WHEN LANGUAGE OVERWRITES VISION: OVER-ALIGNMENT AND GEOMETRIC DEBIASING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) increasingly power high-stakes applications, from medical imaging to autonomous systems, yet they routinely hallucinate, confidently describing content not present in the input. We investigate the root causes of these failure modes with a mechanistic analysis focusing on the decoder-based VLMs. We trace these failure modes to a geometric over-alignment: to bridge the modality gap required by attention mechanisms, decoder-based VLMs over-align visual embeddings with the text manifold, injecting a statistical linguistic bias that systematically overshadows fine-grained visual evidence. While prior work either aggressively closes this gap or suppresses hallucinations through expensive black-box decoding strategies, none addresses the underlying geometric cause. We provide the first quantitative characterization of this over-alignment, demonstrating that linguistic bias concentrates in the top principal components of a universal, dataset-agnostic text subspace. Building on this insight, we propose two complementary remedies: a training-free inference strategy and a bias-aware fine-tuning paradigm, both of which explicitly project out this subspace from visual representations. Our methods significantly reduce hallucinations across POPE, CHAIR, and AMBER benchmarks, and improve CLAIR scores on long-form captioning tasks, with the training-free variant adding no computational overhead over the base model.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable performance across a wide variety of multi-modal tasks, powering applications from medical imaging to autonomous systems (Bai et al., 2025; Liu et al., 2023; Zhu et al., 2025). Standard decoder-based architectures typically consist of a vision encoder, a cross-modal projector, and a large language model (LLM) decoder. A central challenge reported in training and inference of these models is the *modality gap* (Liang et al., 2022; Yu et al., 2026): visual and linguistic representations naturally occupy disjoint manifolds within the shared latent space. This geometric misalignment has been closely linked to critical failure modes in VLMs, most notably *object hallucination* — the confident generation of visual content that is simply not present in the image (Rohrbach et al., 2018).

Prior work has explored the modality gap primarily in the context of CLIP-based dual-encoder architectures. *Mind the Gap* (Liang et al., 2022) provided an early characterization of this geometric separation, while subsequent approaches such as AlignVLM (Masry et al., 2025) and related projection techniques (Yu et al., 2026) have attempted to close it by aggressively compressing visual representations into the text manifold. In decoder-based VLMs, bridging this gap is well motivated, since the attention mechanism requires both modalities to share a consistent representation space to compute ideal cross-modal similarity scores. However, whether closing this gap is the right strategy remains an open question (Dhimoïla et al., 2026).

Parallel to efforts on the modality gap, a line of work has focused on mitigating hallucinations at the decoding stage, without directly addressing the modality gap. Methods such as VCD, SID, and CRoPS (Leng et al., 2024; Huo et al., 2024; Anand et al., 2026) employ contrastive or introspective decoding strategies to suppress hallucinated outputs. While effective to varying degrees, these

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.

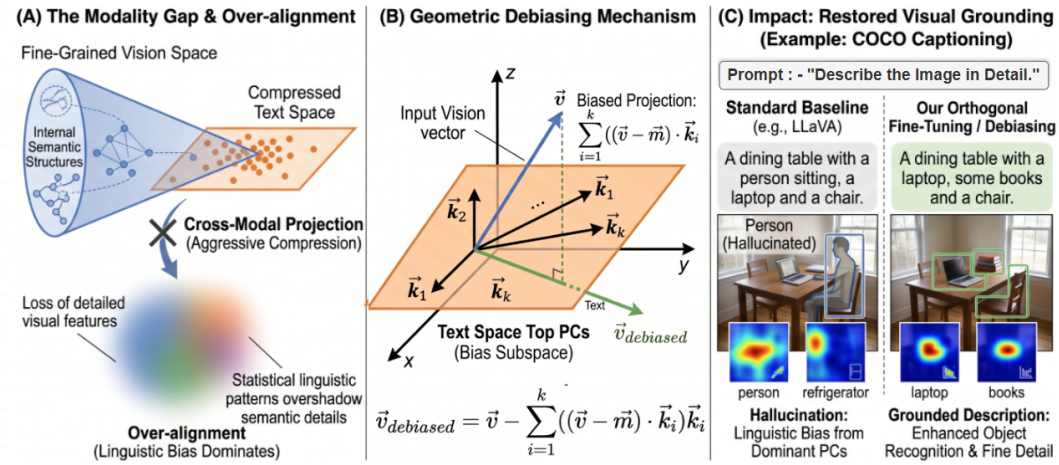


Figure 1: Overview of our geometric debiasing framework. (A) We identify that over-alignment with the text manifold suppresses visual details. (B) We propose a projection-based geometric intervention to isolate and remove statistical linguistic bias. (C) This method unmasks fine-grained visual evidence, directly reducing hallucinations and improving grounding accuracy.

approaches operate as computationally expensive black boxes and treat hallucination as a decoding problem rather than a representational one. Dhimoïla et al. (2026) employs Sparse Autoencoders (SAEs) to show that unimodal atoms explain the modality gap and act as modality-specific biases, but it requires computationally intensive dictionary learning and is validated on dual-encoder models.

Our mechanistic analysis suggests that excessive alignment introduces textual regularities into visual embeddings, obscuring fine-grained visual evidence. Using PCA of pretraining-caption token embeddings, we characterize a text-derived subspace and study its relationship to visual decodability and hallucination. We then remove its leading components from visual embeddings through two complementary interventions: training-free inference and bias-aware projector fine-tuning. Experiments across Qwen and LLaVA models show improvements on hallucination benchmarks and long-form captioning. These findings motivate selective removal of text-associated bias rather than indiscriminately forcing the visual and textual distributions to overlap..

Our main contributions are summarized as follows:

- **Mechanistic analysis.** We characterize over-alignment using a text-derived PCA basis, cross-dataset subspace comparisons, and layer-wise visual probes.
- **Geometric intervention.** We remove leading text-PC components from visual embeddings through training-free inference and bias-aware projector fine-tuning.
- **Empirical evaluation.** We evaluate hallucination, general capabilities, and long-form captioning, and analyze intervention rank, location, and subspace controls.

