REDUCING HALLUCINATIONS IN MULTIMODAL LLMS WITH MODALITY-AWARE SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) often hallucinate visual content absent from the input image. Preference optimization has become a common post-training approach for mitigating hallucinations, but existing preference signals often rely on LLM-as-Judge supervision that can be sensitive to linguistic style rather than visual grounding. We thus introduce *Modality-Aware Sparse Autoencoders* (MA-SAE), an SAE architecture that separates vision-specific, text-specific, and cross-modal features to provide a visual grounding signal for preference labeling. By isolating modality-specific patterns, such as hedging and stylistic scaffolding, MA-SAE learns interpretable cross-modal features from which we derive a grounding score measuring how well a response is supported by the image. Across three MLLM backbones, MA-SAE learns more coherent and better aligned cross-modal features than existing SAE baselines, achieving the best performance across five feature quality metrics. Using its grounding score for preference labeling, MA-SAE substantially reduces hallucinations under Direct Preference Optimization (DPO), consistently outperforming LLM-as-Judge and other preference-labeling methods across three benchmarks. The grounding signal further transfers to mDPO and on-policy preference tuning, where MA-SAE achieves strong overall performance. These results show that interpretable cross-modal features can provide a direct signal of visual grounding, offering an alternative to LLM-as-Judge supervision.

1 INTRODUCTION

Multimodal large language models (MLLMs) achieve strong performance on a wide range of vision-language tasks, yet they commonly hallucinate by generating visual content unsupported by the input image (Liu et al., 2024a; Huang et al., 2024). Preference optimization is a widely used post-training approach for mitigating hallucinations, with Direct Preference Optimization (DPO) emerging as a prominent method as it requires no separate reward model (Rafailov et al., 2023). To efficiently scale DPO to large datasets, many works rely on LLM judges to label preference pairs (Li et al., 2024). However, LLM judges often exhibit linguistic biases that favor verbose or stylistically preferred responses, even when they are weakly grounded in the visual input (Chen et al., 2024a; Lee et al., 2026b). Therefore, training on such labels risks propagating these biases into post-trained models.

We investigate this bias through a controlled stylistic intervention on MMHal-Bench (Appendix A.1). Starting from the same model responses, we inject hedging (e.g., "appears to be") and stylistic language (e.g., "Based on the visual information in the image") while preserving the underlying claims. We then score the original and modified responses with LLM judges using the same evaluation prompts. Despite containing the same image-grounded claims, the modified responses receive statistically significant score increases from all three LLM judges and are less frequently flagged as hallucinated (Table 1). When LLM-judge scores are used to label preference pairs, we find that preferred responses contain substantially more hedging than rejected responses (Appendix A.2). DPO training on these preferences further increases hedging in model outputs (Appendix A.3), indicating that the model learns judge-favored linguistic patterns rather than visual grounding.

To obtain preference labels grounded in the image rather than in linguistic style, we propose *Modality-Aware Sparse Autoencoders* (MA-SAE), an SAE that separates MLLM representations into vision-specific, text-specific, and cross-modal feature banks. This decomposition lets modality-specific banks capture linguistic style and visual-only structure, while encouraging the cross-modal bank to

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.

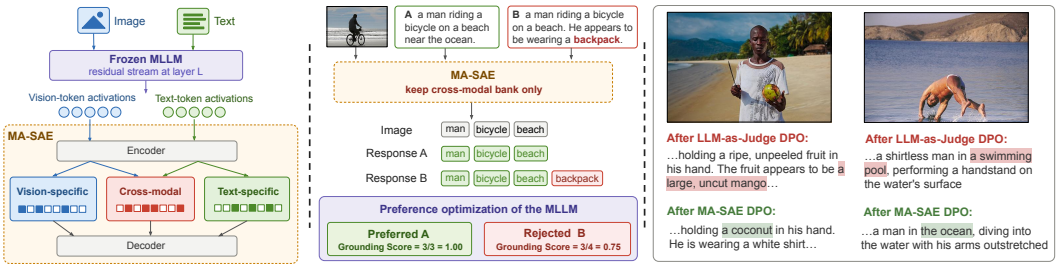


Figure 1: Left: MA-SAE decomposes MLLM representations into vision-specific, text-specific, and cross-modal feature banks. Middle: The cross-modal bank defines a grounding score used to label preference pairs for downstream optimization. Right: AMBER responses after DPO on pairs labeled by LLM-as-Judge vs. MA-SAE, where MA-SAE produces more visually grounded outputs.

capture concepts shared across vision and text. We then derive an interpretable grounding score from the cross-modal bank by measuring the fraction of response-activated features also supported by the image. Using this score to decide which response is preferred provides a direct visual grounding signal for preference supervision without relying on LLM judges (Figure 1).

Across three MLLM backbones, MA-SAE learns more coherent and better-aligned cross-modal features than existing SAE approaches, achieving the best performance across all five feature quality metrics. Using the grounding score for preference labeling, MA-SAE substantially reduces hallucinations under DPO across three benchmarks, consistently outperforming LLM-as-Judge and other preference-labeling methods. These gains further generalize to mDPO (Wang et al., 2024) and on-policy preference tuning (Tajwar et al., 2024), where MA-SAE continues to achieve strong performance. These results show that MA-SAE provides an interpretable cross-modal grounding signal for preference optimization, offering an alternative to supervision from LLM judges.

In summary, we make the following contributions:

1. We demonstrate that LLM judges can reward linguistic style independently of visual grounding, and further show that this bias is reflected in LLM-judge preference labels and propagates through preference optimization to the trained model.
2. We propose a preference-labeling method based on MA-SAE, a modality-aware SAE that separates modality-specific features capturing visual and linguistic patterns from shared cross-modal features. Its cross-modal bank yields an interpretable grounding score used to label preference pairs.
3. We show that MA-SAE learns more coherent cross-modal features than existing SAEs across three MLLM backbones and five feature-quality metrics. Using the grounding score for preference labeling substantially reduces hallucinations across three benchmarks under DPO, with gains extending to mDPO and on-policy preference tuning.

2 Related Work

SAEs for MLLMs. Building on the linear representation hypothesis (Park et al., 2023), SAEs decompose model activations into sparse, interpretable features (Bricken et al., 2023). Recent work has extended SAEs to MLLMs, with the goal of learning semantically coherent features across vision and text (Rajaram et al., 2025; Pach et al., 2026). However, naively applying an SAE does not enforce feature sharing across modalities, and recent methods therefore introduce cross-modal alignment. VL-SAE (Shen et al., 2026) aligns paired vision and text representations using a contrastive autoencoder. GSAE (Kaushik et al., 2026) instead encourages shared features through group-sparse regularization. Shu et al. (2026) aligns cross-modal features using a cosine-similarity objective. However, existing alignment objectives are applied over the full feature representation, which also contains modality-specific information that need not be shared across modalities. MA-SAE instead explicitly separates vision-specific, text-specific, and cross-modal features, allowing modality-specific patterns such as linguistic style to specialize in their own banks while applying alignment only to cross-modal features.

Biases in LLM Judges. LLM-as-Judge has become a widely used approach for scalable evaluation and preference supervision, but recent work has identified systematic biases in LLM-based judgments (Chen et al., 2024b; Zhou et al., 2024). In particular, LLM judges can favor superficial response

	LLM-judge score ↑			Flagged as hallucinated (%) ↓		
Judge	Original	Style-injected	Δ	Original	Style-injected	Δ
Gemini-3.8-Flash	2.38	**2.58**	$+0.20^*$	56.25	**53.12**	-3.12 pp
Claude Opus 5	2.19	**2.35**	$+0.17^*$	62.50	**61.46**	-1.04 pp
GPT-5.5	2.13	**2.29**	$+0.17^*$	63.54	**60.42**	-3.12 pp

Table 1: We apply a stylistic intervention to 96 MMHal-Bench model responses by injecting stylistic language while preserving the underlying claims. Across LLM judges, modified responses receive significant score increases and are flagged as hallucinated less often. $^*p < 0.05$ (paired Wilcoxon test).

properties such as verbosity and fluency, even when these qualities are not directly related to correctness (Zhou et al., 2024; Hu et al., 2025). In multimodal evaluation, LLM judges have also been shown to over-rely on linguistic cues, favoring linguistically plausible or stylistically preferred responses even when they are weakly grounded in the visual input (Chen et al., 2024a; Lee et al., 2026b). Such biases are particularly concerning for preference-based hallucination mitigation, where responses preferred by an LLM judge may be fluent or stylistically appealing while remaining weakly grounded in the image. Our work instead replaces LLM-as-Judge preference signals with an SAE-based grounding score that separates text-specific linguistic patterns from shared cross-modal features, yielding preferences based on visual grounding rather than linguistic style.

Preference Construction for MLLM Hallucination Mitigation. Preference optimization has become a widely used post-training approach for reducing hallucinations in MLLMs, with DPO and its variants becoming particularly prominent for directly optimizing over preference pairs without a separate reward model (Wang et al., 2024; Xie et al., 2024; Li et al., 2026b;a). Recent work has focused on how to construct preference pairs using reliable preference signals. VLFeedback (Li et al., 2024) uses LLM-as-Judge ratings to provide preference supervision, while CLIP-DPO (Ouali et al., 2024) derives preferences using CLIP image–text similarity. SAE-V (Lou et al., 2025) uses vanilla SAE feature weights to score cross-modal alignment for data filtering. DA-DPO (Qiu et al., 2026) combines contrastive CLIP and MLLM likelihood signals to estimate preference-pair difficulty. In contrast, our work labels preference pairs using cross-modal features learned through a modality-aware SAE decomposition, providing an interpretable grounding signal while isolating linguistic patterns that may otherwise bias preference labels.

3 Methodology

3.1 Preliminaries

Multimodal Large Language Models. A general architecture for vision-language MLLMs consists of a vision encoder, a connector, and a language model (Liu et al., 2023; Zhu et al., 2025; Bai et al., 2025). An input image is encoded into a sequence of visual tokens and projected into the language-model representation space through the connector. These visual tokens are then processed jointly with text tokens by the language model for generation. At a fixed layer, we denote the residual-stream activation of each token by $h \in \mathbb{R}^d$ and its modality by $m(h) \in \{V, T\}$, where V and T denote vision and text tokens.

