this premise by showing that the failure lies in the directional geometry of visual key representations, and proposes a spectral correction rather than a scale adjustment.

Table 1: **Key geometry of three LVLMs**, averaged over 500 MSCOCO images, heads and layers. Visual keys are less dispersed and denser than text keys in every model, whether larger or smaller in norm.

Model		Mean norm $\bar{r}$	Ang. disp. D_{ang}	Local density ρ_{local}
LLaVA-1.5-7B	Visual	104.0	0.195	0.957
	Text	95.9	0.375	0.772
	Ratio (Vis. / Text)	1.08×	**0.52**×	**1.24**×
InstructBLIP-7B	Visual	59.7	0.317	0.823
	Text	81.6	0.546	0.454
	Ratio (Vis. / Text)	0.73×	**0.58**×	**1.81**×
Qwen3-VL-4B	Visual	7.8	0.325	0.861
	Text	9.9	0.438	0.649
	Ratio (Vis. / Text)	0.79×	**0.74**×	**1.33**×

3 RETHINKING HALLUCINATION: FROM MODALITY BIAS TO KEY COLLAPSE

3.1 SETUP

We analyze LLaVA-1.5-7B (Liu et al., 2023b), which feeds $N_v = 576$ visual tokens from a CLIP ViT-L/14 encoder into Vicuna-7B, and apply the same protocol to InstructBLIP-7B (Dai et al., 2023) and Qwen3-VL-4B (Bai et al., 2025). On 500 MSCOCO images (Lin et al., 2014) captioned with greedy decoding, we extract the keys after RoPE (Su et al., 2024) at every layer and head, and describe the visual and text keys by their mean ℓ_2 norm, their angular dispersion $D_{\text{ang}} = 1 - \frac{1}{N(N-1)} \sum_{i \neq j} \cos(\mathbf{k}_i, \mathbf{k}_j)$, and their local density ρ_{local}, the mean cosine similarity of each key to its ten nearest neighbours.

3.2 VISUAL KEYS ARE NOT WEAK BUT COLLAPSED

The modality-bias account predicts that the image is under-attended when the model hallucinates. It is not. Hallucinated and faithful captions place the same share of attention on the image (Appendix C), and visual keys match or exceed text keys in norm through most layers of LLaVA-1.5 (Figure 4).

What sets visual keys apart is their direction. In all three models, visual keys are markedly less dispersed and denser than text keys, whether they are larger in norm, as in LLaVA-1.5, or smaller, as in InstructBLIP and Qwen3-VL (Table 1). Their covariance has an effective rank of only six to nine out of 128 dimensions in the intermediate layers of LLaVA-1.5, against 28 to 42 for text keys (Appendix B). Anisotropy is well documented in language models (Ethayarajh, 2019; Gao et al., 2019), but here it affects one modality within the same attention computation. We call this phenomenon *visual key collapse*.

3.3 FROM COLLAPSE TO HALLUCINATION

Collapse matters because it limits which regions a query can select. Writing $\boldsymbol{\mu}$ for the mean visual key, the score of visual token j splits as

$$\frac{\mathbf{q}^\top \mathbf{k}_j}{\sqrt{d}} = \frac{\mathbf{q}^\top \boldsymbol{\mu}}{\sqrt{d}} + \frac{\mathbf{q}^\top (\mathbf{k}_j - \boldsymbol{\mu})}{\sqrt{d}}. \tag{1}$$

The first term is shared by all visual tokens and only sets how much attention the image receives. Which regions receive it is decided by the second term. When the deviations $\mathbf{k}_j - \boldsymbol{\mu}$ lie along a few directions, every query ranks the regions largely by their coordinates along those directions, and different queries produce similar attention over the image.

The mutual information between the query and the attended region makes this precise. Let J be the visual token selected by attention, drawn from the attention distribution over the image. Then

$$I(\mathbf{q}; J) = H(J) - H(J \mid \mathbf{q}) \tag{2}$$

measures how much the query determines where attention goes. When different queries yield similar distributions, $H(J \mid \mathbf{q})$ approaches $H(J)$ and $I(\mathbf{q}; J)$ is small, and it decreases as the collapse deepens (Appendix F). Over the decoding steps of a caption, this mutual information equals the generalized Jensen-Shannon divergence among their attention distributions, which our query-dependence statistic estimates pairwise.

Hallucinated captions show exactly this. Their attention over the image depends less on the word being generated than in faithful captions (Appendix C), and captions contain more hallucinated objects when the visual keys of the image are more collapsed (Figure 1b), an association that holds after controlling for scene complexity and caption length (Appendix D).

The directions that let a query single out particular regions are thus present in the visual keys but carry little weight in the scores. Suppressing the dominant directions, while keeping the shared term and the scale of the scores fixed, should restore selection without changing how much attention the image receives. Section 4 builds this correction.

4 WHITENED ATTENTION STEERING

A correction that follows from Section 3 must satisfy three requirements. (R1) It should restore the weight of the non-dominant directions in the deviation scores, so that different queries can select different regions. (R2) It should keep the shared term and the scale of the scores, so that the image receives as much attention as before. (R3) It should act only where selection has failed. WAS meets them with three steps inside each attention head, applied to the visual scores $\mathbf{s} = \mathbf{K}_v\mathbf{q}/\sqrt{d} \in \mathbb{R}^{N_v}$, where the rows of $\mathbf{K}_v$ are the N_v visual keys and $\mathbf{q}$ is the query of the current token. Text and system scores are left unchanged. Figure 2 illustrates the effect on a single question, and Algorithm 1 summarizes the procedure.

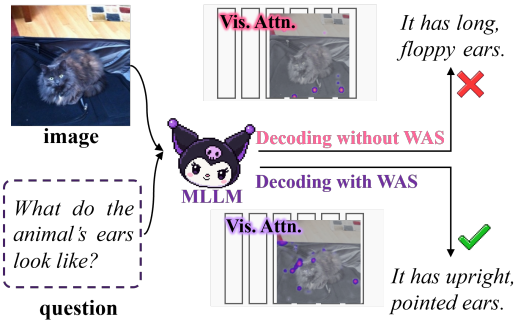


Figure 2: **Overview of WAS.** Without WAS, attention over the image does not follow the question and the answer is wrong. WAS suppresses the dominant key directions in the visual scores, and attention follows the question again.

4.1 Suppressing the Dominant Directions (R1)

We estimate the dominant directions from the visual keys of the current image. After centering the keys at their mean $\boldsymbol{\mu}$,

$$\bar{\mathbf{K}}_v = \mathbf{K}_v - \mathbf{1}_{N_v}\boldsymbol{\mu}^\top, \tag{3}$$

we take the rank-r truncated singular value decomposition

$$\bar{\mathbf{K}}_v \approx \mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{V}_r^\top, \tag{4}$$

whose right singular vectors $\mathbf{V}_r \in \mathbb{R}^{d\times r}$ span the dominant directions and whose singular values $\boldsymbol{\Sigma}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ measure their strength. Each score is then corrected by rescaling its component along direction i by $1/\sigma_i$,

$$\tilde{s}_j = s_j + \frac{1}{\sqrt{d}}\,\mathbf{q}^\top\mathbf{V}_r\big(\boldsymbol{\Sigma}_r^{-1} - \mathbf{I}_r\big)\mathbf{V}_r^\top\mathbf{k}_j. \tag{5}$$

Since $\sigma_i \gg 1$ for the dominant directions, their contribution is almost removed, and the remaining directions, which keep their original weight, now determine the ranking of the regions. The correction amounts to adjusting the query within the dominant subspace, so the keys themselves are never modified. It is applied to the uncentered keys for efficiency, and the resulting offset, shared by all visual tokens, is removed by the next step.

We suppress only the top r directions instead of whitening all d. Full whitening would also rescale the many weak directions, whose variances are poorly estimated from the keys of a single image, and would amplify their noise, which the ablation in which Table 14 confirms. The sharp spectral gap after the first few directions (Appendix B) makes the choice of r uncritical (Appendix H.1). The decomposition is computed once per image during prefill by randomized power iteration and cached for all decoding steps(Appendix G).

4.2 Preserving the Attention Budget (R2)

Suppression changes the mean and spread of the visual scores, and because the softmax normalizes visual and text scores jointly, it would change how much attention the image receives. We therefore map the corrected scores back onto the statistics of the original ones,

$$s_j^\star = \mu_o + \alpha\,\frac{\sigma_o}{\sigma_w}\big(\tilde{s}_j - \mu_w\big), \tag{6}$$

where μ_o, σ_o and μ_w, σ_w are the mean and standard deviation of $\mathbf{s}$ and $\tilde{\mathbf{s}}$ across visual tokens. The mean of the original scores is exactly the shared term $\mathbf{q}^\top\boldsymbol{\mu}/\sqrt{d}$ of Eq. equation 1, so score matching restores the shared term and the spread and takes only the ranking from the corrected scores. It is also invariant to any constant added to $\tilde{\mathbf{s}}$, which absorbs the offset of Eq. equation 5. At $\alpha = 1$ the attention on the image is approximately preserved (Appendix F.3), and a slightly larger α sharpens the corrected ranking with little effect on this budget.

4.3 Entropy-Gated Blending (R3)

Not every head has lost selection. We identify those that have by the normalized entropy of their attention over the image,

$$\bar{H}_{\text{vis}} = -\frac{1}{\log N_v} \sum_{j=1}^{N_v} a_j \log a_j, \qquad a_j = \frac{\exp(s_j)}{\sum_{j'} \exp(s_{j'})}, \tag{7}$$

which approaches 1 when attention is spread uniformly over the image and does not depend on the number of visual tokens. The corrected and original scores are blended by a soft gate,

$$s_j^{\text{final}} = g\, s_j^{\star} + (1-g)\, s_j, \qquad g = \sigma\left(\frac{\bar{H}_{\text{vis}} - \tau}{\beta}\right), \tag{8}$$

where $\sigma(\cdot)$ is the logistic function, τ a threshold and β a temperature. Across heads, the entropy of visual attention is bimodal, with a small mode close to uniform (Figure 3a). With τ in the valley between the two modes and a small β, the gate fires on this sparse set of heads and leaves the others unchanged (Figure 3b).

5 Experiments

We evaluate WAS along four axes: object hallucination on free-form captioning (CHAIR), probing-based hallucination detection (POPE), multi-dimensional hallucination assessment (AMBER), and general multimodal perception and cognition (MME). Experiments span three model families with distinct visual interfaces to test the generality of the geometric diagnosis and correction.

5.1 Setup

Models. We apply WAS to three LVLMs covering distinct visual encoding pipelines: **LLaVA-1.5-7B** (Liu et al., 2023b), which pairs a CLIP ViT-L/14 encoder with a Vicuna-7B backbone via linear projection (576 visual tokens); **InstructBLIP-7B** (Dai et al., 2023), which routes visual features through a Q-Former into Vicuna-7B (32 query tokens); and **Qwen3-VL-4B**, which employs dynamic-resolution visual encoding with a natively multimodal backbone. All experiments use NVIDIA A6000 GPUs.

Benchmarks. **CHAIR** (Rohrbach et al., 2018) evaluates object hallucination in free-form captions on 500 MSCOCO images, reporting CHAIR_S (sentence-level), CHAIR_I (instance-level), and Recall. **POPE** (Li et al., 2023) probes existence hallucination through binary yes/no questions under Random, Popular, and Adversarial splits. **AMBER** (Wang et al., 2023) provides multi-dimensional evaluation spanning generative and discriminative tasks across existence, attribute, and relation hallucinations. **MME** (Fu et al., 2023) measures perception and cognition abilities across 14 subtasks. Formal metric definitions are given in Appendix K.