2 Related Works

2.1 Vision-Language Models and Modality Gap

The rapid advancement of Large Language Models (LLMs) has enabled highly capable Vision-Language Models (VLMs), with diverse applications spanning visual question answering (Agrawal et al., 2016; Tang & Dong, 2024), dataset analysis (Luo et al., 2024), and agentic systems (Zou et al., 2026). These models typically couple a visual encoder with a pretrained language model backbone via a cross-modal projector (Liu et al., 2023; Bai et al., 2025; Zhu et al., 2025). As researchers investigate the training and inference pipelines of these VLMs (Luo et al., 2023; Wadekar et al., 2024), a central challenge emerges in the form of the Modality Gap (Liang et al., 2022; Yu et al., 2026): the persistent geometric separation between vision and language embeddings within a shared latent space. This geometric misalignment has been shown to manifest in critical failure modes, most notably object hallucination (Rohrbach et al., 2018), degraded visual grounding (Li et al., 2023), and an over-reliance on statistical textual priors at the expense of true visual evidence (Leng et al., 2024). While approaches such as AlignVLM (Masry et al., 2025) attempt to close this

gap by enforcing stricter cross-modal alignment, we argue that such forced collapse constitutes over-alignment, introducing a statistical textual bias that actively degrades the semantic integrity of visual representations. One line of work attempts to characterize and quantify linguistic bias directly within multimodal models (Yang et al., 2025; Wang et al., 2025; Chen et al., 2025; Liu et al., 2025b); however, these metrics and diagnostic frameworks remain largely empirical or model-dependent, measuring downstream behavioral discrepancies rather than isolating the foundational representational shortcut within the latent geometry. In parallel, post-hoc decoding interventions such as VCD (Leng et al., 2024), SID (Huo et al., 2024), and CRoPS (Anand et al., 2026) suppress hallucinated tokens during generation, but treat the model as a black box and leave the underlying representation space contaminated. our work instead attempts to eliminate this bias directly at the representation level.

2.2 Mechanistic Interpretability

Mechanistic interpretability seeks to reverse-engineer the internal computations of neural networks by identifying the geometric and algebraic structure of their representations (Sharkey et al., 2025; Zhao et al., 2026). A central organizing principle is the Linear Representation Hypothesis (Elhage et al., 2022), which posits that concepts are encoded as linear directions in activation space, motivating the use of Sparse Autoencoders (SAEs) to decompose representations into interpretable features (Bricken et al., 2023; Cunningham et al., 2023). Within VLMs specifically, this lens has illuminated how visual features propagate through transformer layers (Neo et al., 2024; Liu et al., 2025a), characterized layer-wise token redundancy (Chen et al., 2024), and enabled systematic analysis of generative model outputs (Tang et al., 2025a; 2026a). Beyond analysis, sparse autoencoders have been extended to diverse applications including medical imaging (Abdulaal et al., 2024; Tang et al., 2025b), prompt engineering (Saini et al., 2026), and model editing (Wu et al., 2025). Complementing these empirical advances, rigorous theoretical frameworks have been proposed to characterize the foundations and limitations of sparse dictionary learning (Tang et al., 2026b; Cui et al., 2026). Our work draws on this interpretability tradition to treat cross-modal textual bias as a linearly isolable subspace, enabling principled geometric debiasing of vision embeddings in decoder-based VLMs.

3 Mechanistic Analysis

We examine the geometry, layer-wise evolution, and behavioral consequences of text alignment in visual representations.

3.1 The Mathematical Necessity of Alignment

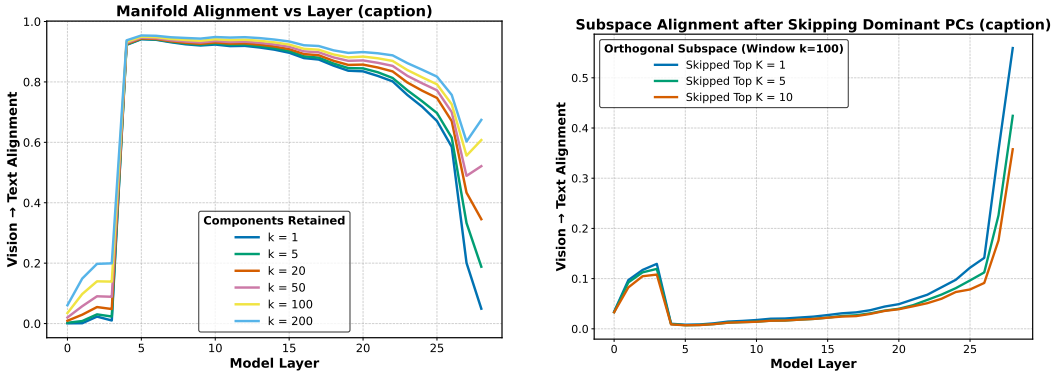
In decoder-based VLMs, bridging the modality gap is theoretically required to support the underlying attention mechanisms. Given a query matrix Q, key matrix K, and value matrix V, scaled dot-product attention is defined as:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

For the dot product $Q \cdot K^T$ to yield mathematically meaningful similarity scores between visual queries and textual keys (or vice versa), both modalities must operate within a shared, dimensionally consistent coordinate system. Thus, some degree of cross-modal alignment is unavoidable if we want to use Decoder-based LLM architecture.

3.2 Defining the Text Manifold and Universal Bias

To investigate the geometry of this shared space, we must first rigorously define the text manifold. We find principal components (PCs) after isolating a representative set of captions directly from the pretraining dataset and taking vector embeddings of them. We compute the top Principal Components from these token embeddings, which allows us to represent the baseline linguistic distribution and cleanly isolate the core textual bias. To verify the stability of this subspace, we compute and plot the Frobenius norm between text manifolds generated from multiple distinct text datasets in 3a. Our analysis reveals that the top PCs form a consistently common manifold across all evaluated datasets. This suggests that standard pretraining injects a universal statistical linguistic bias rather than task-specific semantic structures.



(a) Full vision embeddings

(b) Top k PCs removed

Figure 2: Layer-wise alignment scores of vision tokens onto the text manifold. (a) The alignment trajectory of the full vision embeddings demonstrates that the network aggressively forces the vision tokens to overlap with the textual space in the early to middle layers. (b) The alignment trajectory after removing the top k principal components. The significant drop in alignment score confirms that the statistical textual bias is heavily concentrated in these top directions, leaving the lower PCs to occupy an orthogonal, modality-agnostic subspace.