Sparse Autoencoders. SAEs learn sparse decompositions of model activations into features that are often more monosemantic and human-interpretable. We build on the Top-K SAE (Gao et al., 2025), which encodes an activation $h \in \mathbb{R}^d$ as

$$z = \mathrm{TopK}(\mathrm{ReLU}(W_{\mathrm{enc}}(h - b))), \qquad \hat{h} = W_{\mathrm{dec}} z + b, \tag{1}$$

where $W_{\mathrm{enc}} \in \mathbb{R}^{p \times d}$ and $W_{\mathrm{dec}} \in \mathbb{R}^{d \times p}$ are the encoder and decoder matrices, $b \in \mathbb{R}^d$ is the reconstruction bias, and $z \in \mathbb{R}^p$ is the sparse feature representation. TopK retains the K largest activations and sets the remaining entries to zero. The SAE is trained to reconstruct the original activation by minimizing $\|\hat{h} - h\|_2^2$.

3.2 Modality-aware Sparse Autoencoder

To obtain a grounding signal not confounded by linguistic style, we introduce *Modality-Aware Sparse Autoencoder* (MA-SAE), which extends the SAE with two modifications. First, we partition the

sparse code z into vision-specific, text-specific, and shared banks through hard modality gating. Second, we apply a cross-modal alignment objective only to the shared bank to encourage shared features to capture semantic concepts across vision and text.

Modality-Partitioned Banks and Hard Gating. We partition the sparse code $z \in \mathbb{R}^p$ into three feature banks, $z = [z_V \mid z_T \mid z_C]$, where $z_V \in \mathbb{R}^{p_{\text{vis}}}$ is vision-specific, $z_T \in \mathbb{R}^{p_{\text{text}}}$ is text-specific, and $z_C \in \mathbb{R}^{p_{\text{cross}}}$ is a shared cross-modal bank, with $p = p_{\text{vis}} + p_{\text{text}} + p_{\text{cross}}$. Vision-specific features can activate only for vision tokens, text-specific features only for text tokens, while cross-modal features are accessible to both modalities.

We enforce this structure through modality-conditioned hard gating. Let $u = \text{ReLU}(W_{\text{enc}}(h - b)) = [u_V \mid u_T \mid u_C]$ denote the pre-sparsification activations. For a token h, we compute

$$\begin{aligned} z_V &= \mathbb{I}[m(h) = V]\,\text{TopK}_{K_{\text{mod}}}(u_V), \\ z_T &= \mathbb{I}[m(h) = T]\,\text{TopK}_{K_{\text{mod}}}(u_T), \\ z_C &= \text{TopK}_{K_{\text{cross}}}(u_C), \end{aligned} \tag{2}$$

Unlike an SAE that selects the top K features globally, MA-SAE splits the sparsity budget between the two accessible banks, with K_{mod} features selected from the token's modality-specific bank and K_{cross} from the cross-modal bank ($K = K_{\text{mod}} + K_{\text{cross}}$). This ensures that modality-specific features remain exclusive to their respective modality, while only cross-modal features can be shared across vision and text.

Cross-Modal Feature Alignment. Multimodal SAEs can learn *modality-split features*, where vision and text activate largely separate features rather than shared cross-modal concepts (Lee et al., 2026a). Existing approaches encourage cross-modal sharing through group sparsity or cosine-based alignment (Kaushik et al., 2026; Shu et al., 2026). However, in our partitioned architecture, such objectives can be trivially satisfied by an inactive cross-modal bank. We therefore introduce a Barlow Twins alignment objective (Zbontar et al., 2021) specifically on the cross-modal bank, which aligns corresponding vision–text features while preventing collapse.

For each sample s, we mean-pool the cross-modal activations z_C over vision and text tokens separately and standardize each feature across the batch, yielding $\bar{z}^{(s)}_{C,V}, \bar{z}^{(s)}_{C,T} \in \mathbb{R}^{p_{\text{cross}}}$. We define the cross-correlation between cross-modal features i and j as

$$C_{ij} = \frac{1}{N} \sum_{s=1}^{N} \bar{z}^{(s)}_{C,V,i} \bar{z}^{(s)}_{C,T,j}. \tag{3}$$

The Barlow Twins alignment loss is

$$\mathcal{L}_{\text{BT}} = \frac{1}{p_{\text{cross}}} \sum_i (1 - C_{ii})^2 + \lambda_{\text{off}} \frac{1}{p^2_{\text{cross}} - p_{\text{cross}}} \sum_i \sum_{j \neq i} C^2_{ij}. \tag{4}$$

The diagonal term aligns corresponding cross-modal features across modalities and prevents collapse, while the off-diagonal term reduces redundancy between distinct features. The alignment loss is applied only to the cross-modal bank, leaving the modality-specific banks free to specialize.

Training Objective. The full training objective combines reconstruction losses from both modalities with the Barlow Twins alignment loss,

$$\mathcal{L}_{\text{recon}} = \tfrac{1}{2}\big(\text{MSE}_V + \text{MSE}_T\big), \text{ where } \text{MSE}_m = \frac{1}{|\{h : m(h) = m\}|} \sum_{h : m(h) = m} \|\hat{h} - h\|_2^2. \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{BT}}\, \mathcal{L}_{\text{BT}}. \tag{6}$$

3.3 Grounding Score for Preference Labeling

We next use MA-SAE to derive preference supervision from visual grounding. We compute a grounding score using only the cross-modal bank to measure how well a response is supported by the image. Restricting the score to cross-modal features filters out text-specific patterns in z_T, such as hedging or stylistic scaffolding, as well as vision-specific patterns in z_V that capture visual-only structure.

Given an image and response, we run the frozen MLLM, extract the hidden states, and encode them with the frozen MA-SAE. Let $\mathcal{S}_V, \mathcal{S}_T \subseteq \{1, \ldots, p_{\text{cross}}\}$ denote the sets of cross-modal features in z_C activated by at least one vision or response token, respectively. We define the grounding score as

$$\text{Ground}(\mathcal{S}_V, \mathcal{S}_T) = \frac{|\mathcal{S}_V \cap \mathcal{S}_T|}{|\mathcal{S}_T|}, \tag{7}$$

which measures the fraction of response-activated cross-modal features also supported by the image.

Preference Labeling. We use the grounding score to label preference pairs. Given two candidate responses for the same input, the response with the higher grounding score is treated as preferred and the lower as rejected. The resulting pairs are then used for preference optimization.

4 Experiments

4.1 Experimental Setup

Models and Compute. We use three MLLMs across two model families and multiple model scales: LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024b), and InternVL3-8B (Zhu et al., 2025). SAE training uses a single GPU, and preference optimization uses four GPUs with DeepSpeed ZeRO-1. All experiments are conducted on NVIDIA A100 GPUs.

MA-SAE Training. We train one MA-SAE for each MLLM using residual-stream activations from the midpoint layer: layer 16 of 32 for LLaVA-1.5-7B, layer 20 of 40 for LLaVA-1.5-13B, and layer 14 of 28 for InternVL3-8B. We focus on intermediate layers following prior work suggesting that they capture richer semantic representations than earlier or final layers (Ma et al., 2026). We train the SAE on LLaVA-Instruct-150K (Liu et al., 2023), a multimodal instruction-following dataset of image–text conversations, holding out 5% of the data for evaluation. Full architecture and training hyperparameters, including bank sizes, sparsity level, learning rate, λ_{BT}, are provided in Appendix B.

Preference Optimization. For the second-stage preference optimization, we construct preference pairs from VLFeedback (Li et al., 2024), a large-scale multimodal preference dataset with responses generated by a pool of 12 MLLMs. To isolate the effect of preference labeling, all methods use the same candidate-response pools and training configuration, differing only in the labeling rule. We perform preference optimization with LoRA (Hu et al., 2021) on the language model, while keeping the vision tower and connector frozen. Full preference-pair construction details and training hyperparameters, including learning rate, batch size, and DPO β, are provided in Appendix F.

4.2 Evaluating MA-SAE Representations

Before applying MA-SAE to preference optimization, we first evaluate the quality of its learned representations. By separating modality-specific and shared features, we hypothesize that MA-SAE learns more coherent and better-aligned cross-modal representations. We begin with qualitative examples examining the distinct roles of features learned by the three banks, and then quantitatively compare MA-SAE against existing SAE approaches using established feature-quality metrics.

Qualitative Analysis. Figure 2 provides examples illustrating the distinct roles of features learned by the three MA-SAE banks. Vision-specific features activate on visual content not represented in the paired text, often highlighting background or scene-level regions. Text-specific features capture linguistic patterns such as hedging and inferential expressions (e.g., "seems to"), while cross-modal features encode shared concepts across modalities (e.g., zebra images and text mentions of zebras). These examples qualitatively support the intended decomposition between modality-specific and cross-modal information. Additional examples from all three banks are provided in Appendix C.

We next quantitatively compare MA-SAE cross-modal features against existing MLLM SAE baselines using established feature-quality metrics. We first introduce the baselines and evaluation metrics.

SAE Baselines. We compare MA-SAE against three SAE architectures for MLLMs: (a) *Vanilla SAE* (Rajaram et al., 2025), which applies a vanilla SAE to vision and text activations without explicit cross-modal alignment; (b) *VL-SAE* (Shen et al., 2026), which aligns paired vision and text representations with a contrastive autoencoder, then preserves this similarity with a distance-based encoder; (c) *GSAE* (Kaushik et al., 2026), which promotes cross-modal feature sharing through group-sparse regularization, encouraging paired modalities to activate a common subset of features.



Figure 2: Examples of MA-SAE features. Highlighted regions/tokens reflect high activations. Vision bank features capture visual-only structure absent from the caption, text bank features capture linguistic patterns such as hedging, and cross-modal features encode shared concepts across modalities.

Evaluation Metrics. A desirable MLLM SAE should learn semantically coherent shared features across modalities, where each feature captures a consistent concept while remaining distinct from others. Following Shen et al. (2026), we report Intra-Similarity and Inter-Similarity. Intra-Similarity measures the cosine similarity between the mean-pooled CLIP embeddings of a feature's top-activating vision and text examples, while Inter-Similarity measures the corresponding cross-feature similarity. We additionally report multimodal monosemanticity score (MMS) (Kaushik et al., 2026), which weights semantic similarity by co-activation strength and penalizes modality-split features. We report $\text{MMS}_{v\text{-}t}$ for cross-modal coherence and $\text{MMS}_{v\text{-}v}$ and $\text{MMS}_{t\text{-}t}$ for within-modality coherence. Higher MMS and Intra-Similarity and lower Inter-Similarity indicate better feature quality. For MA-SAE, all metrics are computed over the cross-modal bank. Detailed definitions and additional matched-budget comparisons are provided in Appendices D and E.1.