Algorithm 1 Whitened Attention Steering (WAS)

Require: Query $\mathbf{q} \in \mathbb{R}^d$, visual keys $\mathbf{K}_v \in \mathbb{R}^{N_v \times d}$, original visual logits $\mathbf{s} \in \mathbb{R}^{N_v}$
Sharpening factor α, entropy threshold τ, gate temperature β
Ensure: Corrected visual logits $\mathbf{s}^{\text{final}} \in \mathbb{R}^{N_v}$

// Step 1: Truncated spectral whitening (cached after prefill)
$\boldsymbol{\mu} \leftarrow \frac{1}{N_v}\sum_{j=1}^{N_v} \mathbf{k}_j$ ▷ Eq. 3
$\bar{\mathbf{K}}_v \leftarrow \mathbf{K}_v - \mathbf{1}_{N_v}\boldsymbol{\mu}^\top$
$\mathbf{V}_r, \boldsymbol{\Sigma}_r \leftarrow \text{TruncSVD}(\bar{\mathbf{K}}_v, r=5)$ ▷ Eq. 4
$\tilde{\mathbf{q}} \leftarrow \mathbf{V}_r\left(\boldsymbol{\Sigma}_r^{-1} - \mathbf{I}_r\right)\mathbf{V}_r^\top \mathbf{q}$ ▷ corrected query, $\in \mathbb{R}^d$
$\tilde{\mathbf{s}} \leftarrow \mathbf{s} + \frac{1}{\sqrt{d}}\mathbf{K}_v\tilde{\mathbf{q}}$ ▷ Eq. 5

// Step 2: Score matching
$\mu_o, \sigma_o \leftarrow \text{mean}(\mathbf{s}),\ \text{std}(\mathbf{s})$
$\mu_w, \sigma_w \leftarrow \text{mean}(\tilde{\mathbf{s}}),\ \text{std}(\tilde{\mathbf{s}})$
$\mathbf{s}^\star \leftarrow \mu_o + \alpha\,\frac{\sigma_o}{\sigma_w}(\tilde{\mathbf{s}} - \mu_w)$ ▷ Eq. 6

// Step 3: Entropy-gated blending
$\mathbf{a} \leftarrow \text{softmax}(\mathbf{s})$
$H_{\text{vis}} \leftarrow -\sum_j a_j \log a_j$ ▷ Eq. 7
$\bar{H}_{\text{vis}} \leftarrow H_{\text{vis}} / \log N_v$
$g \leftarrow \sigma\big((\bar{H}_{\text{vis}} - \tau)\,/\,\beta\big)$ ▷ Eq. 8
return $g \cdot \mathbf{s}^\star + (1-g)\cdot \mathbf{s}$

Baselines. We compare against ten representative training-free methods spanning three intervention strategies: contrastive decoding (ICD Wang et al. (2024b), VCD Leng et al. (2024), DoLa Chuang et al. (2024), DeCo Yue et al. (2024)), attention intervention (OPERA Huang et al. (2024), AGLA An et al. (2025), ONLY Wan et al. (2025),

Table 2: **CHAIR evaluation on MSCOCO**. Best in **bold**, second best underlined.

	LLaVA-1.5-7B				InstructBLIP-7B			
Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall↑	CFAR↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall↑	CFAR↑
Greedy	54.2	18.5	72.5	51.9	54.0	18.1	71.1	51.6
+ ICD	56.2	16.3	76.8	53.0	56.5	19.5	71.0	49.8
+ DoLa	48.4	15.9	69.3	54.6	60.0	20.1	71.5	47.7
+ VCD	51.8	16.2	76.8	55.6	57.0	17.0	72.1	50.6
+ VAF	52.6	14.2	76.0	55.5	54.5	16.5	71.5	52.0
+ AGLA	50.0	15.3	76.5	56.8	49.0	12.1	72.5	56.7
+ ONLY	49.8	14.3	78.1	57.9	52.2	15.5	71.0	53.3
+ OPERA	48.5	14.9	78.1	58.5	54.0	12.8	69.8	52.6
+ MEMVR	48.0	14.5	77.1	58.5	51.0	15.0	71.5	54.4
+ DeCo	41.6	13.4	69.7	59.0	44.4	15.2	66.0	55.4
+ SHIELD	36.6	10.3	68.2	61.7	40.4	**10.9**	69.2	60.1
+ WAS	**30.2**	**10.2**	70.7	**66.1**	**36.0**	11.3	71.2	**63.2**

Table 3: **CHAIR on Qwen3-VL-4B.**

Method	$CH_S \downarrow$	$CH_I \downarrow$	Rec↑	CFAR↑
Greedy	26.0	9.4	49.2	54.9
+ VCD	25.8	9.4	49.7	55.3
+ VAF	24.8	8.8	49.5	55.6
+ WAS	**17.8**	**7.2**	**50.2**	**59.1**

Table 4: **Component ablation** (LLaVA-1.5-7B).

Configuration	$CH_S \downarrow$	$CH_I \downarrow$	Rec↑	CFAR↑
Greedy	54.2	18.5	72.5	51.9
+ Spec. whitening	41.2	14.5	67.0	55.7
+ Score matching	32.8	11.2	71.4	64.9
+ Ent. gate (WAS)	**30.2**	**10.2**	**70.7**	**66.1**

SHIELD Huang et al. (2025)), and representation correction (MEMVR Zou et al. (2024), VAF Yin et al. (2025)). All methods use identical inference configurations (greedy decoding, max 512 tokens) with each author's recommended hyperparameters.

WAS configuration. We use a fixed configuration across all models and benchmarks: sharpening factor $\alpha = 1.2$, truncation rank $r = 5$, entropy threshold $\tau = 0.9$, gate temperature $\beta = 0.02$ and intervention window L9–20. Sensitivity analysis in Appendix H.1 confirms that performance varies by less than 2 $CHAIR_S$ points across the tested ranges of each parameter.

5.2 Main Results

We take $CHAIR_S$ as the primary measure of caption hallucination and additionally report CFAR (Appendix K), which also accounts for recall. The improvements over greedy decoding hold under paired bootstrap intervals (Appendix I.1).

CHAIR (Tables 2 and 3). WAS achieves the lowest $CHAIR_S$ and the highest CFAR on all three models while keeping recall close to that of greedy decoding. The size of the gain follows the severity of collapse: $CHAIR_S$ drops by 24.0, 18.0 and 8.2 points on LLaVA-1.5, InstructBLIP and Qwen3-VL-4B, the same order as their collapse in Table 1, and the gain becomes small on the less collapsed Qwen3-VL-8B (Appendix I.3). Since LLaVA-1.5 and InstructBLIP hallucinate equally often without correction, this ordering does not reflect differing room for improvement.

POPE (Table 6). WAS achieves the best F1 on the Popular and Adversarial splits for all three models, and its margin over the strongest baseline is largest on the Adversarial split, where co-occurring objects make hallucination most likely.

AMBER (Table 5). WAS lowers the generative hallucination rate and raises object coverage on all three models, and improves discriminative F1 as well, so the gain holds for both open-ended description and yes/no questions.

MME (Tables 8 and 9). WAS obtains the highest total score on LLaVA-1.5 and Qwen3-VL-4B, so reducing hallucination does not come at the cost of general perception and reasoning.

Table 5: **AMBER evaluation** across three LVLM architectures. Best in **bold**, second best underlined.

	Generative				Discriminative			
Method	CHAIR↓	Cover↑	Hal↓	Cog↓	Acc↑	Prec↑	Rec↑	F1↑
LLaVA-1.5-7B								
Greedy	11.2	51.6	52.0	7.0	68.7	77.9	71.5	74.6
+ VCD	9.8	52.2	47.0	6.8	67.2	76.8	69.9	73.2
+ VAF	11.5	47.8	51.0	5.3	66.5	78.5	65.6	71.5
+ WAS	**5.7**	**59.0**	**24.2**	**2.5**	**83.5**	**90.4**	**84.3**	**87.2**
InstructBLIP-7B								
Greedy	6.4	53.9	31.7	5.0	75.3	92.3	68.5	78.6
+ VCD	5.4	57.0	30.8	2.8	77.5	91.2	72.8	81.0
+ VAF	5.6	56.5	31.5	2.6	77.2	91.5	72.2	80.8
+ WAS	**3.8**	**60.5**	**22.0**	**1.8**	**81.0**	**93.0**	**77.5**	**84.5**
Qwen3-VL-4B								
Greedy	5.9	52.5	31.0	1.9	89.8	91.8	93.0	92.4
+ VCD	5.4	51.0	28.5	1.7	90.2	92.5	93.2	92.8
+ VAF	5.2	51.2	28.2	**1.6**	90.5	92.8	93.5	93.1
+ WAS	**3.2**	**58.5**	**19.0**	**1.2**	**93.5**	**95.0**	**94.8**	**94.9**

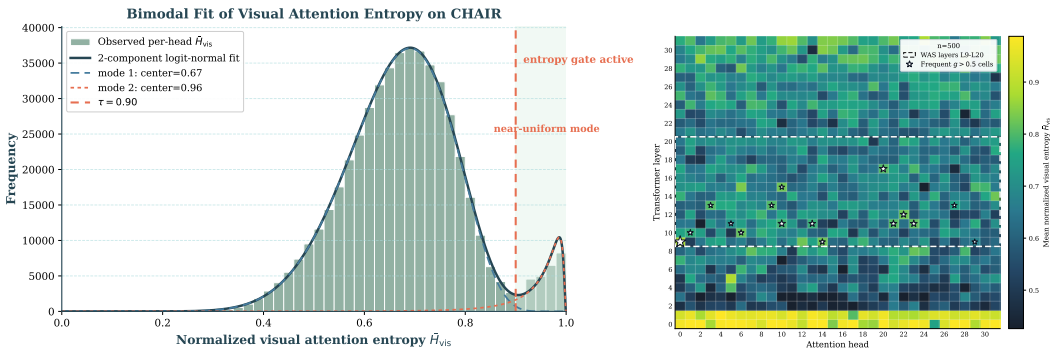
Table 6: **POPE evaluation.** Best in **bold**, second best underlined.

	Random				Popular				Adversarial			
Method	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
LLaVA-1.5-7B												
Greedy	83.49	88.84	76.76	82.28	79.98	82.47	76.76	79.34	76.03	76.11	76.80	76.26
+ ICD	84.87	89.96	78.53	83.27	82.93	86.28	78.46	81.45	81.07	82.64	78.60	79.96
+ DoLa	84.78	87.59	81.27	84.19	79.75	84.11	76.22	80.61	76.32	77.27	75.47	76.16
+ VCD	86.84	87.15	**86.68**	86.83	82.65	87.15	80.59	83.37	77.31	73.43	**86.47**	79.28
+ DeCo	85.80	89.20	82.00	85.42	82.40	85.50	79.50	82.38	78.80	79.60	78.80	79.18
+ OPERA	87.53	94.52	79.80	86.45	84.21	88.00	79.80	83.50	80.88	82.16	79.76	80.69
+ AGLA	**88.54**	94.41	82.08	**87.71**	85.14	87.88	**82.08**	84.68	81.13	81.20	82.10	81.36
+ WAS	88.27	**95.77**	80.07	87.22	**86.30**	**91.47**	80.07	**85.39**	**83.20**	**85.37**	80.13	**82.67**
InstructBLIP-7B												
Greedy	80.42	78.93	83.21	80.94	76.09	73.22	82.94	77.65	72.37	68.78	83.06	75.42
+ ICD	82.50	81.80	84.20	82.80	78.00	75.90	83.60	79.10	73.80	70.20	83.80	76.00
+ DoLa	83.00	83.06	83.13	83.00	78.99	77.12	83.13	79.85	74.67	71.53	83.11	76.68
+ VCD	84.11	84.20	85.36	84.13	79.94	77.84	83.33	80.80	76.32	73.24	84.08	78.08
+ DeCo	84.60	85.40	83.80	84.50	79.90	77.60	84.40	80.80	76.20	72.90	84.30	78.20
+ OPERA	85.07	88.39	80.73	84.39	78.33	73.85	**87.73**	80.20	75.50	70.49	**87.73**	78.17
+ AGLA	87.30	88.83	**85.68**	87.07	81.86	80.17	85.68	82.58	77.29	74.09	85.67	79.16
+ WAS	**89.97**	**94.91**	84.47	**89.38**	**85.73**	**86.66**	84.47	**85.55**	**82.50**	**81.27**	84.47	**82.84**
Qwen3-VL-4B												
Greedy	90.20	96.80	83.20	89.50	87.30	93.20	81.00	86.60	85.80	90.00	81.00	85.20
+ VCD	90.50	96.50	83.80	89.70	87.60	93.50	81.50	87.10	86.00	90.20	81.50	85.60
+ VAF	91.00	97.00	84.80	90.40	88.20	**94.20**	82.20	87.80	86.80	**91.00**	82.50	86.50
+ WAS	**92.30**	**98.10**	**86.27**	**91.81**	**89.73**	92.75	**86.20**	**89.36**	**88.53**	90.76	**85.80**	**88.21**

5.3 Ablation Studies

All ablations use LLaVA-1.5-7B on CHAIR (500 MSCOCO images, greedy decoding). Additional analyses are provided in the appendix, including ablations and layer selection in Appendix H, attention maps in Appendix J, and caption case studies in Appendix O.

Component analysis. Table 4 isolates each component's contribution by incrementally adding them to greedy decoding. Spectral whitening alone reduces CHAIR_S by 13.0 points but degrades Recall by 5.5, reflecting the cross-modal attention imbalance when raw whitened logits replace the original scale. Adding score matching recovers Recall to near-baseline levels while preserving the hallucination reduction, confirming that it successfully decouples the corrected ranking from the logit scale. Entropy-gated blending provides an additional 2.6-point CHAIR_S reduction by confining the correction to genuinely collapsed heads. Full whitening ($r = d$) serves as a negative control, confirming that truncation is essential (Table 14).



(a) (b)

Figure 3: **Where WAS intervenes** (LLaVA-1.5-7B, 500 MSCOCO images). (a) Normalized entropy of visual attention across heads in layers 9 to 20. The distribution is bimodal with a small mode close to uniform; the dashed line marks the gate threshold τ. (b) Mean normalized entropy per layer and head. Stars mark heads on which the gate fires in more than 10% of decoding steps, and dashed lines delimit the intervention window.