3.3 Layer-wise Alignment Trajectory

To understand where and how the textual bias penetrates the vision embeddings, we analyze the layer-wise alignment of vision tokens within the LLM decoder. Building on our defined text manifold, we first center the vision embedding $\vec{v}^{(l)}$ at layer l by subtracting the mean of the text manifold $\vec{m}$. We then project it onto the text-induced PCA basis $\{\vec{k}_1, ..., \vec{k}_K\}$ to obtain the text-aligned representation $\vec{z}^{(l)}$:

$$\vec{z}^{(l)} = \sum_i \left((\vec{v}^{(l)} - \vec{m}) \cdot \vec{k}_i \right) \vec{k}_i$$

The alignment score is then defined as the ratio of the projected norm to the original centered norm:

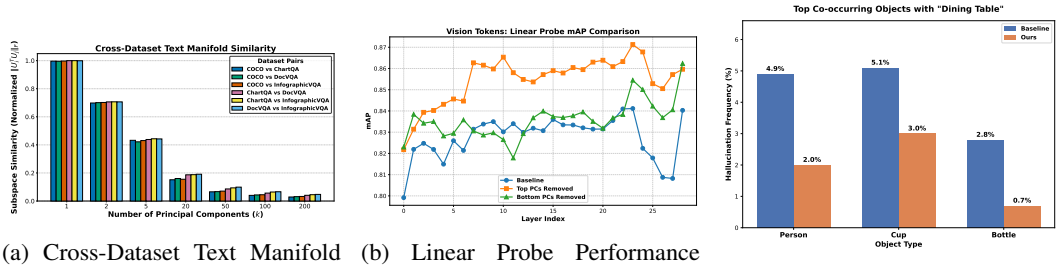
$$Align(l) = \frac{1}{M} \sum_{j=1}^{M} \frac{||\vec{z}_j^{(l)}||_2}{||\vec{v}_j^{(l)} - \vec{m}||_2}$$

By analyzing the ratio of this projected norm to the original centered norm, we track the alignment trajectory. As shown in Figure 2a, the full vision embeddings exhibit an aggressively high alignment score in the early to middle layers, confirming that the network forces the vision tokens to heavily overlap with the text manifold. However, when we restrict this projection to manifolds defined without the top principal components (Figure 2b), the alignment scores drop significantly. This confirms our geometric hypothesis: the top PCs are hijacked to satisfy the alignment objective.

3.4 Over-Alignment and the Overshadowing of Visual Details

While Large Language Models rely on textual biases as essential priors to structure generation, enforcing these dataset biases onto vision embeddings overrides true visual evidence and acts as a structural distraction. We argue that the current pretraining paradigm suffers from over-alignment. To validate this, we conduct token-level linear probing for multi-label visual classification across all layers of the LLM decoder (Figure 3b). By extracting latent vision representations to train a regularized Ridge classifier on 80 MS-COCO object categories, we aggregate token-level logits via max-pooling to calculate image-level mean Average Precision (mAP). This setup measures precisely how linearly decodable the raw visual signal remains as it propagates through the network.

Throughout the decoder layers, the baseline model (blue) consistently yields the lowest decodability. When we perform a spectral intervention to project the vision tokens orthogonally to the top textual PCs (orange), linear decodability is significantly enhanced across the entire depth of the network. This continuous performance gap suggests that the top PCs act as a structural mask overriding the underlying visual data. Ultimately, these results demonstrate that standard projection forces visual details to compete with high-variance linguistic priors; by removing the top PCs, we strip away



(a) Cross-Dataset Text Manifold Similarity

(b) Linear Probe Performance (mAP)

(c) Frequency-Driven Hallucination

Figure 3: Geometric stability of textual bias and its impact on visual decodability. **(a)** Subspace similarity (Normalized $\|U_i^T U_j\|_F$) across diverse datasets. The near-perfect overlap in the top PCs confirms a universal linguistic bias, while lower PCs capture dataset-specific semantics. **(b)** Token-level linear probe performance across decoder layers. The baseline model (blue) degrades in middle-to-late layers due to structural text alignment. Projecting orthogonally to the top text PCs (orange) rescues and preserves fine-grained visual details. **(c)** Hallucination frequency of objects frequently co-occurring with "dining table." The baseline model defaults to textual priors, while our method forces visual grounding and reduces hallucination rates by over 50% on average.

textual overshadowing and continuously unmask the localized visual details necessary for accurate semantic grounding.

3.5 Geometrically Induced Hallucinations

When dataset bias is left unchecked, it acts as a strong gravitational pull within the embedding space. Because the representations are skewed toward textual priors, the VLM undergoes visual overwrite, defaulting to high-probability text patterns instead of visually attending to the actual image. As illustrated in our frequency dependency plots, the model confidently predicts objects that frequently co-occur in text, ignoring their absence in reality. To empirically validate this visual overwrite, we analyze the hallucination frequency of objects that frequently co-occur in the pretraining corpus. We isolate image subsets containing a base object (e.g., "dining table") but strictly lacking common co-occurring objects (e.g., "person," "cup," or "bottle"). As illustrated in Figure 3c, the baseline model frequently hallucinates these missing objects, driven by their statistical proximity in the text manifold as mentioned in paper (Leng et al., 2024). By applying our geometric debiasing strategy ablating the top principal components associated with the modality anchor we effectively sever this textual gravitational pull. Consequently, the hallucination rates drop substantially across all categories.

4 Methodology

Having established that over-alignment injects a statistical textual bias that acts as a structural distraction, we propose a geometric framework to extract and eliminate this bias. In this section, we formalize our text-induced coordinate system and detail our two proposed debiasing methods: a training-free inference strategy and a generalized VLM finetuning paradigm.

4.1 Training-Free Debiasing Inference Strategy

Our empirical analysis demonstrates that the top principal components of the vision embeddings encapsulate artifactual statistical bias. Building on the Linear Representation Hypothesis (Elhage et al., 2022), which posits that distinct concepts are encoded as linear directions within a representation space, we treat this textual bias as a linear subspace that can be cleanly isolated. To mitigate statistically induced hallucinations without requiring further training, we propose a post-hoc bias removal strategy.

Given a vision embedding $\vec{v}$ and the mean of the text manifold $\vec{m}$, we first mean-center the vision embedding and compute its projection onto the top k principal components of the text-aligned subspace. We then subtract this projection from the original embedding to obtain the debiased

Table 1: **Training-free hallucination mitigation and general capabilities.** We compare three VLMs on POPE, AMBER, CHAIR, MME, and MMBench. Arrows indicate the preferred direction for each metric. Entries with uncertainties report mean ± standard deviation. Bold values indicate the best result within each model and metric, not statistical significance. The inference-time PCA ranks for our method are selected on the training set as described in Appendix A.1.