Quantitative Results. Table 2 shows the quantitative comparison of MA-SAE against existing SAE baselines. Across all five feature-quality metrics and all three MLLM backbones, MA-SAE consistently achieves the best performance. This suggests that separating modality-specific information from the shared feature space leads to more coherent and better-aligned cross-modal representations. Appendix E.1 further reports bimodal MMS and a matched-budget comparison, where MA-SAE continues to perform best across all three backbones.

Method	$\text{MMS}_{v\text{-}t}\uparrow$	$\text{MMS}_{v\text{-}v}\uparrow$	$\text{MMS}_{t\text{-}t}\uparrow$	Intra-Sim.↑	Inter-Sim.↓
LLaVA-1.5-7B					
Vanilla SAE (Rajaram et al., 2025)	0.1599	0.4670	0.6180	0.2912	0.2656
VL-SAE (Shen et al., 2026)	0.1596	0.5105	0.5984	0.2735	0.2696
GSAE (Kaushik et al., 2026)	0.1723	0.5066	0.6150	0.2974	0.2741
MA-SAE (ours)	**0.2386**	**0.6509**	**0.6960**	**0.3513**	**0.2313**
LLaVA-1.5-13B					
Vanilla SAE (Rajaram et al., 2025)	0.1589	0.4586	0.6172	0.2934	0.2638
VL-SAE (Shen et al., 2026)	0.1463	0.5170	0.5158	0.2770	0.2589
GSAE (Kaushik et al., 2026)	0.1731	0.5044	0.6162	0.2974	0.2709
MA-SAE (ours)	**0.2407**	**0.6522**	**0.6982**	**0.3526**	**0.2318**
InternVL3-8B					
Vanilla SAE (Rajaram et al., 2025)	0.1843	0.5319	0.6310	0.3043	0.2705
VL-SAE (Shen et al., 2026)	0.1364	0.5210	0.5001	0.2636	0.2513
GSAE (Kaushik et al., 2026)	0.1820	0.5260	0.6262	0.3038	0.2708
MA-SAE (ours)	**0.2387**	**0.6496**	**0.6950**	**0.3509**	**0.2303**

Table 2: Comparison of SAEs on feature-quality metrics across three MLLMs. Higher MMS and Intra-Similarity indicate greater semantic coherence, while lower Inter-Similarity indicates better feature separation. MA-SAE consistently achieves the best performance across all metrics and models.

Ablation Study. We ablate the MA-SAE architecture in Appendix E.2. Results show that adding modality-partitioned banks improves feature quality over the vanilla SAE, with the strongest performance achieved when combined with Barlow Twins alignment in the full MA-SAE.

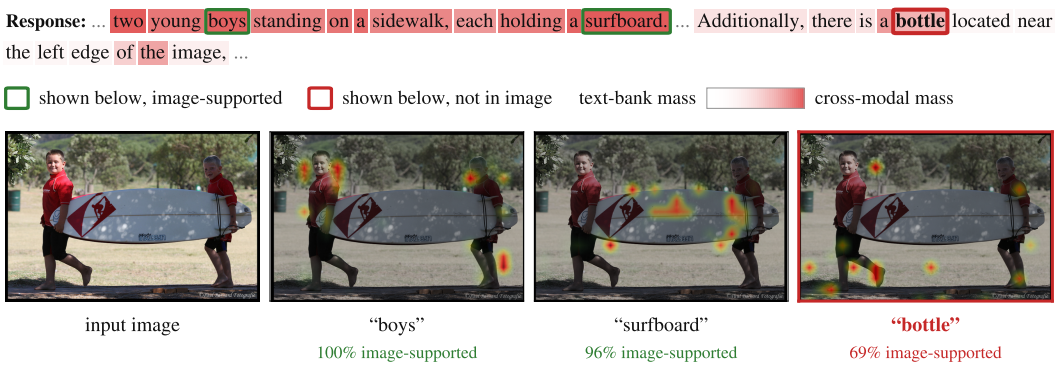


Figure 3: For each object mention, we measure the fraction of cross-modal features also activated by vision tokens and visualize the corresponding image regions. Grounded objects such as "boys" and "surfboard" receive strong image support, whereas the hallucinated "bottle" receives lower support.

4.3 Preference Optimization with MA-SAE

We next apply MA-SAE to preference optimization using the grounding score. We first analyze the grounding signal directly, then evaluate its effectiveness when used to label preference pairs.

Grounding Signal Analysis. We begin by examining whether the MA-SAE grounding signal distinguishes image-grounded objects from hallucinated ones as desired. Figure 3 illustrates how cross-modal features quantify image support for individual object mentions. For each object mention, image support measures the fraction of its cross-modal features that are also activated by vision tokens. Grounded objects such as "boys" and "surfboard" receive strong image support, whereas the hallucinated "bottle" receives substantially lower support. Additional qualitative examples are provided in Appendix G. We then evaluate this signal at scale on 500 MSCOCO samples. We group object mentions as grounded or hallucinated based on the image ground-truth annotations and average image support within each group. As shown in Figure 4, grounded mentions receive significantly higher image support across all three backbones (Mann–Whitney U test, $p < 10^{-3}$ for all backbones), demonstrating that the MA-SAE grounding signal effectively separates grounded from hallucinated content.

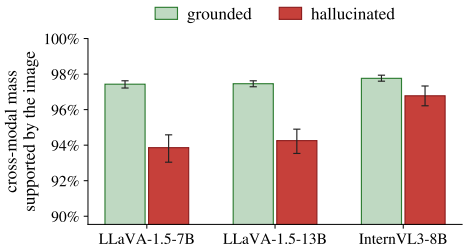


Figure 4: Image support for grounded and hallucinated object mentions across three MLLM backbones.

We next evaluate preference optimization using the MA-SAE grounding score to label preference pairs. We begin by introducing the preference-labeling baselines and evaluation metrics.

Preference Labeling Baselines. We compare MA-SAE against four baselines for labeling preference pairs: (a) *LLM-as-Judge* (Li et al., 2024), which uses the GPT-4V ratings provided by VLFeedback as the preference score; (b) *SAE-V* (Lou et al., 2025), which uses vanilla SAE features and scores responses using cosine-based weights from their top-activating vision and text examples; (c) *CLIP-DPO* (Ouali et al., 2024), which scores each response by the cosine similarity between its CLIP text and image embeddings; and (d) *DA-DPO* (Qiu et al., 2026), which combines the CLIP-based score with response log-probabilities from a pretrained reference MLLM.

Benchmarks and Metrics. We evaluate generative hallucination on three widely used benchmarks (Wang et al., 2024; Ouali et al., 2024; Lu et al., 2025; Qiu et al., 2026): AMBER (Wang et al., 2023), Object HalBench (Rohrbach et al., 2018), and MSCOCO (Hua et al., 2025). AMBER evaluates object hallucination using CHAIR, object coverage (Cover), response-level hallucination rate (HalRate), and cognitive hallucination (Cog). Object HalBench and MSCOCO report sentence-level CHAIR_s and instance-level CHAIR_i. CHAIR_s measures the fraction of generated responses containing at least one hallucinated object, and CHAIR_i measures the fraction of object mentions that are hallucinated. Lower CHAIR, HalRate, and Cog indicate fewer hallucinations, while higher Cover indicates greater coverage of image content. We intentionally exclude MMHal-Bench (Sun et al., 2024), whose evaluation relies on an LLM judge, to avoid the stylistic biases demonstrated in Table 1.

Method	AMBER				Object HalBench		MSCOCO	
	CHAIR↓	Cover↑	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓
DPO (Rafailov et al., 2023)								
LLaVA-1.5-7B								
Untuned	7.7	51.8	35.6	4.4	52.67	13.92	60.80	19.63
LLM-as-Judge	5.8	54.3	32.7	2.4	43.67	10.72	52.00	15.09
CLIP-DPO (Ouali et al., 2024)	5.3	**55.3**	30.4	2.8	42.33	9.99	50.60	14.42
SAE-V (Lou et al., 2025)	7.6	54.1	39.5	4.3	59.33	16.32	65.80	21.03
DA-DPO (Qiu et al., 2026)	5.5	53.5	31.9	2.9	51.33	12.62	56.00	16.00
MA-SAE (ours)	**4.3**	48.0	**17.2**	**1.5**	**18.67**	**6.38**	**29.00**	**10.88**
LLaVA-1.5-13B								
Untuned	6.6	51.8	30.8	3.3	50.67	13.98	58.20	19.01
LLM-as-Judge	6.1	54.8	33.6	2.7	41.33	10.68	49.60	15.16
CLIP-DPO (Ouali et al., 2024)	5.7	**55.5**	31.5	2.5	45.67	10.51	51.40	14.37
SAE-V (Lou et al., 2025)	6.8	53.6	34.4	3.5	51.00	13.79	57.00	18.57
DA-DPO (Qiu et al., 2026)	5.4	53.6	29.8	2.6	46.67	10.74	54.20	15.13
MA-SAE (ours)	**4.1**	49.6	**17.9**	**1.3**	**26.67**	**7.75**	**36.60**	**12.74**
InternVL3-8B								
Untuned	7.5	71.2	62.4	5.4	27.00	7.60	34.40	11.23
LLM-as-Judge	7.9	71.2	63.2	5.7	30.33	7.63	38.00	11.66
CLIP-DPO (Ouali et al., 2024)	7.7	72.3	63.3	6.2	26.67	7.71	36.00	10.84
SAE-V (Lou et al., 2025)	7.2	70.2	59.1	5.2	28.67	7.46	37.20	11.51
DA-DPO (Qiu et al., 2026)	7.4	71.1	61.5	5.5	30.33	7.53	37.60	11.05
MA-SAE (ours)	**6.9**	**72.5**	**57.4**	**5.0**	**21.00**	**6.36**	**29.80**	**10.25**

Table 3: Off-policy preference tuning with DPO. We compare MA-SAE with preference-labeling baselines. MA-SAE achieves the strongest overall performance across metrics and backbones.