Table 7: **Computational overhead** (LLaVA-1.5-7B, A6000, 512 tokens).

Method	**Mem**	**ΔMem**	**Lat**	**ΔLat**
Greedy	15.2 G	—	4.32 s	1.0×
+ ICD	15.4 G	+1%	8.6 s	2.0×
+ VCD	16.6 G	+9%	9.0 s	2.1×
+ SHIELD	18.2 G	+20%	7.3 s	1.7×
+ OPERA	22.8 G	+50%	30.8 s	7.1×
+ WAS	15.2 G	<1%	4.63 s	1.07×

Table 8: **MME** on LLaVA-1.5-7B (↑ for all).

Method	Percep.	Cogn.	Total
Greedy	1380.5	298.2	1678.7
+ ICD	1398.0	300.5	1698.5
+ DoLa	1390.2	295.8	1686.0
+ DeCo	1415.5	296.0	1711.5
+ OPERA	1410.8	**305.2**	1716.0
+ VCD	<u>1418.3</u>	302.4	<u>1720.7</u>
+ WAS	**1430.6**	<u>303.6</u>	**1734.2**

Computational overhead. Table 7 compares resource costs. The truncated SVD is computed once during prefill and cached; the per-step overhead is limited to a rank-r projection and scalar gating. WAS adds less than 1% memory and approximately 7% end-to-end latency, comparable to the lightest baselines and an order of magnitude cheaper than contrastive methods (VCD, ICD) that require a second forward pass.

Where the correction acts. The entropy gate confines WAS to a small set of heads. Across heads, the entropy of visual attention is bimodal (Figure 3a), with a small mode close to uniform separated from the main mode by a low-density valley. The threshold τ lies in this valley, so performance changes little as τ varies (Appendix H). The heads on which the gate fires are sparse and concentrated in the middle of the network (Figure 3b), the same layers in which the layer ablation finds the largest benefit, while restricting WAS to early or late layers weakens it substantially (Appendix H).

6 Conclusion

Hallucination in LVLMs is accompanied not by a shortage of attention to the image but by a loss of selection within it, and we traced this loss to visual key collapse, the concentration of visual keys along a few dominant directions. Whitened Attention Steering suppresses these directions while holding the attention budget fixed, and reduces hallucination in a graded and direction-specific way, with gains that follow the severity of collapse. Since collapse originates in the visual encoder and is amplified by the key projection, training objectives that keep visual keys diverse are a natural next step toward addressing hallucination at its source.

Table 9: **MME** on Qwen3-VL-4B.

Method	Percep.	Cogn.	Total
Greedy	1688.3	583.4	2271.7
+ VCD	1689.5	586.5	2276.0
+ VAF	<u>1690.2</u>	<u>590.2</u>	<u>2280.4</u>
+ WAS	**1690.4**	**621.1**	**2311.5**

AI USE STATEMENT

In this work, we used generative AI tools for language editing and coding assistance. We have not used generative AI tools for formulating the research problem, developing the core methodology, conducting experiments, analyzing results, or drawing scientific conclusions, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to improve the clarity, grammar, and presentation of the manuscript, as well as to assist with code implementation and debugging. We have reviewed all AI-assisted work. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code was inspected, tested, and validated for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29915–29926, 2025.

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*, 2025.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21635–21644, 2025.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*, 2025.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2534–2544, 2025.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*, 2025.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288, 2011.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. Whiteningbert: An easy unsupervised sentence embedding approach. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 238–244, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. Shield: Suppressing hallucinations in lvlm encoders via bias and vulnerability defense. *arXiv preprint arXiv:2510.16596*, 2025.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25004–25014, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *arXiv preprint arXiv:2506.11712*, 2025.

Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832, 2024.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. Intervene-all-paths: Unified mitigation of lvlm hallucinations across alignment formats. *arXiv preprint arXiv:2511.17254*, 2025.

Ruiyang Ren, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Tome: A two-stage approach for model-based retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6102–6114, 2023.

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.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arik, and Tomas Pfister. Mitigating object hallucination via data augmented contrastive tuning. *CoRR*, 2024.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Yuying Shang, Xinyi Zeng, Yutao Zhu, Xiao Yang, Zhengwei Fang, Jingyuan Zhang, Jiawei Chen, Zinan Liu, and Yu Tian. From pixels to tokens: Revisiting object hallucinations in large vision-language models. *arXiv preprint arXiv:2410.06795*, 2024.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29904–29914, 2025.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26147–26159, 2025b.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. *arXiv preprint arXiv:2507.00898*, 2025.

Jiaqi Wang, Yifei Gao, and Jitao Sang. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*, 2024a.

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.

Weixing Wang, Zifeng Ding, Jindong Gu, Rui Cao, Christoph Meinel, Gerard de Melo, and Haojin Yang. Image token matters: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing. *Advances in Neural Information Processing Systems*, 38: 111306–111335, 2026.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024b.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for mllms. In *International Conference on Learning Representations*, volume 2026, pp. 78875–78894, 2026.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14635–14645, 2025a.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10610–10620, 2025b.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14625–14634, 2025.

Dokyoon Yoon, Youngsook Song, and Woomyoung Park. Stop learning it all to mitigate visual hallucination, focus on the hallucination target. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4200–4208, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024a.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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*, pp. 13807–13816, 2024b.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption. 2023.

Kejia Zhang, Keda Tao, Jiasheng Tang, and Huan Wang. Poison as cure: Visual noise for mitigating object hallucinations in lvms. *arXiv preprint arXiv:2501.19164*, 2025.

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. *arXiv preprint arXiv:2310.00754*, 2023.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4189–4199, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

TABLE OF CONTENTS FOR APPENDIX

Appendix A. Layer-wise and Per-Head Key Geometry p. 15
Appendix B. Singular Value Spectra and Effective Rank p. 16
Appendix C. Allocation and Selection in Faithful and Hallucinated Captions p. 17
Appendix D. Collapse and Hallucination across Images p. 18
Appendix E. Why Does Collapse Occur? p. 18
Appendix F. Theoretical Analysis p. 19
F.1 Formalising Visual Key Collapse
F.2 Selective Spectral Correction
F.3 Score Matching Preserves Cross-Modal Balance
F.4 Entropy Gate
F.5 Computational Cost
Appendix G. Power Iteration: Algorithm and Convergence p. 22
Appendix H. Hyperparameter Sensitivity and Layer Selection p. 22
H.1 Hyperparameter Sensitivity
H.2 Intervention Layer Selection
Appendix I. Additional Results p. 24
I.1 Statistical Significance
I.2 SHIELD on Qwen3-VL-4B
I.3 Collapse Severity and Gain across Models
I.4 Counting Repeated Objects
Appendix J. Attention Map Visualization p. 25
Appendix K. Benchmark Evaluation Metrics p. 25
K.1 CHAIR
K.2 POPE
K.3 AMBER
K.4 MME
Appendix L. Extended Related Work p. 30
Appendix M. Limitations p. 30
Appendix N. Broader Impacts p. 30
Appendix O. Case Studies p. 31
Appendix P. Intervention Controls and Dose Response p. 31

A LAYER-WISE AND PER-HEAD KEY GEOMETRY

Figure 4 gives the layer-wise view behind Table 1. In LLaVA-1.5-7B, visual key norms match or exceed text key norms in layers 4 to 24, yet visual keys are the more concentrated modality. In InstructBLIP-7B, text key norms exceed visual key norms at every layer, and the visual keys still concentrate. Qwen3-VL-4B shows the mildest ratios. Figure 5 shows that the norm comparison holds

head by head: within each layer of LLaVA-1.5-7B, visual key norms vary little across heads, while text key norms vary widely.

Figure 1a stacks the 576 visual keys and the text keys of one image at layer 11, head 23, taken after RoPE without centering, normalization or subsampling, and embeds them jointly with t-SNE (perplexity [·], seed [·]).

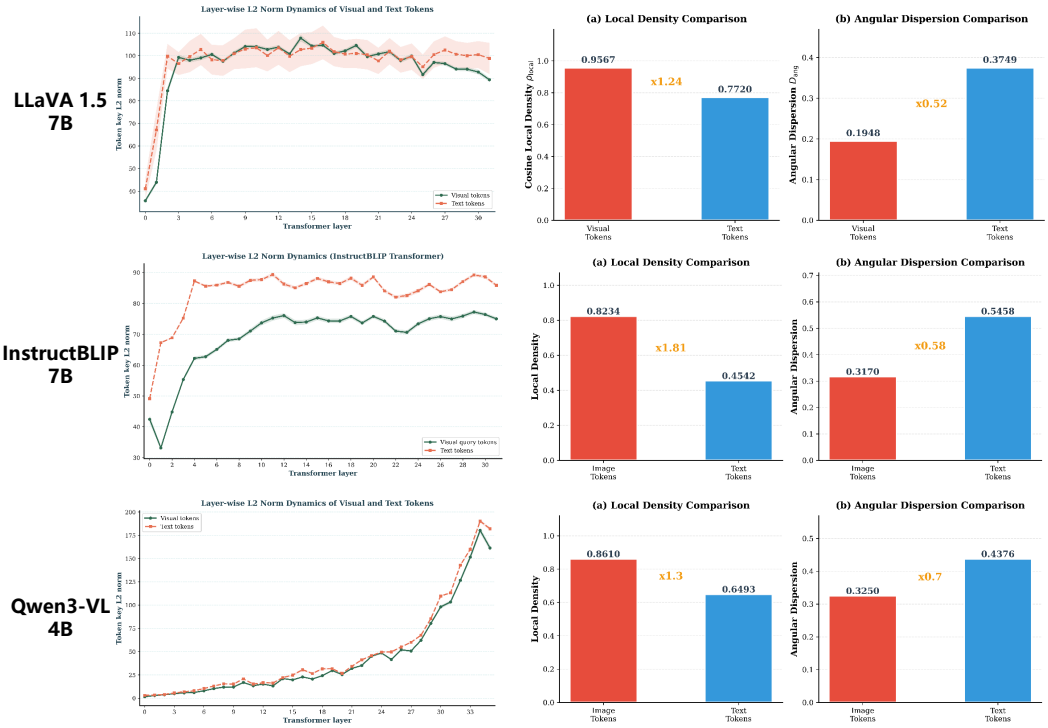


Figure 4: **Layer-wise key geometry of three LVLMs** (500 MSCOCO images). Left: mean ℓ_2 norm of visual and text keys per layer, with ± 1 s.d. bands. Centre and right: local density and angular dispersion aggregated over layers, with the visual-to-text ratio.

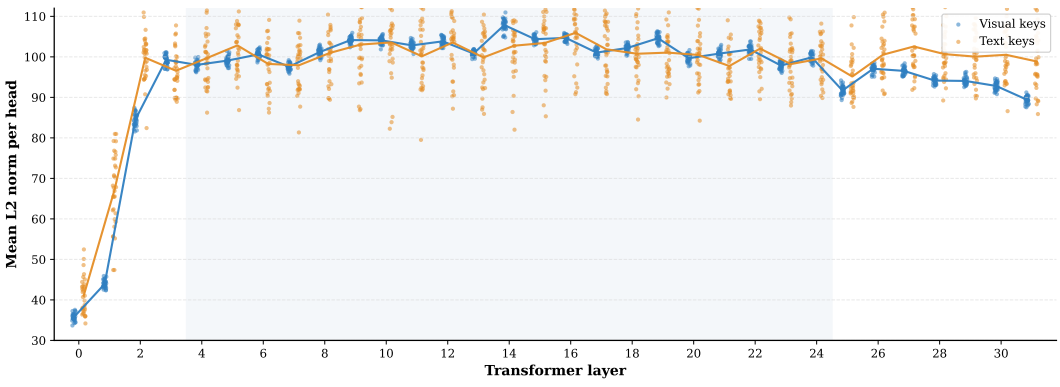


Figure 5: **Per-head mean ℓ_2 norm** of visual (blue) and text (orange) keys across the 32 layers of LLaVA-1.5-7B. Each dot is one head; lines connect head averages.

B Singular Value Spectra and Effective Rank

Figure 6 shows the spectrum of the visual key covariance $\mathbf{\Sigma}_K$ in LLaVA-1.5-7B. The first four to five components carry most of the variance, and the cumulative share reaches 85% at four components.

The spectrum then drops by about two orders of magnitude and decays log-linearly over the remaining directions, with $\sigma_5/\sigma_6 > 3$. Across layers 9 to 16, the top five components explain 88% to 94% of the variance, the effective rank $\mathrm{erank} = \exp(-\sum_i \pi_i \log \pi_i)$ with $\pi_i = \lambda_i / \sum_{i'} \lambda_{i'}$ (Roy & Vetterli, 2007) is 6 to 9 out of $d = 128$, and the number of components needed for 85% of the variance is 3 to 5. For text keys in the same layers, the top five components explain only 42% to 53% and the effective rank is 28 to 42. The spectral gap separates a stable dominant subspace from a tail of poorly separated directions, which motivates the truncation in WAS (Section 4.1) and explains its insensitivity to r (Table 14).