Model	Method	POPE (↑)	AMBER			CHAIR			MME (↑)	MMBench (↑)
			CHAIR (↓)	HAL (↓)	Cover (↑)	$CHAIR_i$ (↓)	$CHAIR_s$ (↓)	Recall (↑)		
Qwen2.5-VL-7B	Zero-shot	87.3	5.14 ±0.45	28.62 ±1.6	63.5 ±0.9	10.40 ±0.82	35.24 ±2.1	67.10 ±1.2	2328.1	81.13
	SID	88.1	4.58 ±0.45	25.90 ±1.6	45.7 ±0.7	8.87 ±0.82	30.60 ±2.0	58.50 ±1.5	2342.8	82.15
	VCD	87.6	4.72 ±0.45	26.35 ±1.6	43.1 ±1.0	8.75 ±0.82	33.10 ±2.1	60.95 ±1.7	2335.7	82.55
	VISTA	88.0	4.46 ±0.45	25.79 ±1.6	56.9 ±0.8	8.95 ±0.84	33.00 ±2.1	60.49 ±1.4	2338.1	82.70
	DMAS	88.3	4.40 ±0.45	25.60 ±1.6	60.0 ±0.9	8.90 ±0.82	31.20 ±2.0	65.30 ±1.5	2346.0	83.20
	Nullu	87.5	4.83 ±0.40	26.90 ±1.9	62.9 ±0.9	8.65 ±0.80	32.9 ±2.1	67.3 ±1.2	2334	82.1
	VTI	87.2	4.60 ±0.35	27.10 ±2.0	**64.2 ±0.8**	9.41 ±0.75	33.4 ±2.1	68.4 ±1.3	2339	82.4
	DeCo	87.8	4.78 ±0.36	27.60 ±2.0	62.7 ±0.9	9.03 ±0.63	33.7 ±2.2	**68.6 ±1.2**	2340	82.6
	ICT	87.3	4.90 ±0.40	27.40 ±1.9	63.4 ±0.8	9.58 ±0.80	34.1 ±2.1	66.9 ±1.2	2330	82.0
	Ours	**88.4**	**4.23 ±0.45**	**25.07 ±1.6**	63.0 ±0.9	**8.20 ±0.82**	**29.80 ±2.0**	67.48 ±1.2	**2355.2**	**83.33**
LLaVA-1.5-7B	Zero-shot	82.8	9.20 ±0.45	38.80 ±1.6	48.1 ±0.8	15.80 ±0.82	52.80 ±2.2	77.90 ±1.0	1762.5	62.00
	SID	83.5	**6.24 ±0.32**	33.73 ±1.5	**53.45 ±0.61**	14.20 ±0.80	50.60 ±2.2	77.94 ±0.96	1769.2	63.65
	VCD	83.3	7.20 ±0.39	33.33 ±1.5	51.66 ±0.60	15.42 ±0.86	51.60 ±2.2	77.44 ±1.0	1766.0	63.10
	VISTA	83.7	6.90 ±0.38	31.94 ±1.5	49.93 ±0.61	13.28 ±0.78	46.40 ±2.2	75.74 ±0.99	1760.0	62.70
	DMAS	83.6	7.10 ±0.45	33.50 ±1.6	51.0 ±0.7	14.40 ±0.82	48.80 ±2.2	77.60 ±1.0	1770.4	63.30
	Nullu	83.2	7.30 ±0.45	32.00 ±1.9	49.0 ±0.8	14.90 ±0.85	50.4 ±2.2	78.0 ±1.0	1765	63.1
	VTI	83.0	9.90 ±0.41	34.88 ±1.7	44.06 ±0.77	14.56 ±0.74	50.6 ±2.1	68.18 ±1.0	1758	62.4
	DeCo	83.6	7.04 ±0.60	**26.67 ±2.0**	45.15 ±0.79	**12.31 ±0.85**	**36.6 ±2.2**	68.31 ±1.0	1768	63.4
	ICT	83.4	7.20 ±0.45	33.10 ±1.9	49.2 ±0.8	14.50 ±0.82	50.8 ±2.2	**78.2 ±1.0**	1766	63.2
	Ours	**83.8**	6.74 ±0.45	29.87 ±1.6	50.1 ±0.84	12.99 ±0.75	45.60 ±2.2	75.11 ±1.0	**1771.3**	**63.70**
LLaVA-1.5-13B	Zero-shot	85.9	6.09 ±0.50	28.67 ±2.0	49.28 ±0.84	13.89 ±0.82	47.4 ±2.2	78.51 ±1.0	1804	67.9
	SID	85.3	5.20 ±0.45	26.00 ±1.9	**52.5 ±0.7**	12.30 ±0.82	47.0 ±2.2	77.90 ±1.0	1806	68.2
	VCD	86.0	5.60 ±0.45	27.00 ±1.9	51.5 ±0.7	13.70 ±0.82	49.5 ±2.2	77.70 ±1.0	1802	68.0
	VISTA	86.1	5.10 ±0.45	25.80 ±1.9	50.5 ±0.7	12.40 ±0.82	44.1 ±2.2	75.70 ±1.0	1806	68.1
	DMAS	86.0	5.30 ±0.45	26.50 ±1.9	51.0 ±0.7	13.40 ±0.82	46.4 ±2.2	77.60 ±1.0	1798	68.3
	Nullu	85.7	5.60 ±0.45	27.00 ±1.9	48.80 ±0.83	13.50 ±0.85	45.5 ±2.2	77.69 ±1.0	1800	67.8
	VTI	83.0	5.38 ±0.38	26.68 ±1.7	44.06 ±0.79	12.94 ±0.65	43.2 ±1.9	66.92 ±1.2	1782	66.6
	DeCo	85.5	5.50 ±0.53	27.00 ±2.0	46.14 ±0.83	12.95 ±0.77	41.4 ±2.2	71.83 ±1.0	1796	67.2
	ICT	85.8	5.80 ±0.50	28.00 ±2.1	49.79 ±0.83	13.60 ±0.78	50.5 ±2.2	**79.00 ±1.0**	1802	67.6
	Ours	**86.4**	**4.71 ±0.45**	**21.88 ±1.8**	47.90 ±0.82	**11.89 ±0.79**	**39.2 ±2.2**	76.75 ±1.0	**1808**	**68.4**

representation:

$$\vec{v}_{debiased} = \vec{v} - \sum_{i=1}^{k} ((\vec{v} - \vec{m}) \cdot \vec{k}_i)\vec{k}_i$$

By projecting out the top PCs associated with the training captions, we explicitly remove the textual bias. This forces the VLM to rely on generalized semantic information, yielding embeddings that are crisp, semantic, and highly grounded.