Method	AMBER				Object HalBench		MSCOCO	
	CHAIR↓	Cover↑	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓
mDPO (Wang et al., 2024)								
LLaVA-1.5-7B								
LLM-as-Judge	4.2	52.4	23.7	2.0	37.33	10.81	45.20	14.26
CLIP-DPO (Ouali et al., 2024)	4.6	52.4	25.6	2.2	37.67	10.29	45.20	14.25
SAE-V (Lou et al., 2025)	4.7	**53.4**	24.9	2.0	36.33	9.06	44.20	13.10
DA-DPO (Qiu et al., 2026)	4.9	51.6	27.2	2.3	44.00	11.73	50.00	15.33
MA-SAE (ours)	**3.8**	51.1	**19.4**	**1.7**	**30.67**	**8.76**	**37.60**	**12.53**
LLaVA-1.5-13B								
LLM-as-Judge	**3.7**	52.1	20.2	**1.5**	35.33	9.49	42.20	13.58
CLIP-DPO (Ouali et al., 2024)	4.0	52.2	21.5	1.7	33.33	9.48	43.00	13.92
SAE-V (Lou et al., 2025)	4.4	**52.5**	23.9	1.8	36.67	9.21	46.20	13.27
DA-DPO (Qiu et al., 2026)	**3.7**	50.2	20.0	1.7	39.33	10.90	46.60	14.80
MA-SAE (ours)	**3.7**	49.7	**18.6**	**1.5**	**27.00**	**7.60**	**37.80**	**12.17**
InternVL3-8B								
LLM-as-Judge	7.9	72.0	63.4	6.3	33.00	7.81	40.00	11.99
CLIP-DPO (Ouali et al., 2024)	7.5	72.0	60.7	5.7	28.67	7.71	38.20	11.90
SAE-V (Lou et al., 2025)	7.5	70.9	59.3	**5.6**	29.67	7.87	38.20	11.79
DA-DPO (Qiu et al., 2026)	7.7	71.7	64.0	6.1	33.67	8.03	40.80	11.96
MA-SAE (ours)	**7.1**	**72.5**	**57.1**	**5.6**	**28.33**	**7.11**	**35.20**	**11.04**

Table 4: Generalization to mDPO. MA-SAE remains competitive with preference-labeling baselines across three MLLM backbones, showing that its benefits extend beyond standard DPO.

Off-policy DPO Results. Table 3 reports performance under DPO using off-policy VLFeedback. MA-SAE achieves the strongest overall performance across all three MLLM backbones and hallucination benchmarks, consistently reducing hallucinations relative to LLM-as-Judge and other baselines. These gains are consistent across backbones, suggesting that the MA-SAE grounding score provides a more reliable signal for preference labeling. Figure 1 (right) shows AMBER responses after DPO on pairs labeled by LLM-as-Judge vs. MA-SAE, with additional examples in Appendix H.

We additionally examine whether preferences labeled by MA-SAE exhibit the hedging bias observed for LLM-derived preferences. As shown in Appendix A.4, preferred and rejected responses labeled by MA-SAE exhibit nearly identical levels of hedging, indicating that MA-SAE preference labeling does not exhibit the same stylistic bias.

Method	AMBER				Object HalBench		MSCOCO	
	CHAIR↓	Cover↑	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓
On-policy DPO (Tajwar et al., 2024)								
LLaVA-1.5-7B								
LLM-as-Judge	5.5	50.8	28.7	2.7	42.33	10.80	49.80	15.26
MA-SAE (ours)	**5.1**	**51.0**	**23.8**	**1.9**	**23.00**	**7.11**	**33.40**	**11.93**
LLaVA-1.5-13B								
LLM-as-Judge	4.9	50.8	23.0	2.1	37.00	9.92	45.60	14.84
MA-SAE (ours)	**4.5**	**52.4**	**20.9**	**1.5**	**25.33**	**6.91**	**34.80**	**11.23**
InternVL3-8B								
LLM-as-Judge	7.3	70.8	60.6	5.2	29.33	8.17	37.40	11.99
MA-SAE (ours)	**5.9**	**71.1**	**45.4**	**3.9**	**19.33**	**5.62**	**27.40**	**9.59**

Table 5: Generalization to on-policy preference tuning. We generate candidate responses from the current policy. MA-SAE consistently achieves the best performance across metrics and backbones, demonstrating that its gains extend to the on-policy setting.

Generalization to mDPO. Because MA-SAE labels preference pairs independently of the optimization objective, it can be paired with different preference-learning methods. We next evaluate this generality with mDPO (Wang et al., 2024), a multimodal DPO variant that augments response preferences with contrastive image preferences. As shown in Table 4, MA-SAE remains competitive across metrics and backbones, suggesting that its benefits extend beyond the standard DPO objective.

Generalization to On-policy Preference Tuning. Experiments above use the off-policy VLFeedback dataset, where candidate responses are generated by a fixed pool of 12 MLLMs. Recent work suggests that preference optimization can benefit from on-policy data, where candidates are generated by the model being optimized (Tajwar et al., 2024; Yang et al., 2025). We therefore evaluate an on-policy setting where candidate responses are sampled from the current policy at temperature 1 and then paired and labeled using the corresponding scoring method. As shown in Table 5, MA-SAE consistently outperforms LLM-as-Judge across all three backbones, achieving better performance across all evaluation metrics. These results show that the benefits of MA-SAE extend beyond fixed off-policy preference data to the on-policy setting.

Ablation of Modality-Specific Banks. Finally, we examine whether separating modality-specific features is needed for the grounding signal. We compare the full MA-SAE against a variant without the vision- and text-specific banks (full results in Appendix I). Removing the modality-specific banks consistently degrades hallucination metrics across all three backbones, indicating that explicitly isolating modality-specific information is important for a reliable cross-modal grounding signal.

5 Conclusion, Limitations and Future Work

We introduce MA-SAE, a modality-aware SAE that enables grounding-based preference labeling to mitigate hallucinations in MLLMs. Motivated by the observation that LLM-judge preference labels can favor linguistic style over visual grounding, MA-SAE explicitly separates vision-specific, text-specific, and cross-modal features, allowing such text-specific patterns to be isolated while using the cross-modal bank to directly measure image–text grounding. Across three MLLM models, MA-SAE consistently learns more coherent cross-modal representations, and its preference labels substantially reduce hallucinations under DPO. Its benefits further extend to mDPO and on-policy preference tuning, suggesting that the learned grounding signal generalizes across preference-optimization objectives.

This study also has several limitations. First, our current grounding formulation and evaluation primarily focus on object-level hallucinations. Multimodal hallucinations can also involve errors in attributes, relations, and knowledge-dependent claims (Wang et al., 2023; Guan et al., 2024). Extending MA-SAE to address these broader forms of hallucination remains an important direction for future work. Second, our experiments consider models up to 13B parameters, and it remains an open question whether the same feature decomposition and grounding signal remain effective in substantially larger models. Lastly, our current formulation focuses on vision and text. Future work could investigate whether the same modality-specific and shared feature decomposition extends to other modalities, such as audio and video (Rubenstein et al., 2023; Ren et al., 2024).

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments and implement methods. We have not used generative AI tools to propose or refine hypotheses, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, and identify relevant literature. We have reviewed all AI-assisted work, including verifying and testing LLM-generated code for correctness, reading and checking AI-suggested literature, and reviewing AI-assisted figures and content for accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the experimental settings, implementation details, datasets, baselines, hyperparameters, and training procedures needed to reproduce our results in the main paper and appendix (Section 4, Appendix B, Appendix F). Anonymous source code is included with the submission and will be publicly released upon acceptance.

REFERENCES

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, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, 2024b.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in llm-based preference evaluations. In *EMNLP (Findings)*, pp. 6763–6794, 2025.

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

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631, 2024.

Chiraag Kaushik, Davis Barch, and Andrea Fanelli. Decomposing multimodal embedding spaces with group-sparse autoencoders. *arXiv preprint arXiv:2601.20028*, 2026.

Chungpa Lee, Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Same concept, different directions: Cross-modal feature heterogeneity in sparse autoencoders. *arXiv preprint arXiv:2606.29888*, 2026a.

Sua Lee, Sanghee Park, and Jinbae Im. Mm-judgebias: A benchmark for evaluating compositional biases in mllm-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25336–25373, 2026b.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, 2024.

Xintong Li, Chuhan Wang, Junda Wu, Rohan Surana, Tong Yu, Julian McAuley, and Jingbo Shang. Importance sampling for multi-negative multimodal direct preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 96851–96867, 2026a.

Xudong Li, Mengdan Zhang, Peixian Chen, Xiawu Zheng, Yan Zhang, Jingyuan Zheng, Yunhang Shen, Ke Li, Chaoyou Fu, Xing Sun, et al. Zooming from context to cue: Hierarchical preference optimization for multi-image mllms. *Advances in Neural Information Processing Systems*, 38: 179–214, 2026b.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

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

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

Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Jinda Lu, Junkang Wu, Jinghan Li, Xiaojun Jia, Shuo Wang, YiFan Zhang, Junfeng Fang, Xiang Wang, and Xiangnan He. Dama: Data-and model-aware alignment of multi-modal llms. *arXiv preprint arXiv:2502.01943*, 2025.

George Ma, Zhongyuan Liang, Irene Y Chen, and Somayeh Sojoudi. Do sparse autoencoders identify reasoning features in language models? *arXiv preprint arXiv:2601.05679*, 2026.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *Advances in Neural Information Processing Systems*, 38:95706–95742, 2026.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Longtian Qiu, Shan Ning, Chuyu Zhang, Jiaxuan Sun, and Xuming He. Da-dpo: Cost-efficient difficulty-aware preference optimization for reducing mllm hallucinations. *Transactions on Machine Learning Research*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Achyuta Rajaram, Sarah Schwettmann, Jacob Andreas, and Arthur Conmy. Line of sight: On linear representations in vllms. *arXiv preprint arXiv:2506.04706*, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 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.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set. *Advances in Neural Information Processing Systems*, 38:45235–45265, 2026.

Huizhen Shu, Xuying Li, Hongxu Lin, Wenjie Sun, and Hui Li. Steering vision-language models with joint sparse autoencoders. *arXiv preprint arXiv:2606.25657*, 2026.

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

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *arXiv preprint arXiv:2404.14367*, 2024.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

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

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620. IEEE, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. Mitigating the bias of large language model evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 1310–1319, 2024.

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.