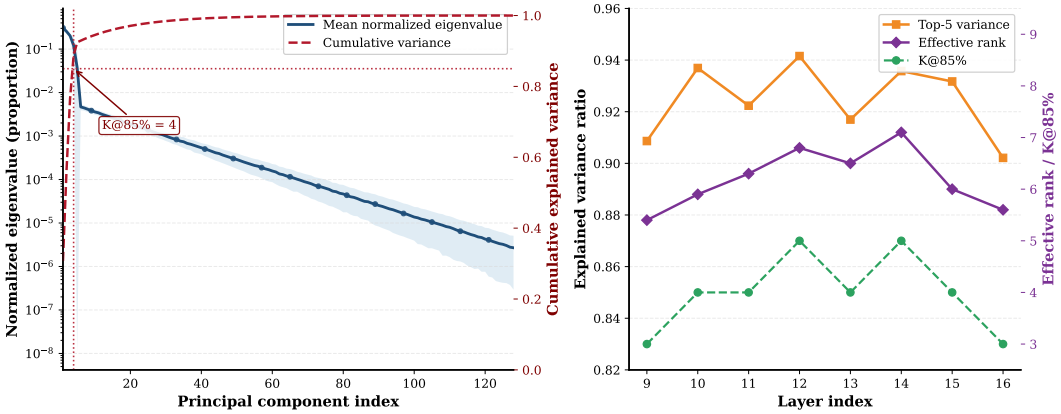


Figure 6: **Spectrum of the visual key covariance in LLaVA-1.5-7B.** Left: normalized eigenvalues (log scale) and cumulative explained variance, averaged over heads in layers 9 to 16 and 500 MSCOCO images. Right: top-5 explained variance, effective rank and components needed for 85% of the variance, per layer.

C Allocation and Selection in Faithful and Hallucinated Captions

The CHAIR pipeline splits the 500 greedy captions of LLaVA-1.5-7B into 229 faithful and 271 hallucinated captions. For a caption of T generated tokens and a head in layers [L_a to L_b], we compute

$$A_t = \sum_{j=1}^{N_v} a_{t,j}, \qquad \bar{H}_t = \frac{H(\mathbf{p}_t)}{\log N_v}, \qquad Q = \frac{2}{T(T-1)} \sum_{t<t'} \mathrm{JSD}(\mathbf{p}_t \,\|\, \mathbf{p}_{t'}), \tag{9}$$

with $\mathbf{p}_t = \mathbf{a}_t / A_t$,the dominant score share $(\sigma_s^{\parallel})^2/\sigma_s^2$ of Eq. equation 13, and average over steps and heads. Q is the pairwise counterpart of the mutual information in Proposition 2. Table 10 reports group means, paired differences with 95% intervals from 10,000 image-level bootstrap resamples, and the ROC-AUC of each quantity for separating the groups.

Table 10: **Faithful versus hallucinated captions** of LLaVA-1.5-7B. AUC is taken in the direction of the observed difference, with 0.5 at chance.

Quantity	Faithful	Hallucinated	Difference [95% CI]	AUC
Allocation A	0.384	0.388	$+0.004$ $[-0.006, +0.014]$	0.52
Query dependence Q	0.094	0.064	-0.030 $[-0.036, -0.024]$	0.71
Normalized entropy $\bar{H}$	0.645	0.712	$+0.067$ $[+0.052, +0.080]$	0.69
Dominant score share	0.612	0.703	$+0.091$ $[+0.073, +0.109]$	0.73
Top-5 concentration (L11)	0.892	0.923	$+0.031$ $[+0.026, +0.039]$	0.72

D COLLAPSE AND HALLUCINATION ACROSS IMAGES

Figure 1b relates the collapse of the visual keys of each image at layer 11 to the number of hallucinated objects in its caption. We test whether this association survives image and caption properties with a negative binomial regression

$$\log \mathbb{E}[Y_n] = \beta_0 + \beta_c z_n + \boldsymbol{\gamma}^\top \mathbf{x}_n, \tag{10}$$

where Y_n is the number of hallucinated objects for image n, z_n a standardized collapse statistic, and $\mathbf{x}_n$ ten covariates: ground-truth instance count, category count, category entropy, box occupancy, small-object ratio, crowd count and aspect ratio of the image, and token count, noun count and object-noun count of the caption. The dispersion is 1.84, the maximum variance inflation factor is 1.92, and intervals come from 10,000 image-level bootstrap resamples. Collapse retains about two thirds of its unadjusted effect (Table 11). Matching more and less collapsed images one to one without replacement on $\mathbf{x}_n$ gives 0.44 more hallucinated objects per image $[0.28, 0.60]$ and 0.180 against 0.123 hallucinations per object noun.

Visual keys of neighbouring patches are naturally similar (Ren et al., 2023). Table 12 stratifies key pairs of a head at layer 11 by distance on the 24×24 patch grid. The excess similarity of hallucinated images is present at every distance and grows with distance, which local redundancy would not produce.

Table 11: **Negative binomial regression** of hallucinated-object count on standardized collapse.

Predictor	Unadj. β	Adj. β [95% CI]	IRR	p
Top-5 conc.	$+0.41$	$+0.27\,[+0.16, +0.38]$	1.31	$< 10^{-4}$
Angular	$+0.35$	$+0.22\,[+0.11, +0.33]$	1.25	10^{-4}
Eff. rank	$+0.30$	$+0.18\,[+0.07, +0.29]$	1.20	0.0014

Table 12: **Mean key cosine by patch distance**.

Distance	Faith.	Hall.	Diff. [95% CI]
Adjacent	0.90	0.94	$+0.04\,[.03, .06]$
Near	0.87	0.92	$+0.05\,[.04, .08]$
Medium	0.83	0.89	$+0.06\,[.05, .07]$
Far	0.79	0.86	$+0.07\,[.05, .09]$

E WHY DOES COLLAPSE OCCUR?

Source 1: encoder anisotropy. CLIP ViT-L/14 is trained to align a global image representation with text, which does not reward discriminability among patches. The encoder output is already concentrated, with $D_{\text{ang}} = 0.14$ against about 0.99 for random vectors in $\mathbb{R}^{1024}$.

Source 2: the key projection amplifies it. The key projections of layers 9 to 20 have condition numbers $\kappa(\mathbf{W}_K) = \sigma_{\max}/\sigma_{\min}$ of 15 to 45, against 5 to 12 for the query and value projections. To separate the effect of the key projection from that of depth, we take at each layer the normalized hidden state entering attention and compare four views of it by how well their top-5 concentration separates faithful from hallucinated captions: a random semi-orthogonal projection to 128 dimensions aggregated like the heads (matched hidden), the value, and the key before and after RoPE (Table 13). The key exceeds the matched hidden control in layers 0 to 20, the value does not, RoPE slightly lowers the gain, and the gain disappears beyond layer 24.

Table 13: **Separability (ROC-AUC) of faithful and hallucinated captions** by four views of the same hidden state (LLaVA-1.5-7B, measured at prefill).

Layer	Matched hidden	Value	Pre-RoPE key	Post-RoPE key
0	0.655	0.663	0.706	0.703
11	0.676	0.684	0.723	0.720
17	0.696	0.702	0.751	0.748
20	0.700	0.707	0.734	0.731
24	0.705	0.709	0.718	0.716
31	0.702	0.705	0.693	0.691

Source 3: weak gradients through flat attention. Training may fail to correct collapse because flat attention transmits little gradient to the visual scores.

Lemma 1 (Softmax Jacobian)**.** *For* $\mathbf{p} = \mathrm{softmax}(\mathbf{s})$*, the Jacobian* $\mathbf{J} = \partial\mathbf{p}/\partial\mathbf{s} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ *satisfies* $\mathbf{0} \preceq \mathbf{J} \preceq \mathrm{diag}(\mathbf{p})$*, so* $\|\mathbf{J}\|_2 \le \max_j p_j$*.*

Proof. $\mathbf{J}$ is the covariance of the one-hot vector $\mathbf{e}_J$ with $J \sim \mathbf{p}$, hence $\mathbf{J} \succeq \mathbf{0}$, and $\mathrm{diag}(\mathbf{p}) - \mathbf{J} = \mathbf{p}\mathbf{p}^\top \succeq \mathbf{0}$. □

For the attention output $\mathbf{o} = \sum_j p_j \mathbf{v}_j$ and a loss $\mathcal{L}$, $\partial\mathcal{L}/\partial\mathbf{s} = \mathbf{J}\mathbf{V}\nabla_{\mathbf{o}}\mathcal{L}$, so $\|\partial\mathcal{L}/\partial\mathbf{s}\|_2 \le \max_j p_j \, \|\mathbf{V}\nabla_{\mathbf{o}}\mathcal{L}\|_2$. Under near-uniform attention $\max_j p_j \approx 1/N_v$, so the signal that would reshape the visual keys is weakest exactly where collapse is strongest. We state this as a hypothesis about training dynamics.

Why text keys are spared. Text keys inherit representations from a language model pretrained to distinguish a large vocabulary, whereas visual tokens come from an encoder trained for global alignment. Training objectives that keep the spectrum of visual keys spread out are a natural complement to WAS.

F Theoretical Analysis

This section formalizes three aspects of the analysis and of WAS. Collapsed visual keys concentrate the logit variation in a few spectral modes, which limits how much the query determines where attention goes. Truncated spectral whitening attenuates these modes. Score matching preserves the cross-modal attention budget. Throughout, d is the per-head dimension, N_v the number of visual tokens, $s_j = \mathbf{q}^\top\mathbf{k}_j/\sqrt{d}$, and

$$\boldsymbol{\mu} = \frac{1}{N_v}\sum_{j=1}^{N_v}\mathbf{k}_j, \qquad \boldsymbol{\Sigma}_K = \frac{1}{N_v}\sum_{j=1}^{N_v}(\mathbf{k}_j - \boldsymbol{\mu})(\mathbf{k}_j - \boldsymbol{\mu})^\top = \sum_{i=1}^{d}\lambda_i\mathbf{v}_i\mathbf{v}_i^\top, \qquad \lambda_1 \ge \cdots \ge \lambda_d \ge 0. \tag{11}$$

F.1 Formalising Visual Key Collapse

Definition 1 ((r,ϵ,γ)-collapse)**.** *Visual keys* $\{\mathbf{k}_j\}_{j=1}^{N_v}$ *exhibit* (r,ϵ,γ)*-collapse if*

$$\sum_{i=1}^{r}\lambda_i \ge (1-\epsilon)\,\mathrm{tr}(\boldsymbol{\Sigma}_K), \qquad \mathrm{CV}\big(\{\|\mathbf{k}_j\|\}_{j=1}^{N_v}\big) \le \gamma, \tag{12}$$

where $\mathrm{CV} = \mathrm{std}/\mathrm{mean}$*. Collapsed heads of LLaVA-1.5-7B satisfy* $r = 5$*,* $\epsilon \le 0.15$ *and* $\gamma \approx 0.05$*.*

The mean score $\bar{s} = \mathbf{q}^\top\boldsymbol{\mu}/\sqrt{d}$ is the term shared by all visual tokens (Eq. equation 1), so the variation of the scores across positions comes only from the deviations $\mathbf{k}_j - \boldsymbol{\mu}$ and decomposes in the eigenbasis of $\boldsymbol{\Sigma}_K$:

$$\sigma_s^2(\mathbf{q}) \triangleq \frac{1}{N_v}\sum_j (s_j - \bar{s})^2 = \frac{1}{d}\,\mathbf{q}^\top\boldsymbol{\Sigma}_K\mathbf{q} = \underbrace{\frac{1}{d}\sum_{i\le r}\lambda_i(\mathbf{q}^\top\mathbf{v}_i)^2}_{(\sigma_s^{\parallel})^2} + \underbrace{\frac{1}{d}\sum_{i>r}\lambda_i(\mathbf{q}^\top\mathbf{v}_i)^2}_{(\sigma_s^{\perp})^2}. \tag{13}$$

Proposition 1 (Expected residual logit energy)**.** *If* $\mathbf{q}/\|\mathbf{q}\|$ *is uniform on the unit sphere and independent of* $\{\mathbf{v}_i\}$*, then under* (r,ϵ,γ)*-collapse*

$$\frac{\mathbb{E}_{\mathbf{q}}[(\sigma_s^{\perp})^2]}{\mathbb{E}_{\mathbf{q}}[\sigma_s^2(\mathbf{q})]} = \frac{\sum_{i>r}\lambda_i}{\mathrm{tr}(\boldsymbol{\Sigma}_K)} \le \epsilon. \tag{14}$$

Proof. By rotational invariance $\mathbb{E}[(\mathbf{q}^\top\mathbf{v}_i)^2] = \|\mathbf{q}\|^2/d$ for all i. Substituting into Eq. equation 13 gives $\mathbb{E}[(\sigma_s^{\perp})^2] = \frac{\|\mathbf{q}\|^2}{d^2}\sum_{i>r}\lambda_i$ and $\mathbb{E}[\sigma_s^2] = \frac{\|\mathbf{q}\|^2}{d^2}\sum_i\lambda_i$. The ratio and Eq. equation 12 give the bound. □

The isotropic query is an idealization, since trained queries co-adapt with keys. Measured across all heads in layers 9 to 20 of LLaVA-1.5-7B on 500 MSCOCO images, $(\sigma_s^{\perp})^2/\sigma_s^2$ has median 0.11 and mean 0.14, consistent with $\epsilon \leq 0.15$. The residual directions thus carry little of the score variation, which motivates attenuating the dominant modes rather than amplifying all directions.