4.2 Bias-Aware Fine-Tuning Paradigm

While our training-free strategy is effective at inference, we can further solidify this geometric debiasing by integrating it directly into the training loop. We propose a generalized VLM fine-tuning paradigm that explicitly prevents textual biases from abstracting visual embeddings during parameter updates. During the fine-tuning forward pass, we apply the identical top-k PC component bias removal to the output of the cross-modal projector before it enters the LLM decoder. By structurally depriving the network of these top principal components during training, we force the VLM to rely on generalized semantic information rather than exploiting statistical textual shortcuts.

5 Results

We evaluate geometric debiasing in two complementary settings: training-free inference and bias-aware fine-tuning. We first measure hallucination and general-capability performance, then assess long-form captioning, inspect patch-level semantic readouts, and analyze the sensitivity and specificity of the subspace intervention.

Table 2: **Long-form captioning with bias-aware fine-tuning.** CLAIR scores (↑) on VRSBench and TextCaps for zero-shot inference, standard supervised fine-tuning (SFT), and our bias-aware fine-tuning. The two fine-tuning configurations use the same training data and budget. Bold indicates the best score for each dataset–model pair.

Dataset	Model	Zero-shot	Standard SFT	Ours
VRSBench	Qwen2.5-VL-7B	63.90	72.30	**74.70**
	LLaVA-1.5-7B	54.10	63.50	**65.35**
TextCaps	Qwen2.5-VL-7B	69.95	77.10	**78.75**
	LLaVA-1.5-7B	56.60	60.30	**63.80**

5.1 Training-Free Bias Removal

We test whether removing the leading directions of the text-induced PCA basis reduces hallucination without requiring model fine-tuning. Our evaluation covers Qwen2.5-VL-7B, LLaVA-1.5-7B, and LLaVA-1.5-13B on POPE (Li et al., 2023), AMBER (Wang et al., 2023), CHAIR (Rohrbach et al., 2018), MME (Fu et al., 2023), and MMBench (Liu et al., 2024). Table 1 compares our method with the zero-shot model and SID (Huo et al., 2024), VCD (Leng et al., 2024), VISTA (Li et al., 2025), DMAS (Yin et al., 2026), Nullu, VTI, DeCo, and ICT.

Selecting the intervention rank. We tune the number of removed top text PCs on the training set itself, selecting $k = 10$ for Qwen2.5-VL-7B, $k = 50$ for LLaVA-1.5-7B, and $k = 150$ for LLaVA-1.5-13B. Appendix A.1 reports the sweep and the associated hallucination–recall trade-off. These ranks apply to training-free inference, not to the separate fine-tuning configuration in Section 4.2.

Hallucination and capability results. Relative to the zero-shot baseline, our method reduces both AMBER hallucination measures and both CHAIR hallucination measures for all three models. On **Qwen2.5-VL-7B**, AMBER CHAIR decreases from 5.14 to 4.23 and HAL from 28.62 to 25.07. CHAIR_i decreases from 10.40 to 8.20 and CHAIR_s from 35.24 to 29.80, while recall remains similar (67.10 to 67.48). POPE, MME, and MMBench also improve.

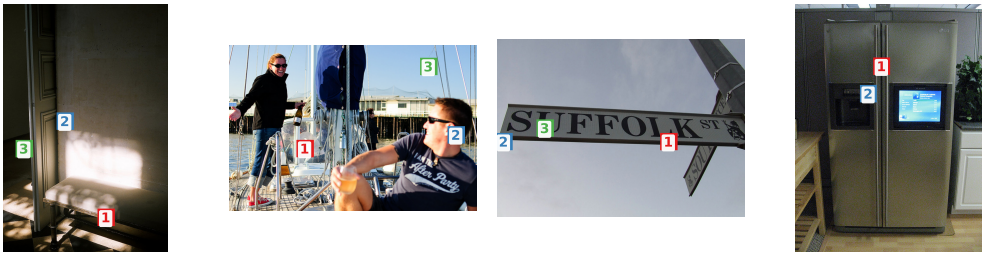
On **LLaVA-1.5-7B**, AMBER CHAIR decreases from 9.20 to 6.74 and HAL from 38.80 to 29.87; CHAIR_i decreases from 15.80 to 12.99 and CHAIR_s from 52.80 to 45.60. These improvements accompany a recall decrease from 77.90 to 75.11. Our method is not uniformly best on every hallucination metric: SID obtains lower AMBER CHAIR, and DeCo obtains lower HAL, CHAIR_i, and CHAIR_s. However, our method retains higher recall than DeCo (75.11 versus 68.31), illustrating the importance of assessing hallucination and coverage jointly.

The pattern extends to **LLaVA-1.5-13B**: CHAIR_i decreases from 13.89 to 11.89 and CHAIR_s from 47.4 to 39.2, while AMBER CHAIR and HAL decrease from 6.09 to 4.71 and from 28.67 to 21.88, respectively. Recall decreases from 78.51 to 76.75, and AMBER coverage decreases from 49.28 to 47.90. Across all three models, our method achieves the highest POPE, MME, and MMBench point estimates among the listed methods. These results support the usefulness of targeted text-subspace removal, while the recall and coverage changes show that the intervention is not lossless in every setting.

5.2 Bias-Aware Fine-Tuning: Long-Form Captioning

We next incorporate the intervention into the fine-tuning forward pass, using the configuration in Section 4.2. We evaluate on **VRSBench** (Li et al., 2024), which requires detailed descriptions of remote-sensing imagery, and **TextCaps** (Sidorov et al., 2020), which requires integrating visible text with visual context. Following the semantic evaluation setting of VRSBench, we report **CLAIR** (Chan et al., 2023) to assess agreement between generated and reference captions beyond surface-level lexical overlap. CLAIR is a reference-based caption-quality metric; its gains complement, rather than replace, direct hallucination measurements.