Appendix

Table of Contents

A Linguistic Bias in LLM-Derived Preferences **15**
- A.1 Style Injection for the LLM-Judge Probe . . . 15
- A.2 Preference Labels Favor Hedged Responses . . . 16
- A.3 DPO Propagates Hedging into Model Outputs . . . 17
- A.4 MA-SAE Preference Labeling Does Not Favor Hedging . . . 18

B MA-SAE Architecture and Training Details **19**

C Qualitative Analysis of Learned MA-SAE Features **20**
- C.1 Cross-modal Bank . . . 20
- C.2 Vision-specific Bank . . . 24
- C.3 Text-specific Bank . . . 25

D SAE Feature Evaluation Metrics **28**
- D.1 Intra- and Inter-Similarity . . . 28
- D.2 Multimodal Monosemanticity . . . 28

E Additional MA-SAE Experiments **30**
- E.1 Cross-Modal Coverage and Matched-Budget Comparison . . . 30
- E.2 Component Ablation . . . 31

F DPO and mDPO Training Details **32**

G Additional Qualitative Visualizations of MA-SAE Grounding **33**

H Additional Qualitative Comparison of MA-SAE and LLM-as-Judge on AMBER **35**

I Ablation of Modality-Specific Banks for Preference Optimization **36**

J Response Length and Object Coverage Analysis **37**

A Linguistic Bias in LLM-Derived Preferences

A.1 Style Injection for the LLM-Judge Probe

To test whether LLM judges are sensitive to linguistic style independently of visually grounded content, we construct a controlled style-injection intervention on the widely used MMHal-Bench (Sun et al., 2024). We choose MMHal-Bench because its evaluation relies on an LLM judge to score model responses, making it a natural testbed for probing whether stylistic changes can influence multimodal hallucination evaluation.

We use the 96 MMHal-Bench questions and the corresponding responses generated by the untuned LLaVA-1.5-7B baseline. We apply a deterministic rule-based style injection designed to modify only surface-level linguistic style.

The Style Injection. We apply three families of edits, none of which introduces a new verifiable claim about the image.

1. *Hedging.* Eight regular-expression rules convert assertions into hedged versions of the same assertions. For example, "there is X" becomes "there appears to be X"; "the *noun* is" becomes "the *noun* appears to be"; and "I can see" becomes "I can observe what appears to be."
2. *Preamble.* We prepend the fixed opener "Based on the visual information in the image," unless the response already begins with a similar phrase. This addition restates that the response is describing the image without adding new image-specific content.
3. *Stylistic filler.* We append one sentence from a closed set of five stylistic filler sentences that do not introduce verifiable image content, such as "The image captures this moment with clarity, allowing for a detailed observation."

Importantly, the style injection changes only linguistic presentation by making the response more verbose and hedged; it does not add or remove objects, attributes, or other image-grounded claims. The example below illustrates the response before and after style injection.

Original	No one is sitting on the bench. It is empty.
Style-injected	**Based on the visual information in the image,** no one **appears to be** sitting on the bench. It **appears to be** empty. **The image captures this moment with clarity, allowing for a detailed observation.**

LLM-Judge Scoring. We evaluate the original and style-injected responses using three LLM judges: Gemini-3.8-Flash, Claude Opus 5, and GPT-5.5. For each judge, we use the MMHal-Bench evaluation prompts unchanged and score the same 96 response pairs under the original and style-injected conditions. Table 1 shows that all three judges assign higher scores to the style-injected responses and flag them as hallucinated less often, despite the underlying factual content being unchanged. This provides direct evidence that LLM-judge evaluations are sensitive to linguistic style, favoring stylistically preferred responses even when the underlying image-grounded content is held fixed.

A.2 Preference Labels Favor Hedged Responses

Building on the finding that LLM judges can reward more verbose and hedged responses even when visually grounded content is held fixed (Table 1), we next examine whether this stylistic preference is reflected in preference pairs labeled with LLM-judge ratings.

We analyze VLFeedback (Li et al., 2024), a widely used multimodal preference dataset whose response ratings are provided by GPT-4V. To mitigate differences arising from response length, we restrict the analysis to preference pairs whose preferred and rejected responses differ in word count by at most 15%. As shown in Table 6, preferred responses contain substantially more hedging than rejected responses across nearly all expressions, with $1.48\times$ more hedge occurrences overall. This suggests that the stylistic preferences of the LLM judge are reflected in the resulting preference labels when its ratings are used to construct the dataset.

	Preferred	Rejected	Ratio
Phrase-level occurrences			
"suggest(s) / suggesting"	549	321	$1.71\times$
"appear(s) to (be)"	381	206	$1.85\times$
"could be"	231	193	$1.20\times$
"might be"	252	220	$1.15\times$
"indicate(s) / indicating"	367	256	$1.43\times$
"likely"	260	196	$1.33\times$
"possibly"	241	165	$1.46\times$
"may be"	85	50	$1.70\times$
other hedges	298	187	$1.59\times$
Summary			
All hedges	**2664**	**1794**	$\mathbf{1.48}\times$
Responses with hedge (%)	60.6	55.5	$1.09\times$

Table 6: LLM-derived preference labels favor hedged wording. After controlling for response length, preferred responses still contain substantially more hedging than rejected responses, suggesting that LLM-derived preference signals favor linguistic style beyond length alone.

A.3 DPO Propagates Hedging into Model Outputs

We next analyze whether DPO training on LLM-derived preference data transfers linguistic biases from the judge model into the optimized model, focusing on hedging and uncertainty expressions. We compare LLaVA-1.5-7B before and after DPO across AMBER, Object HalBench, and MMHal-Bench, measuring both phrase-level hedge occurrences and the fraction of responses containing at least one hedge. As shown in Table 7, DPO consistently increases hedging across all three benchmarks, suggesting that stylistic preferences encoded by the LLM judge can be propagated through preference optimization into the trained model.

	AMBER		Obj. HalBench		MMHal-Bench		Ratio
	Base	DPO	Base	DPO	Base	DPO	
Phrase-level occurrences							
appear(s) to (be)	649	1252	143	376	6	53	2.1×
possibly	348	414	128	168	5	11	1.2×
seem(s) to (be)	82	315	17	83	1	9	4.1×
suggest(s) / suggesting	78	233	37	112	3	19	3.1×
likely	89	232	49	113	2	17	2.6×
indicate(s) / indicating	37	174	28	69	2	8	3.7×
slightly	62	182	10	42	0	5	3.2×
relatively	20	76	1	12	0	3	4.3×
other hedges	48	197	28	71	1	17	3.7×
Summary							
All hedges	1413	**3075**	441	**1046**	20	**142**	2.3×
Responses with hedge (%)	79	**98**	77	**100**	11	**55**	1.29×

Table 7: DPO on LLM-derived preferences substantially increases hedging. We compare LLaVA-1.5-7B before and after DPO training on preference pairs labeled using LLM-judge ratings. Hedging increases consistently across all three hallucination benchmarks, both in phrase-level occurrences and in the fraction of responses containing hedging.

A.4 MA-SAE Preference Labeling Does Not Favor Hedging

We next repeat the analysis of Appendix A.2 using preference pairs labeled by the MA-SAE grounding score rather than LLM-judge ratings. As shown in Table 8, preferred and rejected responses exhibit nearly identical levels of hedging overall ($1.02\times$), with no consistent preference across individual expressions. In contrast to LLM-derived labels, MA-SAE preference labeling does not systematically favor hedged wording.

	Preferred	Rejected	Ratio
Phrase-level occurrences			
"suggest(s) / suggesting"	601	503	$1.19\times$
"appear(s) to (be)"	329	368	$0.89\times$
"could be"	409	398	$1.03\times$
"might be"	410	408	$1.00\times$
"indicate(s) / indicating"	368	357	$1.03\times$
"likely"	337	307	$1.10\times$
"possibly"	287	299	$0.96\times$
"may be"	110	116	$0.95\times$
other hedges	181	221	$0.82\times$
Summary			
All hedges	**3032**	**2977**	$\mathbf{1.02}\times$
Responses with hedge (%)	45.9	47.1	$0.98\times$

Table 8: MA-SAE preference labeling does not favor hedged wording. After controlling for response length, preferred and rejected responses exhibit nearly identical levels of hedging, in contrast to preferences labeled with GPT-4V ratings.

B MA-SAE ARCHITECTURE AND TRAINING DETAILS

Table 9 summarizes the MA-SAE architecture and training hyperparameters for the three backbones. For a controlled comparison, all SAE baselines (Vanilla SAE, VL-SAE, and GSAE) are trained on the same backbone layer and LLaVA-Instruct-150K training split, and evaluated on the same held-out split. Each baseline uses the same dictionary size p and per-token sparsity budget $K = 64$ as MA-SAE, ensuring matched parameter and sparsity budgets across methods.

Hyperparameter	**LLaVA-1.5-7B**	**LLaVA-1.5-13B**	**InternVL3-8B**
MA-SAE Architecture			
SAE layer	16	20	14
Input dimension	4096	5120	3584
Expansion factor	$5\times$	$5\times$	$5\times$
Dictionary size p	20,480	25,600	17,920
Vision-only bank	6,144 (30%)	7,680 (30%)	5,376 (30%)
Text-only bank	6,144 (30%)	7,680 (30%)	5,376 (30%)
Cross-modal bank	8,192 (40%)	10,240 (40%)	7,168 (40%)
Shared Training Settings			
Activation	ReLU + Top-K		
Top-K budget	$K = 64$		
K_{mod}	27		
K_{cross}	37		
Learning rate	5×10^{-5}		
Warmup steps	1,000		
LR schedule	Warmup $\rightarrow$ constant		
Training epochs	3		
Batch size	32 samples/step		
Adam β_1, β_2	0.9, 0.999		
Total training steps	32,967		
λ_{BT}	5.0		
Barlow Twins off-diagonal weight	0.005		
Alignment warmup steps	5,000		

Table 9: MA-SAE architecture and training hyperparameters.

C Qualitative Analysis of Learned MA-SAE Features

Qualitative analysis of the features learned by MA-SAE at layer 16 of LLaVA-1.5-7B, trained on LLaVA-Instruct-150K activations and evaluated on a held-out split. We show features from all three banks: vision-specific, text-specific, and the shared cross-modal bank. For each feature, we present its top-activating examples, visualized as heatmaps over the image patch grid for vision activations and per-token shading for text activations. We observe qualitatively consistent feature patterns in MA-SAE dictionaries trained on LLaVA-1.5-13B and InternVL3-8B.

C.1 Cross-modal Bank

Cross-modal bank: features encode shared concepts across modalities.

Each block shows one feature from the cross-modal bank. The images show its strongest vision-side activations as heatmaps over the patch grid, while the corresponding responses show its strongest text-side activations with per-token shading.

Feature 17870 **double-decker bus**



1 2 3 4

1. Double-decker buses, with their distinctive two-level design, have been an iconic symbol of ...
2. A double-decker bus is a type of bus that has two levels or decks, ...
3. A double decker bus offers a unique and enhanced experience for passengers compared to regular ...
4. ... as a contrast to the busy city life with vehicles like the double-decker bus and cars driving around, creating a balance between urban and natural...