The consequence for selection is captured by the mutual information between the query and the attended position. Let $\mathbf{q}_1, \ldots, \mathbf{q}_T$ be the queries of a caption, let the step t be uniform on $\{1, \ldots, T\}$, and let $J \mid t \sim \mathbf{p}_t$, the attention distribution over visual tokens at step t, with $\bar{\mathbf{p}} = \frac{1}{T}\sum_t \mathbf{p}_t$.

Proposition 2 (Mutual information and query dependence)**.**

$$I(\mathbf{q}; J) = H(\bar{\mathbf{p}}) - \frac{1}{T}\sum_t H(\mathbf{p}_t) = \mathrm{JSD}(\mathbf{p}_1, \ldots, \mathbf{p}_T) \leq (1 - \bar{H}) \log N_v, \tag{15}$$

where $\bar{H} = \frac{1}{T}\sum_t H(\mathbf{p}_t)/\log N_v$.

Proof. The first equality is $I = H(J) - H(J \mid t)$ with marginal $\bar{\mathbf{p}}$, and the second is the definition of the generalized Jensen-Shannon divergence with uniform weights. For the bound, $I = \frac{1}{T}\sum_t \mathrm{KL}(\mathbf{p}_t \| \bar{\mathbf{p}}) = \frac{1}{T}\sum_t \mathrm{KL}(\mathbf{p}_t \| \mathbf{u}) - \mathrm{KL}(\bar{\mathbf{p}} \| \mathbf{u}) \leq \frac{1}{T}\sum_t (\log N_v - H(\mathbf{p}_t))$ for the uniform distribution $\mathbf{u}$. □

Under collapse, the scores of every query are dominated by the same few modes, so the distributions $\mathbf{p}_t$ differ little and $I(\mathbf{q}; J)$ is small. When the attention of a head becomes nearly uniform, Eq. equation 15 bounds $I(\mathbf{q}; J)$ directly. The query dependence of Section 3 is the pairwise counterpart of this quantity.

F.2 Selective Spectral Correction

Eigenvalue form. Let $\mathbf{T} = \mathbf{I}_d + \mathbf{V}_r(\mathbf{\Lambda}_r^{-1/2} - \mathbf{I}_r)\mathbf{V}_r^{\top}$ with $\mathbf{\Lambda}_r = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, so that $\mathbf{T}\mathbf{v}_i = \lambda_i^{-1/2}\mathbf{v}_i$ for $i \leq r$ and $\mathbf{T}\mathbf{v}_i = \mathbf{v}_i$ otherwise. With corrected keys $\tilde{\mathbf{k}}_j = \boldsymbol{\mu} + \mathbf{T}(\mathbf{k}_j - \boldsymbol{\mu})$,

$$\tilde{s}_j^{(\mathrm{cov})} = s_j + \frac{1}{\sqrt{d}}\,\mathbf{q}^{\top}\mathbf{V}_r(\mathbf{\Lambda}_r^{-1/2} - \mathbf{I}_r)\mathbf{V}_r^{\top}(\mathbf{k}_j - \boldsymbol{\mu}), \qquad \tilde{s}_j^{(\mathrm{cov})} - \tilde{s}_{j'}^{(\mathrm{cov})} = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} w_i\, q_i(\bar{k}_{ji} - \bar{k}_{j'i}), \tag{16}$$

with $w_i = \lambda_i^{-1/2}$ for $i \leq r$, $w_i = 1$ for $i > r$, $q_i = \mathbf{q}^{\top}\mathbf{v}_i$ and $\bar{k}_{ji} = \mathbf{v}_i^{\top}(\mathbf{k}_j - \boldsymbol{\mu})$.

SVD form. Algorithm 1 uses the truncated SVD $\bar{\mathbf{K}}_v \approx \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^{\top}$, which has the same $\mathbf{V}_r$ and $\sigma_i = \sqrt{N_v\lambda_i}$. It differs from the eigenvalue form in (a) rescaling by $\mathbf{\Sigma}_r^{-1}$ instead of $\mathbf{\Lambda}_r^{-1/2}$ and (b) acting on uncentered keys. Both differences vanish after score matching.

Proposition 3 (SVD–eigenvalue decomposition)**.** *The SVD-form logit (Algorithm 1, line 6) decomposes as*

$$\tilde{s}_j^{(\mathrm{svd})} = \underbrace{\frac{1}{\sqrt{d}}\sum_{i \leq r} N_v^{-1/2}\lambda_i^{-1/2} q_i \bar{k}_{ji}}_{\Delta_j^{\parallel}} + \underbrace{\frac{1}{\sqrt{d}}\sum_{i > r} q_i \bar{k}_{ji}}_{\Delta_j^{\perp}} + C_{\mathrm{svd}}, \quad C_{\mathrm{svd}} = \frac{1}{\sqrt{d}}\Big(\sum_{i \leq r}\sigma_i^{-1} q_i\mu_i + \sum_{i>r} q_i\mu_i\Big), \tag{17}$$

with $\mu_i = \mathbf{v}_i^{\top}\boldsymbol{\mu}$*. The eigenvalue form has the same decomposition with* $\lambda_i^{-1/2}$ *in place of* $N_v^{-1/2}\lambda_i^{-1/2}$ *and* $C_{\mathrm{cov}} = \mathbf{q}^{\top}\boldsymbol{\mu}/\sqrt{d}$*.*

Proof. Line 6 gives $\tilde{s}_j^{(\mathrm{svd})} = \frac{1}{\sqrt{d}}\mathbf{q}^{\top}\mathbf{M}\mathbf{k}_j$ with $\mathbf{M} = \mathbf{I}_d + \mathbf{V}_r(\mathbf{\Sigma}_r^{-1} - \mathbf{I}_r)\mathbf{V}_r^{\top}$, which scales $\mathbf{v}_i$ by $\sigma_i^{-1} = N_v^{-1/2}\lambda_i^{-1/2}$ for $i \leq r$ and leaves the rest unchanged. Writing $\mathbf{k}_j = (\mathbf{k}_j - \boldsymbol{\mu}) + \boldsymbol{\mu}$ and expanding in $\{\mathbf{v}_i\}$ separates the j-dependent terms from C_{svd}. The eigenvalue form follows in the same way from Eq. equation 16. □

By Proposition 4, score matching removes C_{svd} and C_{cov}, which resolves (b). The pairwise differences that set the ranking are then

$$\tilde{s}_j - \tilde{s}_{j'} = \frac{1}{\sqrt{d}}\Big[\sum_{i\le r} \omega_i\, q_i(\bar{k}_{ji} - \bar{k}_{j'i}) + \sum_{i>r} q_i(\bar{k}_{ji} - \bar{k}_{j'i})\Big], \qquad \omega_i^{(\mathrm{svd})} = N_v^{-1/2}\,\omega_i^{(\mathrm{cov})} = N_v^{-1/2}\lambda_i^{-1/2}. \tag{18}$$

The residual sum is identical in both forms, and (a) reduces to a common factor on already suppressed terms. In the intervention layers of LLaVA-1.5-7B, $\omega_i^{(\mathrm{cov})} < 0.03$ and $\omega_i^{(\mathrm{svd})} < 0.002$ for all $i \le r$, so the ranking is governed by the residual subspace in both forms. Across all heads in layers 9 to 20 on 500 MSCOCO images, the Spearman correlation between $\tilde{\mathbf{s}}^{(\mathrm{svd})}$ and $\tilde{\mathbf{s}}^{(\mathrm{cov})}$ exceeds 0.999.

F.3 Score Matching Preserves Cross-Modal Balance

Proposition 4 (Score-matching invariance)**.** *For* $\mathrm{SM}(\mathbf{x})_j = \mu_o + \alpha\sigma_o\big(x_j - \mathrm{mean}(\mathbf{x})\big)/\mathrm{std}(\mathbf{x})$ *and any* $\beta > 0$, $\gamma \in \mathbb{R}$, $\mathrm{SM}(\beta\mathbf{x} + \gamma\mathbf{1}) = \mathrm{SM}(\mathbf{x})$.

Proof. $\mathrm{mean}(\beta\mathbf{x} + \gamma\mathbf{1}) = \beta\,\mathrm{mean}(\mathbf{x}) + \gamma$ and $\mathrm{std}(\beta\mathbf{x} + \gamma\mathbf{1}) = \beta\,\mathrm{std}(\mathbf{x})$, so β and γ cancel. □

The matched scores satisfy $\mathrm{mean}(\mathbf{s}^\star) = \mu_o$, which is exactly the shared term, and $\mathrm{std}(\mathbf{s}^\star) = \alpha\sigma_o$. The visual attention mass is $A = \mathrm{sig}\big(\mathrm{LSE}(\mathbf{s}) - \mathrm{LSE}(\mathbf{s}^{\mathrm{txt}})\big)$ with $\mathrm{sig}(x) = (1 + e^{-x})^{-1}$, and since $\mathrm{sig}' \le \frac{1}{4}$,

$$|A^\star - A| \le \tfrac{1}{4}\big|\mathrm{LSE}(\mathbf{s}^\star) - \mathrm{LSE}(\mathbf{s})\big|. \tag{19}$$

If the visual logits are close to Gaussian, which the Shapiro-Wilk test fails to reject at $p = 0.05$ for 82% of head-layer pairs in the intervention window, then $\mathbb{E}\exp(S) = \exp(\mu + \sigma^2/2)$ for $S \sim \mathcal{N}(\mu, \sigma^2)$ gives

$$\mathrm{LSE}(\mathbf{s}) \approx \log N_v + \mu + \frac{\sigma^2}{2}, \qquad \mathrm{LSE}(\mathbf{s}^\star) - \mathrm{LSE}(\mathbf{s}) \approx \frac{(\alpha^2 - 1)\sigma_o^2}{2}. \tag{20}$$

The shift vanishes at $\alpha = 1$. At $\alpha = 1.2$ and a typical $\sigma_o \approx 0.5$ it is 0.055, far below the visual-text LSE gap of 3 to 5. Direct measurement gives an average shift of 0.06 across the intervention layers, with a maximum of 0.3.

F.4 Entropy Gate

For the visual attention distribution $\mathbf{a}$ and the uniform distribution $\mathbf{u}$,

$$D_{\mathrm{KL}}(\mathbf{a} \,\|\, \mathbf{u}) = \sum_j a_j \log(N_v a_j) = \log N_v - H_{\mathrm{vis}} = (1 - \bar{H}_{\mathrm{vis}}) \log N_v, \tag{21}$$

so $\bar{H}_{\mathrm{vis}} \to 1$ means vanishing visual selectivity, and by Proposition 2 it also bounds the mutual information between the query and the attended position. The gate $g = \mathrm{sig}\big((\bar{H}_{\mathrm{vis}} - \tau)/\beta\big)$ satisfies, for $\varepsilon \in (0, \frac{1}{2})$,

$$g \le \varepsilon \iff \bar{H}_{\mathrm{vis}} \le \tau - \beta \log \tfrac{1-\varepsilon}{\varepsilon}, \qquad g \ge 1 - \varepsilon \iff \bar{H}_{\mathrm{vis}} \ge \tau + \beta \log \tfrac{1-\varepsilon}{\varepsilon}. \tag{22}$$

At $\tau = 0.9$ and $\beta = 0.02$, $g < 0.01$ for $\bar{H}_{\mathrm{vis}} < 0.81$ and $g > 0.95$ for $\bar{H}_{\mathrm{vis}} > 0.96$. The midpoint $g = 0.5$ lies at $\bar{H}_{\mathrm{vis}} = \tau$, where $D_{\mathrm{KL}}(\mathbf{a}\|\mathbf{u}) = 0.1 \log N_v \approx 0.64$ for $N_v = 576$. The two modes of Figure 3a, near 0.67 and 0.96, receive $g \approx 10^{-5}$ and $g \approx 0.95$. Because the modes are separated by a density valley in $[0.82, 0.90]$ and β is small, any threshold in the valley gives nearly the same partition of heads, which matches the weak sensitivity to τ in Table 14.