As shown in Table 2, bias-aware fine-tuning improves over both zero-shot inference and standard SFT for both architectures and both datasets. On VRSBench, Qwen2.5-VL-7B improves from 72.30



(a) (b) (c) (d)

Patch-level logit-lens readouts

Image	Patch	Baseline	Ours
(a)	1	`","`	→ `"bench"`
	2	`"a"`	→ `"doorway"`
	3	`"a"`	→ `"open"`
(b)	1	`","`	→ `"deck"`
	2	`"in"`	→ `"man"`
	3	`","`	→ `"blue"`
(c)	1	`"<|object_ref_start|>"`	→ `"Street"`
	2	`"1"`	→ `"sign"`
	3	`"[empty]"`	→ `"U"`
(d)	1	`"by"`	→ `"refrigerator"`
	2	`"and"`	→ `"water"`

Text-PC token neighbors

Category	Top-PC tokens
Punctuation	".", `U+3002`, "...", ","; newline
Articles	"a", "A", "the", "The", "an"
Prep./Conj.	"of", "with", "on", "and", "for", "to", "from"
Pron./Copula	"you", "your", "his", "are", "am", "is"

Figure 4: **Patch-level semantic readouts and interpretation of the removed text subspace.** Top: images with indexed, color-coded patches. Bottom left: the highest-scoring logit-lens token for each marked patch before and after top-text-PC removal. In these examples, removal shifts the readout from structural tokens toward visually relevant concepts. Bottom right: categories of input-token embeddings nearest to the leading text-PC directions under cosine similarity. These neighbors emphasize punctuation and frequent function words; `U+3002` denotes the ideographic full stop.

with standard SFT to 74.70, and LLaVA-1.5-7B improves from 63.50 to 65.35. On TextCaps, the corresponding improvements are 77.10 to 78.75 and 60.30 to 63.80. Under the shared training-data and training-budget setting, these comparisons support incorporating geometric debiasing into the fine-tuning forward pass to improve semantic caption quality.

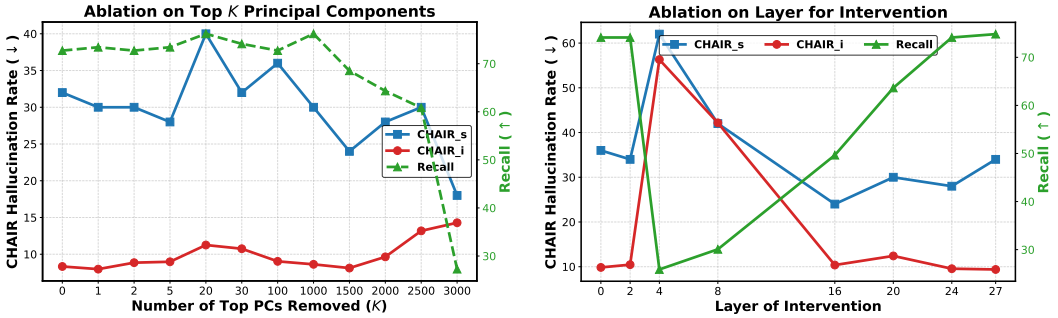
5.3 Logit-Lens Analysis

To inspect the semantic information carried by visual tokens, we apply the logit lens to patch-level hidden states at decoder layers 26–28, reading them out through the language model's unembedding matrix. Figure 4 contrasts the highest-scoring tokens before and after top-text-PC removal. In the illustrated patches, baseline readouts often favor punctuation or function words, whereas the debiased readouts include visually relevant tokens such as `bench`, `deck`, and `refrigerator`. These are diagnostic token readouts, not generated captions.

We additionally inspect the nearest input-token embeddings to the leading text-PC directions using cosine similarity. Their association with punctuation and frequent function words is consistent with a linguistic-structure component in the removed subspace. Together with the patch-level readouts, this provides qualitative support for the textual-bias interpretation, without establishing that every removed direction is exclusively linguistic.

5.4 Ablation Study

Sensitivity to rank and intervention layer. Figure 5 examines Qwen2.5-VL-7B on CHAIR while varying the number of removed top text PCs and the decoder layer at which the intervention is applied. We report CHAIR_i, CHAIR_s, and recall jointly to expose the trade-off between hallucination



(a) Number of removed top text PCs.

(b) Decoder layer of intervention.

Figure 5: **Sensitivity of geometric debiasing on Qwen2.5-VL-7B.** CHAIR results when varying (a) the number of removed top text principal components and (b) the decoder layer of intervention. Lower CHAIR_i and CHAIR_s and higher recall are better. The training-set results used for model-specific rank selection are given in Appendix A.1.

Table 3: **Matched-dimension subspace controls on LLaVA-1.5-7B.** Each removal condition suppresses a $k = 50$ subspace at the projector/merger output; the baseline has no removal. Entries report mean ± standard error. Bold values highlight the deterioration under vision-PCA removal, not the best scores. This experiment uses its own baseline, distinct from Table 1; it does not include a matched top-text-PC-removal condition.

Condition	**CHAIR**			**AMBER**		
	CHAIR_s (↓)	**CHAIR_i (↓)**	**Recall (↑)**	**CHAIR (↓)**	**HAL (↓)**	**Cover (↑)**
Baseline	52.2 ±2.2	16.05 ±.84	76.0 ±1.0	7.76 ±.54	33.3 ±2.1	47.8 ±.8
Random	52.0 ±2.2	15.43 ±.85	76.5 ±1.0	7.25 ±.51	30.5 ±2.0	48.0 ±.8
Bottom	49.0 ±2.2	15.47 ±.86	76.4 ±1.0	6.63 ±.50	30.9 ±2.1	48.9 ±.8
Middle	52.8 ±2.2	15.45 ±.80	77.1 ±1.0	7.42 ±.50	34.9 ±2.1	48.4 ±.9
Vision-PCA	49.6 ±2.2	**57.34 ±2.4**	**16.1 ±.8**	**64.59 ±2.6**	**69.0 ±2.1**	**9.6 ±.5**

reduction and information loss. The per-model training-set sweeps used to select the inference ranks are reported separately in Appendix A.1.

Subspace controls at the merger. To examine whether the choice of removed subspace matters beyond its dimension, Table 3 reports a separate LLaVA-1.5-7B control experiment on CHAIR and AMBER. Each non-baseline condition removes a $k = 50$ subspace at the cross-modal projector/merger output. We compare random, bottom-PC, middle-PC, and vision-PCA removal with the unmodified baseline from this control experiment.