Feature 18556 **zebra**



1 2 3 4

1. Preserving the zebra's natural habitat involves various factors that help maintain a healthy ecosystem in ...
2. ... in a zoo, wildlife park, or sanctuary. In such facilities, animals like zebras are kept in contained areas to ensure their safety, the safety of ...
3. In their natural habitat, zebras typically graze on grasses and other vegetation. Their diet primarily consists of ...
4. ... welfare advocates might have concerns regarding the living conditions of the two zebras seen in the image. Their enclosure, which consists of paved surfaces and ...

Feature 15849 giraffe



1 2 3 4

1. ... a grassland or field environment, which is a natural habitat for both giraffes and zebras.
2. ... picture was taken at a wildlife park, zoo, or sanctuary, as the giraffes are enclosed in a fenced area, hinting at a man-made environment designed...
3. An ostrich might be housed in the same enclosure as giraffes in a zoo because both animals are native to similar habitats, primarily...
4. ... or savanna, which is a characteristic environment for giraffes in the wild. Giraffe populations are predominantly found in sub-Saharan African countries, spanning across countries like ...

Feature 17382 elephant



1 2 3 4

1. In the image, there are two elephants - a mother elephant and a baby elephant - standing together on...
2. Elephant habitats and conservation efforts are crucial to the survival and well-being of ...
3. ... elephant might be walking with its baby for several reasons. First, adult elephants play a critical role in the upbringing and protection of their young,...
4. ... engaged in shared activities such as swimming or playing in a river. Elephants are known for their intelligence, intricate communication, and strong family ties. They ...

Feature 15772 sheep

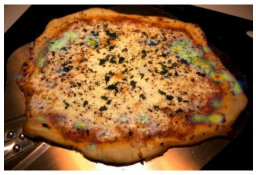


1 2 3 4

1. The sheep in the foggy and rocky environment might face several challenges: ...
2. There are several reasons for raising sheep in a field with tall grass: Natural diet: Sheep ...
3. ... maintains a balanced ecosystem. Soil fertility and erosion control: Sheep's natural grazing behavior contributes to soil health, as the even distribution of ...
4. The presence of the sheep grazing in the grass-covered field suggests a rural, agricultural, or pastoral setting.

Feature 13809 **pizza**



1 2 3 4

1. The pizza with cheese and black olives in the image could be suitable for ...
2. Several factors could have contributed to the burnt crust of the pizza in the image. These factors may include: Overcooking: The ...
3. To enhance the flavor of the cheese pizza, one can add various toppings and seasonings to suit their preferences.
4. The appeal of the pizza in the image primarily comes from its presentation, variety of toppings, and...

Feature 18036 **cake**






1 2 3 4

1. ... to determine the exact occasion based on the image alone, the decorated cake could be for a variety of occasions, such as a birthday, anniversary,...
2. When planning to use this beautifully designed cake for a special occasion, several considerations should be made: ...
3. The elaborate decoration of the cake, which features a woman in a carriage and a bride and groom...
4. The cake in the image is decorated with an Angry Birds theme, featuring green...

Feature 17460 **fruit**






1 2 3 4

1. Consuming a diverse range of fruits provides several health benefits, as different fruits offer a unique mix ...
2. The neat organization of fruits in the market is beneficial to both sellers and customers in several ...
3. ... shorter distance from the farm to the point of sale. Conversely, imported fruits and vegetables may spend more time in transit, affecting their freshness. ...
4. Consuming a variety of fruits on a regular basis offers numerous health benefits, mainly due to their ...

Feature 14445 **baseball batter**






1 2 3 4

1. ... situation, several factors play a crucial role in the success of the batter. These include: Timing: The batter's ability to anticipate the ...
2. ... game. You can witness various elements of the game, such as the batter at the plate with a baseball bat and a player with a...
3. ... out by the defense. During their turn at bat, the batter tries to hit the ball thrown by the pitcher. The pitcher's objective ...
4. When a pitch is thrown during a baseball game, the batter has a few options for their actions: Swing ...

Feature 18595 **snow sports (skiing, snowboarding)**






1 2 3 4

1. ... and equipment for snowboarders who wish to safely and enjoyably navigate the snowy mountain terrain.
2. While snowboarding in a wooded area as depicted in the image, there are a ...
3. Snowboarding and performing tricks in the air can be thrilling and enjoyable, but...
4. When skiing in remote areas, it is crucial for the skier to take several ...

C.2 VISION-SPECIFIC BANK

Vision-specific bank: features capture modality-specific visual content.

Each block shows one feature from the visual-only bank. The images show its strongest activations as heatmaps over the patch grid. These features are intended to capture visual content not represented in the paired text; qualitatively, they primarily respond to background, spatial, and scene-level patterns rather than salient objects mentioned in the text.

Feature 1860



1 2 3 4 5

Feature 2020



1 2 3 4 5

Feature 2411



1 2 3 4 5

Feature 916



1 2 3 4 5

Feature 3208



1 2 3 4 5

Feature 350



1 2 3 4 5

C.3 TEXT-SPECIFIC BANK

Text-specific bank: features capture modality-specific linguistic patterns.

Each block shows one feature from the text-only bank. The responses show its strongest activations with per-token shading. These features are intended to capture linguistic patterns not grounded in the paired image; qualitatively, they respond to modality-specific language such as hedging, uncertainty, and inferential expressions.

Feature 6863

1. ... that require writing or manual note-taking. Furthermore, the cup on the table implies that the person might be consuming a beverage, such as coffee or ...
2. ... for tasks such as opening packages or cutting herbs. The multiple bottles imply that they likely use various liquids, sauces, or seasonings during their cooking...
3. ... dishes. The presence of the tray or the covered food try also implies that the food is ready to be served or consumed, which may...
4. ... facility that holds importance in its community. The crane in the background implies that this area might be going through some transformation, preserving or renovating...

Feature 8974

1. The image suggests that the purpose of the vehicle is to transport passengers and their...
2. The scene suggests that someone may be in the process of moving into the cabin, ...
3. The image suggests that the man might have some skill or experience in playing frisbee, ...
4. The image suggests that the surfers are likely on their way to a beach or ...

Feature 7634

1. The group of people dressed in suits, who are likely boys, seem to be gathered for a graduation assembly, and they are sitting down...
2. ... is someone in the driver's seat of the old truck cabin. They seem to be driving the truck, which is not attached to a trailer.
3. The cat seems to be curiously exploring items in the closet area of the home,...
4. ... small plane flying high above the trees. They are gathered together and seem to be enjoying the sight of the aircraft.

Feature 8532

1. One giraffe is extending its neck across the field, appearing to interact or look at the other giraffes.
2. The teddy bear is sitting down, appearing to be reading a book, with the book placed in its lap...
3. The calico cat is lying down in the grass, appearing stealthy and alert, possibly sneaking up on an insect. It seems to...
4. The man appears surprised, and he is making faces at the wall of cuckoo clocks.

Feature 12038

1. ... a recreational activity together. They are playing with a frisbee in what appears to be a backyard setting. The presence of a dining table in...
2. The setting of the image appears to be a restaurant or dining establishment, as evidenced by the presence...
3. ... is traveling on a roadway in a city area, next to what appears to be a Victorian-style building.
4. The setting appears to be a gathering or a party where people are enjoying drinks....

Feature 7928

1. ... one notable feature is the presence of many bars on it. This may suggest the use of architectural elements to create an interesting visual effect...
2. Yes, there is a tag in the display, which might contain information about the floral arrangements, their creator, or their inspiration. The...
3. ... as seen from the overhead view in the image. This organized layout might suggest that the area is well-maintained and possibly set up by a...
4. ... is set in a big room with a brick wall visible. This may be an indoor space, such as a party hall, conference room, or...

Feature 8007

1. ... standing next to his luggage: Travel arrangements: The man might be discussing his travel plans or coordinating travel-related logistics. He could be ...
2. ... table. Some possible explanations include: Wine tasting: The man might be participating in a wine tasting event, assessing and comparing the flavors, ...
3. ... his personal belongings, such as clothes, toiletries, and souvenirs. The cardboard box might hold items that could not fit into the suitcases or require special ...
4. ... Some possible reasons include: Arrival or departure: The man might have just arrived in the city or is about to depart and ...

Feature 11807

1. ... rock-covered beach, looking out at the ocean while holding an umbrella. They could be using the umbrella for various reasons, such as to shield themselves...
2. The woman could be carrying a tray full of carrots in this setting for various...
3. The man could be gesturing near the hazard sign to bring attention to it or...
4. The young boy could be wearing the oversized shirt for various reasons, such as playing dress-up,...

Feature 10439

1. ... man's stern facial expression and crossed arms body language in the image could indicate several possible reasons. It is essential to understand that although these ...
2. ... where the standing cow presses its head against the lying cow's head, could be attributed to several reasons. Some possibilities include social bonding, grooming, or...
3. ... jackets and riding horses on the flattened sandy area, likely a beach, could be a team of mounted police officers patrolling or conducting a special...
4. The gathering of bikers, as seen in the image, can have multiple potential reasons. It could be a casual meetup of friends...

Feature 12063

1. While enjoying their wine, the two men in the image are potentially bonding over a shared interest in watching a football game, as evidenced...
2. In a typical home setting, a person reaching under a bed could potentially find a variety of items, such as: Lost or ...
3. The two sandwiches displayed in an appealing manner on a plate could potentially be for a lunch, dinner, or an appetizer at a gathering or...
4. ... actively playing a game. Additionally, numerous books are present throughout the room, potentially indicating that the man may also enjoy reading, and he could have...

D SAE Feature Evaluation Metrics

We provide detailed definitions of the SAE feature-quality metrics used in Section 4.2. All metrics are computed on the held-out evaluation split using an external CLIP encoder $\phi = (\phi_V, \phi_T)$. For feature $j \in \{1, \dots, p\}$, let $\mathcal{V}_j$ and $\mathcal{T}_j$ denote the vision and text samples on which the feature activates, respectively, and let $a_{s,j} \geq 0$ denote its activation on sample s. We ℓ_2-normalize the CLIP embeddings so that their inner products correspond to cosine similarity.