F.5 Computational Cost

Randomized power iteration (Halko et al., 2011) with oversampling $p = 5$ and $q = 2$ iterationscosts $\mathcal{O}(qN_v d(r + p))$ per head at prefill. With $\mathbf{K}_v\mathbf{V}_r$ cached at prefill, each decoding step costs

$$\mathcal{O}(dr) + \mathcal{O}(N_v r) + \mathcal{O}(N_v) = \mathcal{O}\big((d + N_v)r + N_v\big) \tag{23}$$

per head, for the rank-r projection, the correction and the score matching with entropy. This is dominated by standard attention. The cached $\mathbf{V}_r$ and $\mathbf{\Sigma}_r^{-1}$ take fewer than 700 floats per head, and $\mathbf{K}_v\mathbf{V}_r$ takes $N_v r$ more.

G Power Iteration: Algorithm and Convergence

We compute the truncated SVD $\bar{\mathbf{K}}_v \approx \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top$ by randomized power iteration (Halko et al., 2011), which suits our setting because $r \ll \min(N_v, d)$ and the spectral gap is large. Algorithm 2 draws a Gaussian matrix $\mathbf{\Omega} \in \mathbb{R}^{d \times (r+p)}$ with oversampling $p = 5$, refines the sample matrix $\mathbf{Y} = \bar{\mathbf{K}}_v \mathbf{\Omega}$ through q power iterations, and extracts the SVD of the small projected matrix.

Algorithm 2 Randomized truncated SVD via power iteration

Require: Centered visual keys $\bar{\mathbf{K}}_v \in \mathbb{R}^{N_v \times d}$, rank r, oversampling p, power iterations q
Ensure: $\mathbf{U}_r, \mathbf{\Sigma}_r, \mathbf{V}_r$

```
Draw Ω ~ N(0,1)^{d×(r+p)}
Y ← K̄_v Ω                                   ▷ N_v × (r + p)
for i = 1, ..., q do
    Y ← K̄_v (K̄_v^⊤ Y)
end for
Q, _ ← QR(Y)
B ← Q^⊤ K̄_v                                 ▷ (r + p) × d
Û, Σ_{r+p}, V_{r+p} ← SVD(B)
U_{r+p} ← QÛ; truncate to the top r components
```

Convergence. With $q = 2$ and $p = 5$, the relative error $\|\bar{\mathbf{K}}_v - \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top\|_F / \|\bar{\mathbf{K}}_v\|_F$ stays below 0.08 across all intervention layers and heads. The large spectral gap ($\sigma_5/\sigma_6 > 3$) ensures rapid convergence, and $q = 1$ gives nearly identical results.

Cost. Each power iteration costs $\mathcal{O}(N_v d(r + p))$. With $N_v = 576$, $d = 128$, $r = 5$, $p = 5$ and $q = 2$, the total is about 1.5M multiply-adds per head, negligible compared with attention itself. The decomposition is computed once at prefill and does not add per-token latency.

H Hyperparameter Sensitivity and Layer Selection

All experiments in this section use LLaVA-1.5-7B on CHAIR (500 MSCOCO images, greedy decoding) with the default configuration ($\alpha = 1.2$, $r = 5$, $\tau = 0.9$, $\beta = 0.02$, layers 9 to 20) unless noted.

H.1 Hyperparameter Sensitivity

Table 14 sweeps each hyperparameter with the others held at their defaults.

Truncation rank r. Setting $r = d$ reduces WAS to full whitening $\mathbf{\Sigma}_K^{-1/2}$, which equalizes all $d = 128$ spectral directions, including the roughly 120 tail directions whose eigenvalues are poorly estimated from $N_v = 576$ keys. Inverting these small eigenvalues amplifies estimation noise, and CHAIR_S degrades to 46.8, worse than spectral whitening alone (41.2, Table 4), because score matching faithfully preserves a noise-corrupted ranking. For r from 3 to 12, CHAIR_S varies by at most 0.6 points, reflecting the sharp spectral gap between the few inflated eigenvalues and the remainder (Appendix B). The benefit of WAS therefore comes from targeted attenuation of the collapse subspace rather than from generic isotropization.

Sharpening factor α. Performance has a mild optimum at $\alpha = 1.2$. At $\alpha = 1.0$, the corrected scores reproduce the original spread, which collapse has already compressed, and a modest $\alpha > 1$ amplifies the ranking exposed by the correction. Values above 1.6 over-amplify the spread, shift the visual-text balance, and slightly increase hallucination.

Entropy threshold τ. Performance is stable over $[0.75, 0.95]$, with a gentle optimum at 0.9. The entropy of visual attention is bimodal, and a low-density valley in $[0.82, 0.90]$ separates collapsed heads from the rest (Figure 3a, Appendix F.4), so thresholds near this valley give similar partitions.

Gate temperature β. Within $[0.005, 0.05]$ the gate is sharp enough to separate the two entropy modes, and CHAIR$_S$ varies by at most 0.4 points. At $\beta = 1.0$, $g = \mathrm{sig}(\bar{H}_{\mathrm{vis}} - \tau)$ stays between 0.40 and 0.53 for $\bar{H}_{\mathrm{vis}} \in [0.5, 1]$, so all heads are blended almost equally and head-level selectivity is lost. CHAIR$_S$ then rises by 1.2 points, which shows that correcting collapsed heads selectively outperforms uniform blending.

Excluding the negative control $r = d$, no hyperparameter changes CHAIR$_S$ by more than two points over its tested range, which supports a single configuration across all models and benchmarks.

Table 14: **Hyperparameter sensitivity** on LLaVA-1.5-7B (CHAIR). Each parameter is swept with the others at their defaults ($\alpha = 1.2$, $r = 5$, $\tau = 0.9$, $\beta = 0.02$). Full whitening ($r = d$) amplifies estimation noise and degrades performance.

Param	Value	CHAIR$_S$↓	CHAIR$_I$↓	Recall↑	CFAR↑
r	3	30.8	10.4	70.1	65.0
	5	**30.2**	**10.2**	**70.7**	**66.1**
	8	30.4	10.2	70.5	65.8
	12	30.4	10.3	70.4	65.5
	d (full whitening)	46.8	15.6	66.8	53.5
α	1.0	31.8	10.6	70.2	64.4
	1.2	**30.2**	**10.2**	**70.7**	**66.1**
	1.4	30.6	10.3	70.5	65.6
	1.6	30.8	10.4	70.3	65.2
	1.8	31.4	10.5	70.0	64.6
τ	0.75	31.5	10.8	70.0	64.1
	0.80	31.0	10.6	70.2	64.8
	0.85	30.5	10.3	70.5	65.6
	0.90	**30.2**	**10.2**	**70.7**	**66.1**
	0.95	30.6	10.4	70.3	65.4
β	0.005	30.6	10.3	70.4	65.5
	0.02	**30.2**	**10.2**	**70.7**	**66.1**
	0.05	30.4	10.3	70.6	65.8
	1.0 *(no temperature)*	31.4	10.6	70.2	64.7

H.2 Intervention Layer Selection

Table 15 varies the intervention window with all other parameters fixed. Applying WAS to layers 9 to 16 already reduces CHAIR$_S$ by 22.2 points relative to greedy decoding, and extending the window to layer 20 adds a further 1.8 points while recall rises from 69.5 to 70.7. This matches the middle layers in which the entropy gate fires most often (Figure 3b). Extending the window to layer 28 or to the full depth reduces the gain, since it perturbs layers where the correction brings little benefit. Early-only (layers 1 to 8) and late-only (layers 20 to 28) windows are substantially weaker. The layer selection is therefore tied to where collapse matters rather than acting as a generic decoding bias, and we use layers 9 to 20 for all architectures and benchmarks.

Table 15: **Intervention layer range** on LLaVA-1.5-7B (CHAIR). L9–20 (shaded) is used in all main experiments.

Layer range	CHAIR$_S$↓	CHAIR$_I$↓	Recall↑	CFAR↑
L1–8 *(early)*	42.5	14.8	73.2	53.6
L5–12	35.2	12.0	68.5	58.8
L9–16	32.0	10.8	69.5	63.8
L9–20 *(ours)*	**30.2**	**10.2**	70.7	**66.1**
L9–28	31.8	10.6	70.2	64.6
L0–31 *(all)*	34.5	12.2	69.0	60.8
L20–28 *(late)*	48.0	16.2	73.0	52.5

I Additional Results

I.1 Statistical Significance

Table 16 reports paired differences between WAS and greedy decoding with 95% intervals from 10,000 bootstrap resamples, taken over images for CHAIR and AMBER, over questions stratified by split for POPE, and over samples stratified by subtask for MME, with Holm correction across the family of tests. Every interval excludes zero.

Table 16: **Paired bootstrap tests**, WAS minus greedy decoding. Differences are computed from Tables 2, 6, 5, 8 and 9.

Model	Benchmark, metric	Difference	95% CI	Holm p
LLaVA-1.5-7B	CHAIR, CHAIR_S	-24.0	$[-28.4, -19.6]$	0.0016
LLaVA-1.5-7B	POPE Adv., F1	$+6.41$	$[+4.70, +8.10]$	0.0016
InstructBLIP-7B	POPE Adv., F1	$+7.42$	$[+5.60, +9.20]$	0.0016
Qwen3-VL-4B	POPE Adv., F1	$+3.01$	$[+1.80, +4.20]$	0.0016
LLaVA-1.5-7B	AMBER, Hal	-27.8	$[-31.8, -23.8]$	0.0016
LLaVA-1.5-7B	MME, Total	$+55.5$	$[+24.0, +87.0]$	0.0030
Qwen3-VL-4B	MME, Total	$+39.8$	$[+15.4, +64.2]$	0.0030

I.2 SHIELD on Qwen3-VL-4B

SHIELD is the strongest baseline on LLaVA-1.5 and InstructBLIP (Table 2). We port it to Qwen3-VL-4B under the same protocol (Table 17). WAS is ahead on every hallucination metric, with intervals excluding zero on POPE and AMBER. The two methods are statistically indistinguishable on MME Total, a measure of general ability rather than hallucination. The latency added by WAS is about one sixth of that added by SHIELD.

Table 17: **SHIELD versus WAS on Qwen3-VL-4B.** WAS values match Tables 3, 6, 5 and 9; latency is relative to greedy decoding on this model.

Metric	SHIELD	WAS	WAS − SHIELD [95% CI]
$\text{CHAIR}_S \downarrow$	21.6	**17.8**	-3.8
$\text{CHAIR}_I \downarrow$	8.1	**7.2**	-0.9
Recall↑	48.9	**50.2**	$+1.3$
CFAR↑	56.5	**59.1**	$+2.6$
POPE Adv. F1↑	87.40	**88.21**	$+0.81$ $[+0.10, +1.52]$
AMBER Hal↓	23.0	**19.0**	-4.0 $[-6.8, -1.2]$
MME Total↑	2302.0	**2311.5**	$+9.5$ $[-3.0, +22.0]$
Latency↓	$1.43\times$	$\mathbf{1.07}\times$	–

I.3 Collapse Severity and Gain across Models

Table 18 adds Qwen3-VL-8B under the same protocol. From LLaVA-1.5 to Qwen3-VL-8B, the three collapse indicators and the CHAIR_S gain decrease together without exception. The largest gains occur under the most severe collapse. On Qwen3-VL-8B, the least collapsed model, the gain is small but its interval excludes zero, and recall is unchanged. LLaVA-1.5 and InstructBLIP hallucinate equally often under greedy decoding, so the difference in their gains does not reflect differing room for improvement.

I.4 Counting Repeated Objects

Repeated instances of one category produce correlated visual keys, so suppressing dominant directions could in principle cause undercounting. We build COCO-Repeat500 from MSCOCO images that contain repeated instances of the same category, caption them with the CHAIR prompt and greedy decoding, and score counts against ground-truth instance counts (Table 19). No counting measure

Table 18: **Collapse severity and gain.** Disp. ratio is the visual-to-text angular dispersion ratio of Table 1; Top-5 conc. is the share of visual key variance in the top five directions; Gate firing is the share of head-steps in layers 9 to 20 with $g > 0.5$.

Model	Disp. ratio	Top-5 conc.	Gate firing	CHAIR_S (greedy $\rightarrow$ WAS)	ΔCHAIR_S	ΔRecall
LLaVA-1.5-7B	$0.52\times$	0.89	6.4%	$54.2 \rightarrow 30.2$	-24.0	-1.8
InstructBLIP-7B	$0.58\times$	0.82	4.9%	$54.0 \rightarrow 36.0$	-18.0	$+0.1$
Qwen3-VL-4B	$0.74\times$	0.74	3.8%	$26.0 \rightarrow 17.8$	-8.2	$+1.0$
Qwen3-VL-8B	$0.84\times$	0.69	2.1%	$21.0 \rightarrow 18.8$	-2.2 $[-4.0, -0.4]$	$+0.2$ $[-1.0, +1.4]$

changes significantly, the undercount rate does not rise and the signed error does not become more negative, while the hallucinated-instance rate falls by 7.5 points. On MME Count, WAS scores 160.0 against 158.3 for greedy decoding.

Table 19: **Counting on COCO-Repeat500** (LLaVA-1.5-7B).