Random and bottom-PC removal produce modest changes, with bottom-PC removal improving several hallucination measures while maintaining recall. Middle-PC removal has mixed effects: CHAIR_s and AMBER HAL worsen, despite slightly higher recall. In contrast, removing the vision-PCA subspace sharply degrades visual grounding: CHAIR_i increases from 16.05 to 57.34, recall falls from 76.0 to 16.1, AMBER CHAIR increases from 7.76 to 64.59, and AMBER coverage falls from 47.8 to 9.6. The lower CHAIR_s value in this condition should therefore not be interpreted in isolation as better grounding.These controls demonstrate that removing an equal-dimensional subspace can have very different effects, and that visual principal directions cannot be treated as interchangeable with the targeted textual subspace.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, Claude, and Claude Code to implement methods, write and edit code, and assist in drafting the manuscript. We did not use generative AI tools to generate synthetic data, formulate or prove mathematical claims, or propose our core hypotheses, methodology, or experimental design; these were carried out by the authors. We have reviewed all AI-assisted work: code was checked against expected behavior and the underlying data, and all AI-drafted text and result interpretations were verified against experimental outputs and revised where necessary. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ahmed Abdulaal, Hugo Fry, Nina Montaña-Brown, Ayodeji Ijishakin, Jack Gao, Stephanie Hyland, Daniel C. Alexander, and Daniel C. Castro. An x-ray is worth 15 features: Sparse autoencoders for interpretable radiology report generation, 2024. URL https://arxiv.org/abs/2410.03334.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. URL https://arxiv.org/abs/1505.00468.

Neeraj Anand, Samyak Jha, Udbhav Bamba, and Rahul Rahaman. Crops: A training-free hallucination mitigation framework for vision-language models. *arXiv preprint arXiv:2601.00659*, 2026.

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.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13638–13646, 2023.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4209–4221. IEEE, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy, 2026. URL https://arxiv.org/abs/2506.15963.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Picard. Cross-modal redundancy and the geometry of vision-language embeddings. *arXiv preprint arXiv:2602.06218*, 2026.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

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.

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.

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.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024.

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

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025.

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.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. *arXiv preprint arXiv:2510.04819*, 2025a.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Yifan Luo, Yiming Tang, Chengfeng Shen, Zhennan Zhou, and Bin Dong. Prompt engineering through the lens of optimal control, 2023. URL `https://arxiv.org/abs/2310.14201`.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pp. 235–252. Springer, 2024.

Ahmed Masry, Juan A Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, et al. Alignvlm: Bridging vision and language latent spaces for multimodal document understanding. *arXiv preprint arXiv:2502.01341*, 2025.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.

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.

Harshvardhan Saini, Yiming Tang, and Dianbo Liu. Bridging mechanistic interpretability and prompt engineering with gradient ascent for interpretable persona control, 2026. URL https://arxiv.org/abs/2601.02896.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pp. 742–758. Springer, 2020.

Yiming Tang and Bin Dong. Demonstration notebook: Finding the most suited in-context learning example from interactions, 2024. URL https://arxiv.org/abs/2406.10878.

Yiming Tang, Abhijeet Sinha, and Dianbo Liu. How does my model fail? automatic identification and interpretation of physical plausibility failure modes with matryoshka transcoders, 2025a. URL https://arxiv.org/abs/2511.10094.

Yiming Tang, Wenjia Zhong, Rushi Shah, and Dianbo Liu. Cxr-lanic: Language-grounded interpretable classifier for chest x-ray diagnosis, 2025b. URL https://arxiv.org/abs/2510.21464.

Yiming Tang, Arash Lagzian, Srinivas Anumasa, Qiran Zou, Yingtao Zhu, Ye Zhang, Trang Nguyen, Yih-Chung Tham, Ehsan Adeli, Ching-Yu Cheng, Yilun Du, and Dianbo Liu. Human-like content analysis for generative ai with language-grounded sparse encoders, 2026a. URL https://arxiv.org/abs/2508.18236.

Yiming Tang, Harshvardhan Saini, Zhaoqian Yao, Zheng Lin, Yizhen Liao, Qianxiao Li, Mengnan Du, and Dianbo Liu. A unified theory of sparse dictionary learning in mechanistic interpretability: Piecewise biconvexity and spurious minima, 2026b. URL https://arxiv.org/abs/2512.05534.

Shakti N. Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. The evolution of multimodal model architectures, 2024. URL https://arxiv.org/abs/2405.17927.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025.

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.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.17148.

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

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2602.21704*, 2026.

Xiaomin Yu, Yi Xin, Wenjie Zhang, Chonghan Liu, Hanzhen Zhao, Xiaoxing Hu, Xinlei Yu, Ziyue Qiao, Hao Tang, Xue Yang, et al. Modality gap-driven subspace alignment training paradigm for multimodal large language models. *arXiv preprint arXiv:2602.07026*, 2026.

Haiyan Zhao, Zirui He, Yiming Tang, Fan Yang, Ali Payani, Dianbo Liu, and Mengnan Du. Rep2text: Decoding full text from a single llm token representation, 2026. URL https://arxiv.org/abs/2511.06571.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Qiran Zou, Hou Hei Lam, Wenhao Zhao, Yiming Tang, Tingting Chen, Samson Yu, Tianyi Zhang, Chang Liu, Xiangyang Ji, and Dianbo Liu. Fml-bench: Benchmarking machine learning agents for scientific research, 2026. URL https://arxiv.org/abs/2510.10472.

A Appendix

A.1 Selecting the Number of Principal Components

The number of removed principal components, k, controls the extent of the training-free intervention: removing too few directions may leave textual bias intact, whereas removing too many may discard information needed for visual grounding. We tune k on the training set itself, rather than on the evaluation set, and examine CHAIR_i, CHAIR_s, and object recall jointly.

Table 4 reports the training-set sweep. Among the evaluated settings, we select $k = 10$ for Qwen2.5-VL-7B, $k = 50$ for LLaVA-1.5-7B, and $k = 150$ for LLaVA-1.5-13B. Each selected setting achieves the lowest CHAIR_i and CHAIR_s for its model while retaining recall close to that of the smaller-k setting. These are empirical choices among the evaluated configurations, not claims of a globally optimal rank.