D.1 Intra- and Inter-Similarity

We adopt Intra-Similarity and Inter-Similarity from Shen et al. (2026). For each feature j, we collect its top-n_0 activating examples from each modality. Let $\mathcal{V}_j^{(n_0)}$ and $\mathcal{T}_j^{(n_0)}$ denote these sets. We first mean-pool their CLIP embeddings,

$$u_j^V = \frac{1}{|\mathcal{V}_j^{(n_0)}|} \sum_{s \in \mathcal{V}_j^{(n_0)}} \phi_V(s), \qquad u_j^T = \frac{1}{|\mathcal{T}_j^{(n_0)}|} \sum_{s \in \mathcal{T}_j^{(n_0)}} \phi_T(s), \tag{8}$$

and normalize the pooled representations,

$$\bar{u}_j^V = \frac{u_j^V}{\|u_j^V\|_2}, \qquad \bar{u}_j^T = \frac{u_j^T}{\|u_j^T\|_2}. \tag{9}$$

Let $\mathcal{H}$ denote the set of features that are active in both modalities and let $H = |\mathcal{H}|$. Intra-Similarity measures cross-modal semantic consistency within the same feature,

$$\mathrm{Sim}_{\mathrm{intra}} = \frac{1}{H} \sum_{j \in \mathcal{H}} (\bar{u}_j^V)^\top \bar{u}_j^T, \tag{10}$$

while Inter-Similarity measures cross-modal similarity across distinct features,

$$\mathrm{Sim}_{\mathrm{inter}} = \frac{1}{H(H-1)} \sum_{j \in \mathcal{H}} \sum_{\substack{\ell \in \mathcal{H} \\ \ell \neq j}} (\bar{u}_j^V)^\top \bar{u}_\ell^T. \tag{11}$$

Higher Intra-Similarity indicates that the vision and text examples associated with the same feature are semantically consistent, while lower Inter-Similarity indicates better separation between different features. In our implementation, we use the top $n_0 = 50$ activating examples per modality.

D.2 Multimodal Monosemanticity

We adopt multimodal monosemanticity (MMS) from Kaushik et al. (2026). MMS measures whether a feature activates on semantically similar examples across modalities, while weighting each pair by the strength of the feature's activations.

For feature j and modalities $m, n \in \{V, T\}$, let $\boldsymbol{a}_j^{(m)} \in \mathbb{R}^M$ and $\boldsymbol{a}_j^{(n)} \in \mathbb{R}^N$ contain the non-zero activations of feature j on its activating samples from the two modalities. We compute the pairwise CLIP cosine-similarity matrix

$$S_{rq} = \phi_m(s_r)^\top \phi_n(s_q), \tag{12}$$

and the corresponding co-activation matrix

$$A_{rq} = \left| a_{j,r}^{(m)} a_{j,q}^{(n)} \right|. \tag{13}$$

For within-modality comparisons, self-pairs are excluded. We normalize the co-activation weights,

$$\widetilde{A}_{rq} = \frac{A_{rq}}{\sum_{r',q'} A_{r'q'}}, \tag{14}$$

and define the feature-level score as

$$\mathrm{MMS}_j(m, n) = \sum_{r,q} \widetilde{A}_{rq} S_{rq}. \tag{15}$$

Thus, MMS is the co-activation-weighted average semantic similarity of the examples associated with a feature. If feature j has no activating samples in one of the two modalities, its cross-modal score is $\mathrm{MMS}_j(V,T) = 0$, so modality-split features are penalized rather than excluded.

We report the dictionary-level averages

$$\mathrm{MMS}_{v\text{-}t} = \frac{1}{p}\sum_{j=1}^{p}\mathrm{MMS}_j(V,T),\ \mathrm{MMS}_{v\text{-}v} = \frac{1}{p}\sum_{j=1}^{p}\mathrm{MMS}_j(V,V),\ \mathrm{MMS}_{t\text{-}t} = \frac{1}{p}\sum_{j=1}^{p}\mathrm{MMS}_j(T,T). \tag{16}$$

$\mathrm{MMS}_{v\text{-}t}$ measures cross-modal semantic coherence, while $\mathrm{MMS}_{v\text{-}v}$ and $\mathrm{MMS}_{t\text{-}t}$ measure within-modality coherence. Higher values indicate greater semantic coherence.

E ADDITIONAL MA-SAE EXPERIMENTS

E.1 CROSS-MODAL COVERAGE AND MATCHED-BUDGET COMPARISON

Table 2 reports MMS over the features scored by each method. MMS assigns zero cross-modal score to features that are active in only one modality, whereas for MA-SAE the score is computed over its designated cross-modal bank. Because this bank comprises only 40% of the full dictionary, this raises the question of whether MA-SAE's advantage reflects a smaller evaluation set rather than higher feature quality. We therefore report three complementary comparisons in Table 10. *Coverage* is the fraction of the full dictionary that activates in both modalities. *Bimodal MMS* averages $\text{MMS}_{v\text{-}t}$ only over features active in both modalities, while *Matched MMS* restricts each method to the same number of highest co-activating features as MA-SAE: 8,192 for LLaVA-1.5-7B, 10,240 for LLaVA-1.5-13B, and 7,168 for InternVL3-8B.

As shown in Table 10, baseline coverage is already high (87–100%), so few baseline features are penalized with a zero cross-modal score, while MA-SAE is capped at 40% by construction. MA-SAE remains best under both the bimodal and matched MMS comparisons across all three backbones. This shows that its advantage is not explained by differences in cross-modal coverage or feature budget, but instead reflects higher-quality cross-modal features.

Method	Coverage (%)	Bimodal MMS ↑	Matched MMS ↑
LLaVA-1.5-7B			
Vanilla SAE	93.1	0.1796	0.1897
VL-SAE	99.8	0.1610	0.1717
GSAE	99.0	0.1763	0.1914
MA-SAE (ours)	39.9	**0.2400**	**0.2386**
LLaVA-1.5-13B			
Vanilla SAE	91.3	0.1828	0.1943
VL-SAE	93.3	0.1711	0.1777
GSAE	98.3	0.1793	0.1949
MA-SAE (ours)	39.9	**0.2416**	**0.2407**
InternVL3-8B			
Vanilla SAE	99.6	0.1864	0.2003
VL-SAE	87.5	0.1698	0.1783
GSAE	99.4	0.1847	0.2002
MA-SAE (ours)	39.9	**0.2402**	**0.2387**

Table 10: Cross-modal coverage, bimodal-feature MMS quality, and matched-budget MMS comparison. MA-SAE remains best when $\text{MMS}_{v\text{-}t}$ is averaged only over features active in both modalities and when the number of evaluated features is matched across methods.

E.2 Component Ablation

We ablate the two components introduced by MA-SAE: modality-partitioned banks and Barlow Twins alignment. Starting from the vanilla SAE, we first add the modality-partitioned banks and then the Barlow Twins alignment objective to obtain the full MA-SAE. Table 11 reports feature-quality metrics across all three MLLMs. Adding modality-partitioned banks improves feature quality over the vanilla SAE, with the strongest performance achieved when combined with Barlow Twins alignment in the full MA-SAE.

Method	$\text{MMS}_{v\text{-}t}$ ↑	$\text{MMS}_{v\text{-}v}$ ↑	$\text{MMS}_{t\text{-}t}$ ↑	Intra-Sim.↑	Inter-Sim.↓
LLaVA-1.5-7B					
Vanilla SAE	0.1599	0.4670	0.6180	0.2912	0.2656
+ Modality Banks	0.1733	0.5033	0.6158	0.3016	0.2780
+ Barlow Twins **(MA-SAE)**	**0.2386**	**0.6509**	**0.6960**	**0.3513**	**0.2313**
LLaVA-1.5-13B					
Vanilla SAE	0.1589	0.4586	0.6172	0.2934	0.2638
+ Modality Banks	0.1726	0.4972	0.6187	0.3022	0.2758
+ Barlow Twins **(MA-SAE)**	**0.2407**	**0.6522**	**0.6982**	**0.3526**	**0.2318**
InternVL3-8B					
Vanilla SAE	0.1843	0.5319	0.6310	0.3043	0.2705
+ Modality Banks	0.1829	0.5224	0.6266	0.3103	0.2785
+ Barlow Twins **(MA-SAE)**	**0.2387**	**0.6496**	**0.6950**	**0.3509**	**0.2303**

Table 11: Ablation of MA-SAE components across three MLLMs.

F DPO AND MDPO TRAINING DETAILS

Preference-Pair Construction. For the off-policy experiments, each VLFeedback prompt contains between four and eight candidate responses generated by different MLLMs. For each prompt, we randomly sample two candidate responses and score both using the corresponding preference-labeling method. We construct exactly one preference pair per prompt, assigning the higher-scoring response as preferred and the lower-scoring response as rejected. In the case of an exact score tie, the preferred response is selected at random. For the on-policy experiments, we sample four candidate responses per prompt from the current policy at temperature 1. We then randomly sample two responses from this candidate set and construct one preference pair using the same procedure described above. All compared preference-labeling methods again operate on the same sampled candidate pairs.

Training Configuration. Table 12 summarizes the shared training configuration and objective-specific hyperparameters for DPO and mDPO.

Hyperparameter	**Value**
Training Settings	
Optimizer	AdamW
Learning rate	1×10^{-5}
LR schedule	WarmupDecayLR
Warmup steps	93
Training epochs	3
Total optimizer steps	937
Global batch size	32
Adam β_1, β_2	0.9, 0.999
Weight decay	0.01
LoRA rank r	64
LoRA α	128
LoRA dropout	0.05
DPO Settings	
β	0.1
mDPO Settings	
Loss terms	Response + image-conditional + anchor
Term weights	$1:1:1$
β	0.1 for all terms
Negative image	Random crop
Crop area	1%–20% of original image

Table 12: DPO and mDPO training hyperparameters. The same training configuration is used across all backbone models and preference-labeling methods.

G Additional Qualitative Visualizations of MA-SAE Grounding

Figures 5, 6, 7, 8, and 9 provide additional qualitative examples of the MA-SAE grounding signal. Across examples, visually supported concepts receive high image support, while hallucinated concepts receive substantially lower support.

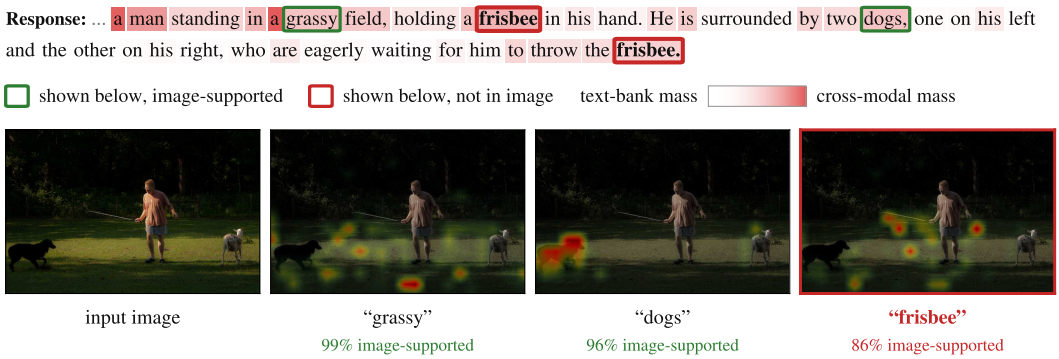


Figure 5: Additional qualitative visualization of MA-SAE grounding. "Grassy" and "dogs" receive strong image support, whereas the hallucinated concept "frisbee" receives lower support.