Metric	Greedy	WAS	Difference [95% CI]
Mean absolute count error↓	1.230	1.180	-0.050 $[-0.130, +0.030]$
Signed count error	-0.42	-0.40	$+0.02$ $[-0.08, +0.12]$
Undercount rate↓	44.7%	44.2%	–
Exact count accuracy↑	38.2%	39.0%	$+0.8$ $[-2.4, +4.0]$
Within-one accuracy↑	70.8%	72.0%	$+1.2$ $[-2.0, +4.4]$
Repeated-category recall↑	70.4%	69.8%	-0.6 $[-2.6, +1.4]$
Hallucinated-instance rate↓	18.1%	10.6%	-7.5 $[-10.3, -4.7]$

J Attention Map Visualization

We visualize attention maps before and after WAS as qualitative evidence that the correction restores spatially selective grounding. For each example, we select a head in layers 9 to 20 with $\bar{H}_{\text{vis}} > 0.9$ at the decoding step where greedy decoding generates a hallucinated object. Its visual attention weights are reshaped to the 24×24 ViT-L/14 patch grid and upsampled to image resolution by bilinear interpolation.

Figures 7 and 8 show ten MSCOCO validation examples on LLaVA-1.5-7B. The original maps (centre column) are diffuse and provide little grounding for the generated token. After WAS (right column), attention concentrates on relevant regions: salient objects, such as the elephant and the person in row 1 of Figure 7 and the cat in row 1 of Figure 8; scene structures, such as the train at the platform in row 2 of Figure 7 and the horses and trailer in row 2 of Figure 8; and small details, such as the cell phone in row 4 and the birds on the water in row 5 of Figure 7. The same contrast holds across diverse scene types, consistent with WAS correcting the geometry identified in Section 3 rather than introducing content-specific biases.

K Benchmark Evaluation Metrics

We provide formal definitions of the evaluation metrics used across all four benchmarks. All metrics are computed per-sample and averaged over the evaluation set unless otherwise noted.

K.1 CHAIR

CHAIR (Rohrbach et al., 2018) (Caption Hallucination Assessment with Image Relevance) quantifies object hallucination by comparing mentioned objects in generated captions against ground-truth annotations from MSCOCO (Lin et al., 2014). Let $\mathcal{C} = \{c_1, \ldots, c_M\}$ denote the generated sentences, $\mathcal{O}_i$ the objects mentioned in sentence c_i (extracted via noun chunking and mapped to MSCOCO categories), and $\mathcal{G}$ the ground-truth objects. An object $o \in \mathcal{O}_i$ is *hallucinated* if $o \notin \mathcal{G}$. CHAIR defines two granularities:

$$\text{CHAIR}_S = \frac{\left|\{\, c_i \in \mathcal{C} \mid \exists\, o \in \mathcal{O}_i,\ o \notin \mathcal{G} \,\}\right|}{|\mathcal{C}|}, \qquad \text{CHAIR}_I = \frac{\sum_{i=1}^{M} \left|\{\, o \in \mathcal{O}_i \mid o \notin \mathcal{G} \,\}\right|}{\sum_{i=1}^{M} |\mathcal{O}_i|}, \tag{24}$$

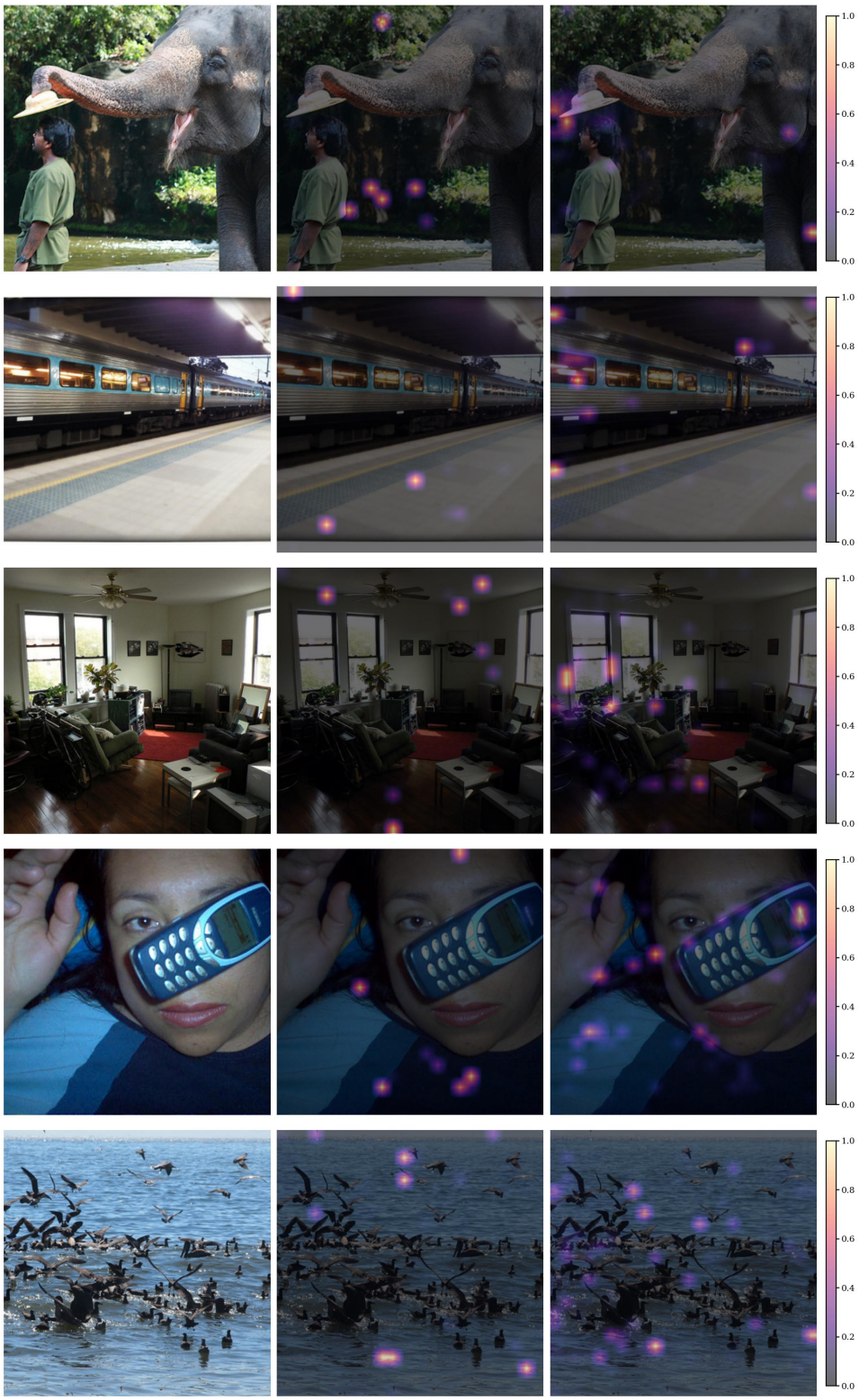


Figure 7: **Attention maps before and after WAS** on five MSCOCO examples (LLaVA-1.5-7B). Left: input image. Centre: original attention. Right: attention after WAS. Warmer colors indicate higher weight; all maps share the same color scale.

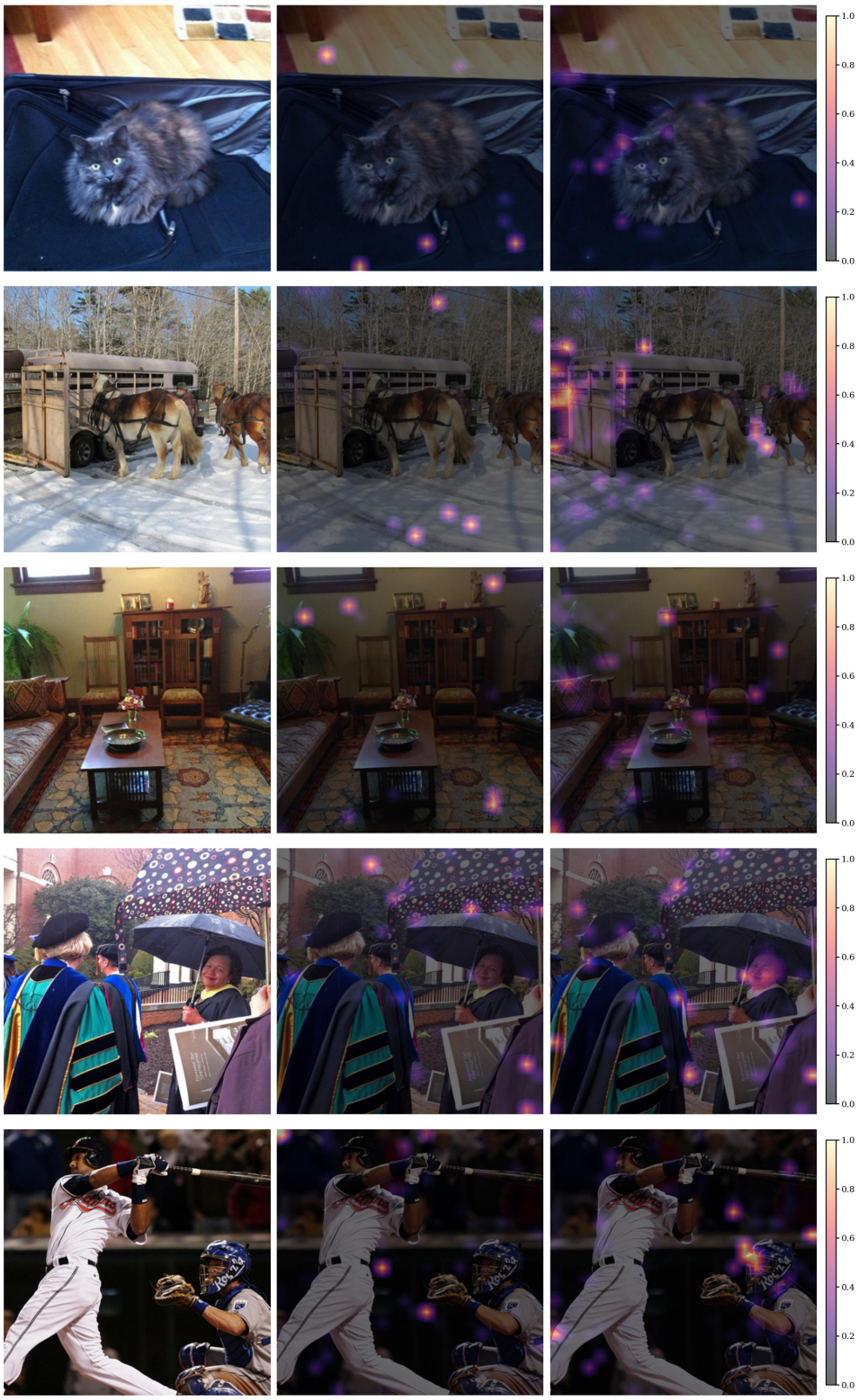


Figure 8: **Attention maps before and after WAS** (continued). Five additional MSCOCO examples with the same setup as Figure 7.

measuring sentence-level and instance-level hallucination rates respectively. Recall quantifies object coverage:

$$\text{Recall} \;=\; \frac{\left|\left(\bigcup_{i=1}^{M} \mathcal{O}_i\right) \,\cap\, \mathcal{G}\right|}{|\mathcal{G}|}\,. \tag{25}$$

Composite Faithfulness–Awareness Rate (CFAR). CHAIR_S and CHAIR_I penalise hallucination but are blind to recall: a model that suppresses all object mentions achieves $\text{CHAIR} = 0$ trivially. Analogous to how F1 balances precision and recall into a single score, we define CFAR to capture the three-way trade-off among object-level precision, recall, and sentence-level faithfulness:

$$\text{CFAR} \;=\; F_{\text{obj}} \,\cdot\, F_s^{\kappa} \,\times\, 100\,, \qquad F_{\text{obj}} \;=\; \frac{2PR}{P+R}\,, \tag{26}$$

where $P = 1 - \text{CHAIR}_I/100$ is the object-level precision, $R = \text{Recall}/100$, $F_s = 1 - \text{CHAIR}_S/100$ is the sentence-level faithfulness, and κ controls the relative weight of sentence-level faithfulness. The F_{obj} term mirrors F1 in treating precision and recall symmetrically; the F_s^{κ} term applies a multiplicative penalty for sentence-level hallucination. When κ is large, CFAR heavily penalises any sentence containing a hallucinated object, favouring conservative generation; when κ is small, CFAR emphasises the precision-recall balance and is more tolerant of occasional sentence-level errors. We set $\kappa = 0.5$ throughout, which provides a balanced regime where neither faithfulness nor coverage dominates the score.

K.2 POPE

POPE (Li et al., 2023) (Polling-based Object Probing Evaluation) converts hallucination assessment into a binary classification task. The model is asked "Is there a {object} in this image?" and must answer "yes" or "no". Questions are constructed under three increasingly challenging sampling strategies:

- **Random**: negative objects are sampled uniformly from the MSCOCO vocabulary, creating easy negatives.
- **Popular**: negative objects are sampled from the most frequently occurring objects in MSCOCO, testing robustness against dataset frequency bias.
- **Adversarial**: negative objects are sampled from those that most frequently co-occur with the ground-truth objects in the image, directly probing susceptibility to co-occurrence hallucination.