Table 4: **Training-set selection of the inference-time PCA rank.** We vary the number k of removed top text principal components. Entries report mean $\pm$ standard deviation; lower CHAIR_i and CHAIR_s and higher recall are better. Bold k values mark the selected setting for each model. This sweep is used for hyperparameter selection, not as a held-out evaluation.

Model	k	CHAIR_i (↓)	CHAIR_s (↓)	Recall (↑)
Qwen2.5-VL-7B	2	8.95 ±0.81	31.10 ±2.0	67.40 ±1.2
	10	8.20 ±0.82	29.80 ±2.0	67.48 ±1.2
	500	18.40 ±1.10	48.20 ±2.8	42.10 ±2.1
LLaVA-1.5-7B	10	13.85 ±0.78	47.50 ±2.2	75.60 ±1.0
	50	12.99 ±0.75	45.60 ±2.2	75.11 ±1.0
	1000	24.10 ±1.30	62.80 ±3.1	48.50 ±2.4
LLaVA-1.5-13B	50	12.60 ±0.78	41.20 ±2.1	77.10 ±1.0
	150	11.89 ±0.79	39.20 ±2.2	76.75 ±1.0
	1500	21.50 ±1.20	58.40 ±2.9	51.20 ±2.2

Aggressive removal degrades both hallucination metrics and recall: recall falls to 42.10 at $k = 500$ for Qwen2.5-VL-7B, 48.50 at $k = 1000$ for LLaVA-1.5-7B, and 51.20 at $k = 1500$ for LLaVA-1.5-13B. Thus, the benefit does not come from indiscriminately suppressing a large fraction of the representation.

The selected rank is model-dependent, with the larger LLaVA model requiring a larger removed subspace in this sweep. These observations motivate model-specific tuning rather than a fixed rank across architectures; they do not establish a proportional scaling law between k and hidden dimension. This selection procedure concerns training-free inference and does not change the separate fine-tuning configuration described in Section 4.2.

A.2 Implementation Details and Hyperparameters

For our fine-tuning experiments (Section 4.2), we adopt a parameter-efficient approach designed to solely optimize the cross-modal projection space. Specifically, we freeze both the vision encoder and the large language model (LLM) backbone, limiting all gradient updates exclusively to the modality projector (the Merger in Qwen2.5-VL and the MLP adapter in LLaVA-1.5). During the forward pass, we apply our bias removal module to orthogonally decouple the top 2 principal components of the textual manifold, forcing the projector to learn fine-grained visual semantics.

The models are trained using the AdamW optimizer with a cosine learning rate scheduler. To maintain memory efficiency and training stability, we utilize DeepSpeed ZeRO-3 optimization and Mixed Precision (bfloat16) with TF32 enabled. Table 5 details the exact hyperparameters used for fine-tuning both the Qwen2.5-VL-7B and LLaVA-1.5-7B architectures on the descriptive captioning datasets.

Table 5: Training hyperparameters for the geometric debiasing fine-tuning phase.

Hyperparameter	**Qwen2.5-VL-7B-Instruct**	**LLaVA-1.5-7B**
Trainable Modules	Merger (Projector)	MLP Adapter (Projector)
Frozen Modules	Vision Encoder, LLM	Vision Encoder, LLM
Number of Top PCs Removed	2	2
Training Epochs	2	2
Global Batch Size	126	48
Optimizer	AdamW	AdamW
Learning Rate	1e-3	2e-5
Learning Rate Schedule	Cosine	Cosine
Warmup Ratio	0.03	0.03
Weight Decay	0.1	0.0
Precision	`bfloat16` (+ TF32)	`bfloat16` (+ TF32)
DeepSpeed Stage	ZeRO-3 (Offload)	ZeRO-3
Gradient Checkpointing	True	True

A.3 Qualitative Analysis: Attention Score and Captions

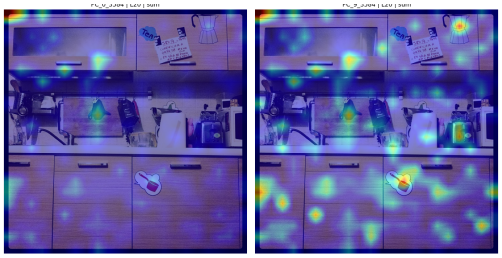
To further understand the effect of removing the modality anchor, we conduct a qualitative analysis using images from the COCO dataset. For each image, we prompt the model to list all objects that are clearly visible in the scene. We then compare the outputs produced by the baseline model and by our method after removing the top principal components associated with the modality anchor.

In addition to comparing the generated object lists, we analyze the cross-modal attention between generated textual tokens and visual patches. For each generated token, we aggregate the attention weights assigned to image patches and visualize them as spatial heatmaps over the input image. These heatmaps allow us to observe which regions of the image the model attends to while producing object descriptions.

In our visualizations (Figure 6), the left side of each panel corresponds to the baseline model, while the right side shows the model after removing the modality anchor. We observe that the baseline model often concentrates attention on a limited set of dominant foreground regions, overlooking smaller or less salient visual elements in the scene. In contrast, our method distributes attention across a broader set of semantically meaningful regions, enabling the model to detect additional objects that were previously ignored. In the captions accompanying each example, the **bolded phrases correspond to the objects** that are newly detected by our method but absent in the baseline output.



(a) Park scene with improved grounding of distant objects.



(b) Kitchen scene with recovery of fine-grained objects.

(a) Baseline: People sitting on a bench, bench with a bag on it, trash can, water body, trees and foliage, bridge in the background.
Ours: People sitting on a bench, bench with a bag placed on it, body of water, trees and foliage in the background, bridge over the water, trash can near the bench, **birds flying above**, **staircase to the left**, **people walking in the distance**.

(b) Baseline: Wooden cabinets, glass door cabinet, cutting board.
Ours: Wooden cabinets, glass door cabinet, cutting board, **kitchen utensils**, **dish rack**, **coffee maker**, **tea kettle**, **stickers (speech bubbles)**, **notes**, **sink**, **dish soap bottle**, **sponge holder**.

Figure 6: **Qualitative comparison of visual grounding via attention heatmaps.** The baseline model (left maps) suffers from attention smearing and focuses disproportionately on dominant foreground regions due to structural text bias. Removing the top principal components (right maps) sharpens the cross-modal attention, allowing the model to accurately detect and ground smaller, fine-grained objects that were previously ignored.