Each split contains 3,000 balanced yes/no questions (500 images $\times$ 6 questions per image). We report four standard classification metrics. Let TP, FP, TN, FN denote true positives, false positives, true negatives, and false negatives respectively, where a "positive" corresponds to a "yes" answer:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}\,, \tag{27}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}\,, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,, \tag{28}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\,. \tag{29}$$

A model biased toward "yes" achieves high recall but low precision (many false positives); a model biased toward "no" shows the reverse. The F1 score captures the trade-off between the two failure modes.

K.3 AMBER

AMBER (Wang et al., 2023) (An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation) provides a multi-dimensional hallucination assessment spanning both generative and discriminative tasks. It contains 1,004 images sourced from held-out test sets to avoid train-test contamination, with annotations covering object existence, attributes (state, number, action), relations, and cognitively plausible hallucination targets.

K.3.1 Generative Task Metrics

The model is prompted with "Describe this image." and the response R is parsed to extract mentioned objects R_{obj}. These are filtered against an annotated object vocabulary X_{obj} to yield $R'_{\text{obj}} = R_{\text{obj}} \cap X_{\text{obj}}$. Let A_{obj} denote the ground-truth objects annotated for the image and H_{obj} the set of cognitively plausible hallucination targets (objects absent from the image but likely to be imagined by the model given scene context).

CHAIR. The proportion of mentioned objects that are hallucinated:

$$\text{CHAIR}(R) \;=\; 1 - \frac{|R'_{\text{obj}} \cap A_{\text{obj}}|}{|R'_{\text{obj}}|}\,. \tag{30}$$

Cover. Object coverage measuring descriptive completeness:

$$\text{Cover}(R) \;=\; \frac{|R'_{\text{obj}} \cap A_{\text{obj}}|}{|A_{\text{obj}}|}\,. \tag{31}$$

Hal. The binary indicator of whether a response contains any hallucination, averaged over all responses:

$$\text{Hal}(R) \;=\; \begin{cases} 1 & \text{if CHAIR}(R) \neq 0\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{32}$$

Cog. The fraction of extracted objects belonging to the cognitively plausible hallucination set, measuring how much the model's errors resemble human cognitive biases:

$$\text{Cog}(R) \;=\; \frac{|R'_{\text{obj}} \cap H_{\text{obj}}|}{|R'_{\text{obj}}|}\,. \tag{33}$$

K.3.2 Discriminative Task Metrics

The model is asked targeted yes/no questions probing specific hallucination dimensions:

- **Existence**: "Is there a {object} in this image?"
- **Attribute—State**: "Is the {object} {state} in this image?"
- **Attribute—Number**: "Are there {number} {object} in this image?"
- **Attribute—Action**: "Does the {object} {action} in this image?"
- **Relation**: "Is there direct contact between {object 1} and {object 2} in this image?"

Standard classification metrics (accuracy, precision, recall, F1) are computed for each category, where precision and recall are defined with respect to the hallucinatory (ground-truth "no") class:

$$\text{Precision}_{\text{hal}} \;=\; \frac{\text{TN}}{\text{TN} + \text{FN}}\,, \qquad \text{Recall}_{\text{hal}} \;=\; \frac{\text{TN}}{\text{TN} + \text{FP}}\,, \tag{34}$$

where TN counts correct rejections of hallucinated content and FP counts cases where the model incorrectly affirms hallucinated content. This convention ensures that higher precision and recall correspond to better hallucination detection.

AMBER Score. A composite metric integrating generative and discriminative performance:

$$\text{AMBER Score} \;=\; \tfrac{1}{2}\big(1 - \text{CHAIR} + \text{F1}\big)\,. \tag{35}$$

K.4 MME

MME (Fu et al., 2023) (Multimodal Large Language Model Evaluation) is a comprehensive benchmark assessing general multimodal capabilities across 14 subtasks, grouped into perception and cognition categories.

Perception subtasks (10 total) evaluate visual understanding: existence, count, position, colour, OCR, poster, celebrity, scene, landmark, and artwork. Each subtask contains image-question pairs with binary yes/no answers.

Cognition subtasks (4 total) evaluate higher-order reasoning: commonsense reasoning, numerical calculation, text translation, and code reasoning.

For each subtask, the score is computed as:

$$\text{Score}_{\text{subtask}} \;=\; \text{Acc}_{\text{yes}} + \text{Acc}_{\text{no}}\,, \tag{36}$$

where Acc_{yes} and Acc_{no} are the accuracies on positive and negative samples respectively. Each subtask score has a maximum of 200. The aggregate scores are:

$$\text{Perception} \;=\; \sum_{i=1}^{10} \text{Score}_i\,, \qquad \text{Cognition} \;=\; \sum_{j=1}^{4} \text{Score}_j\,, \qquad \text{Total} \;=\; \text{Perception} + \text{Cognition}\,. \tag{37}$$

This design prevents models from achieving high scores by indiscriminately answering "yes" or "no", as both positive and negative accuracy contribute equally. We report all three aggregate scores (perception, cognition, total) to verify that hallucination correction does not degrade general multimodal reasoning.

L Extended Related Work

Representation anisotropy. Contextual embeddings in language models occupy a narrow cone (Ethayarajh, 2019), which degrades generation quality (Gao et al., 2019), and whitening corrects it for sentence embeddings (Su et al., 2021; Huang et al., 2021). Related collapse phenomena appear in mixture-of-experts routing (Chi et al., 2022) and in the terminal phase of classifier training (Papyan et al., 2020). In vision-language models, image and text embeddings occupy separate regions of the shared space (Liang et al., 2022), spectral analyses of contrastive learning explain concentration along a few eigenvectors (HaoChen et al., 2021; Saunshi et al., 2022), and vision transformers contain artifact tokens (Darcet et al., 2023). Neighbouring image tokens are redundant enough to be merged (Ren et al., 2023).

Position of this work. We identify an anisotropy specific to visual keys inside the attention heads of LVLMs, relate it to a loss of selection rather than of allocation, and correct it inside live attention during decoding. Unlike sentence-level whitening, WAS whitens only the dominant subspace and restores the score statistics that set allocation. Other recent work edits visual token representations in latent space (Wang et al., 2026) or uses multimodal grounding to correct outputs (Wu et al., 2026); WAS instead acts on the geometry of the keys that determines where attention goes.

M Limitations

The matched intervention controls are carried out on LLaVA-1.5; for the other models we rely on the correspondence between collapse severity and gain (Appendix I.3). WAS uses one layer window and one threshold for all models, and its gain becomes small when collapse is mild. We evaluate models up to 8B parameters. Appendix E proposes causes of collapse during training without testing them.

N Broader Impacts

Reducing hallucination makes vision-language models more reliable in applications such as accessibility and document understanding, and WAS can be added to deployed models without retraining. It reduces but does not eliminate hallucination, so human verification remains necessary in high-stakes settings.

O CASE STUDIES

Figure 9 shows one CHAIR and one AMBER example on LLaVA-1.5-7B with greedy decoding. In the CHAIR example, greedy decoding adds a dining table, knives and a bowl, objects that commonly co-occur with food; WAS removes them and describes the melted cheese and toasted bread. In the AMBER example, greedy decoding adds a book and a cell phone with specific positions; WAS removes both and describes the cartoon character on the mug.

P INTERVENTION CONTROLS AND DOSE RESPONSE

We test whether changing the geometry of the visual keys changes hallucination on LLaVA-1.5-7B (CHAIR, 500 MSCOCO images, greedy decoding). A strength λ multiplies the spectral correction of Eq. equation 5: $\lambda = 1$ is WAS, $\lambda = 0$ is greedy decoding, and $\lambda < 0$ amplifies the dominant directions and deepens the collapse. Score matching and the entropy gate are kept in all settings.

Allocation is fixed while selection changes. WAS leaves the total attention on the image unchanged, from 0.386 to 0.381 with difference interval $[-0.012, +0.002]$, and raises the query dependence Q of Eq. equation 9 from 0.074 to 0.118. Any change in hallucination therefore comes from where attention goes.

Dose response. As λ increases from -1 to 1, CHAIR_S and CHAIR_I decrease monotonically while Q increases monotonically (Table 20). At $\lambda = -1$, hallucination exceeds that of the unmodified model by 5.4 CHAIR_S points $[+1.4, +9.6]$.

Matched controls. Each control matches WAS in rank and correction scale: r random orthonormal directions (five seeds), the r weakest directions, the same correction applied to heads whose entropy lies below the gate threshold, and score matching with the gate but without the spectral correction. None recovers more than two fifths of the reduction achieved by WAS (Table 21), and every paired comparison with WAS survives Holm correction. WAS raises precision from 81.5 to 89.8 while recall changes by less than two points, so it removes absent objects rather than shortening descriptions.

Table 20: **Dose response** on LLaVA-1.5-7B (CHAIR). λ scales the spectral correction; $\lambda < 0$ deepens the collapse.

Strength λ	-1.00	-0.50	0.00	0.25	0.50	0.75	1.00
$\text{CHAIR}_S \downarrow$	59.6	57.0	54.2	45.8	38.4	32.6	**30.2**
$\text{CHAIR}_I \downarrow$	20.2	19.4	18.5	15.9	13.3	11.1	**10.2**
Query dependence $Q \uparrow$	0.060	0.067	0.074	0.085	0.097	0.109	**0.118**

Table 21: **Matched intervention controls** on LLaVA-1.5-7B (CHAIR). Precision is $100 - \text{CHAIR}_I$; intervals are paired bootstrap 95% intervals.

Configuration	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall↑	Precision↑	ΔCHAIR_S [95% CI]
Greedy	54.2	18.5	72.5	81.5	–
Reverse WAS ($\lambda = -1$)	59.6	20.2	71.3	79.8	$+5.4$ $[+1.4, +9.6]$
Random directions (5 seeds)	49.6	16.7	71.8	83.3	-4.6 $[-8.2, -1.0]$
Weakest directions	51.0	17.1	71.9	82.9	-3.2 $[-6.8, +0.4]$
Non-collapsed heads	47.8	15.9	71.4	84.1	-6.4 $[-10.2, -2.6]$
Sharpening only	44.8	14.7	70.9	85.3	-9.4 $[-13.2, -5.6]$
WAS ($\lambda = 1$)	**30.2**	**10.2**	70.7	**89.8**	$\mathbf{-24.0}$ $[-28.4, -19.6]$

CHAIR

Visual Input:



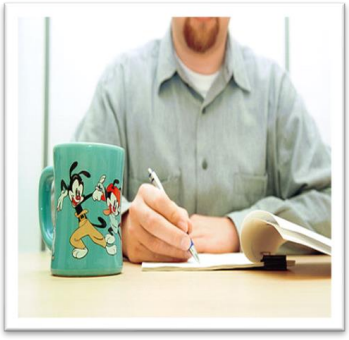
Query:
Describe this photo in detail.

Regular: The image displays two pieces of deli sandwiches prepared and placed on trays, likely aluminum foil, in a kitchen. Each sandwich is cut in half, making a total of four pieces of the delicious sandwiches. The trays are placed on **a dining table** that occupies the lower part of the kitchen. Various cutlery items can be seen around the dining area, with **a couple of knives** and two spoons located in different areas of the table. There is also **a bowl** on the right side of **the table**, contributing to the busy atmosphere of the kitchen.

WAS: The image features a delicious-looking sandwich cut in half and placed on a piece of aluminum foil. The sandwich appears to be a grilled cheese sandwich, with the cheese melted and the bread toasted. The sandwich is placed on a silver tray, which is resting on a countertop. In the background, there is a spoon visible, likely used for eating the sandwich. The scene suggests a casual and enjoyable meal, with the sandwich being a popular choice for a quick and satisfying snack.

AMBER

Visual Input:



Query:
Describe this image.

Regular: The image features **a man** sitting at **a desk**, writing on a piece of paper with **a pen**. He is wearing a gray shirt and appears to be focused on his task. The desk is cluttered with various items, including **a cup**, **a book**, and **a cell phone**. The cup is placed near the center of the desk, while **the book** is located towards the right side. **The cell phone** is situated on the right side of the desk, closer to the edge.

WAS: The image features **a man** sitting at **a desk**, writing on a piece of paper. He is holding **a pen** in his hand, focused on his task. The man is wearing a gray shirt and appears to be in a comfortable environment. On the desk, there is **a cup** with a cartoon character on it, likely a mug. The cup is placed near the man, possibly for a beverage while he works. The scene suggests a casual and relaxed atmosphere, with the man enjoying his writing session.

Figure 9: **Case studies on CHAIR (top) and AMBER (bottom)** (LLaVA-1.5-7B, greedy decoding). Red: hallucinated objects. Blue: correctly grounded objects.