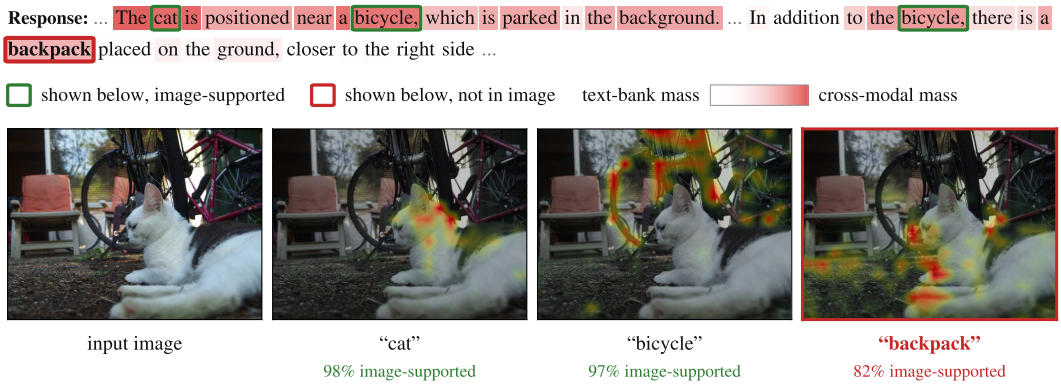


Figure 6: Additional qualitative visualization of MA-SAE grounding. "Cat" and "bicycle" receive strong image support, whereas the hallucinated concept "backpack" receives lower support.

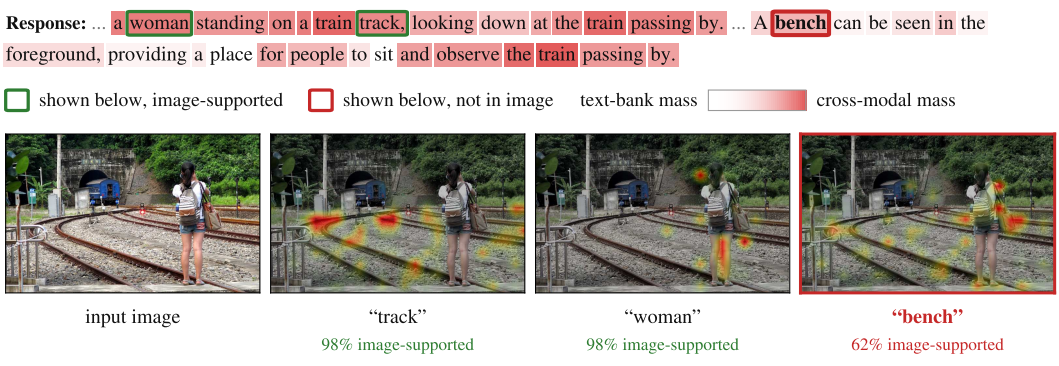


Figure 7: Additional qualitative visualization of MA-SAE grounding. "Track" and "woman" receive strong image support, whereas the hallucinated concept "bench" receives lower support.

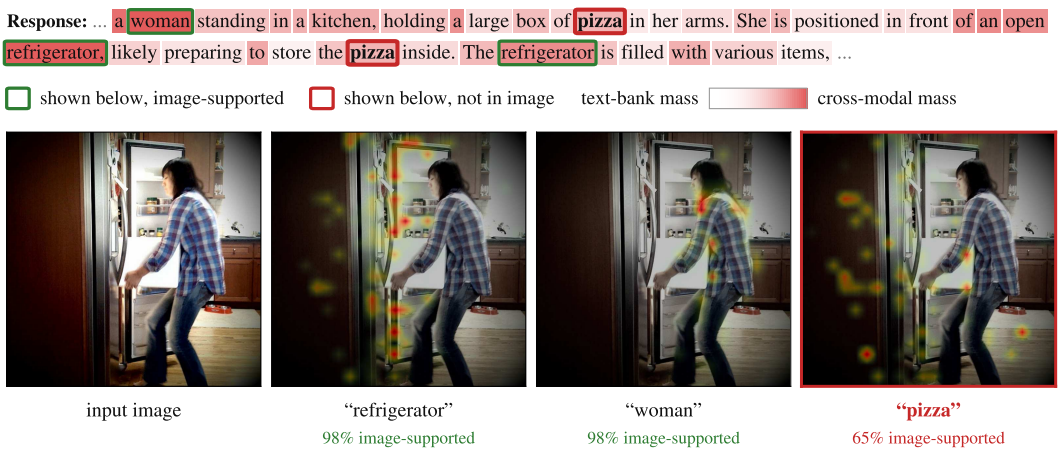


Figure 8: Additional qualitative visualization of MA-SAE grounding. "Refrigerator" and "woman" receive strong image support, whereas the hallucinated concept "pizza" receives lower support.

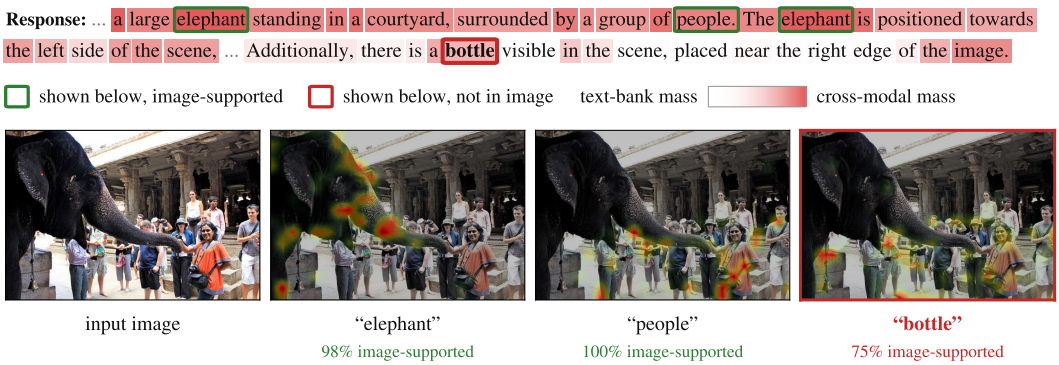


Figure 9: Additional qualitative visualization of MA-SAE grounding. "Elephant" and "people" receive strong image support, whereas the hallucinated concept "bottle" receives lower support.

H Additional Qualitative Comparison of MA-SAE and LLM-as-Judge on AMBER

Figure 10 provides additional AMBER examples comparing LLM-as-Judge DPO and MA-SAE DPO. Across these examples, MA-SAE removes hallucinated objects or attributes and produces responses that are more closely grounded in the image.

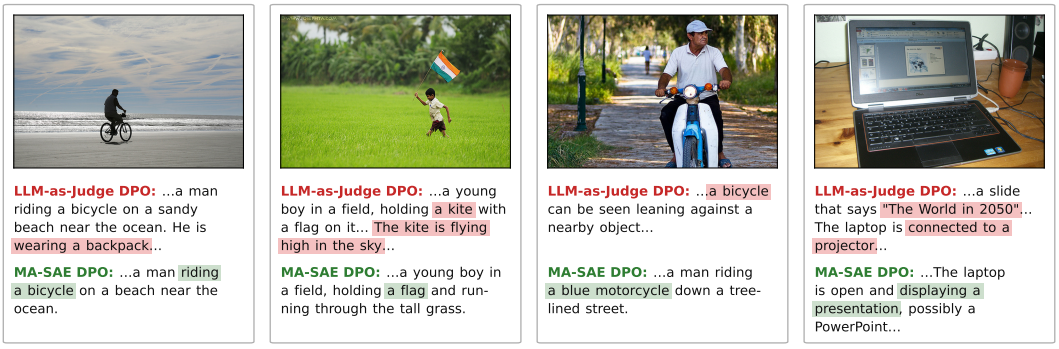


Figure 10: Additional AMBER examples comparing LLM-as-Judge DPO and MA-SAE DPO. MA-SAE removes hallucinated objects or attributes and produces more visually grounded responses.

I Ablation of Modality-Specific Banks for Preference Optimization

To evaluate whether the modality-specific banks are critical for downstream preference optimization, we compare on-policy DPO performance using the full MA-SAE against a variant that removes the vision- and text-specific banks while retaining the same grounding-score formulation. As shown in Table 13, removing the modality-specific banks consistently degrades performance across all three MLLM backbones. These results indicate that explicitly isolating modality-specific information is important for constructing a reliable cross-modal grounding signal for preference labeling.

Method	AMBER				Object HalBench		CHAIR-MSCOCO	
	CHAIR↓	Cover↑	HalRate↓	Cog↓	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$
LLaVA-1.5-7B								
w/o modality-specific banks	6.4	50.9	28.6	2.9	42.33	11.51	50.20	16.71
MA-SAE	**5.1**	**51.0**	**23.8**	**1.9**	**23.00**	**7.11**	**33.40**	**11.93**
LLaVA-1.5-13B								
w/o modality-specific banks	5.4	50.7	25.4	2.4	40.33	11.40	47.80	16.43
MA-SAE	**4.5**	**52.4**	**20.9**	**1.5**	**25.33**	**6.91**	**34.80**	**11.23**
InternVL3-8B								
w/o modality-specific banks	7.3	**71.5**	60.9	5.7	27.00	7.93	34.80	11.69
MA-SAE	**5.9**	71.1	**45.4**	**3.9**	**19.33**	**5.62**	**27.40**	**9.59**

Table 13: Ablation of modality-specific banks under on-policy DPO. Removing the vision and text specific banks consistently degrades performance across all three MLLM backbones, demonstrating that explicitly separating modality-specific information is critical for constructing an effective grounding signal.

J Response Length and Object Coverage Analysis

Because the MA-SAE grounding score measures the fraction of response-activated cross-modal features supported by the image, a potential concern is that it may favor shorter or more conservative responses that mention fewer objects. To examine this possibility, we analyze responses generated after DPO and compare MA-SAE with LLM-as-Judge in terms of response length and object-level content.

For each generated response, we measure the number of words, the number of distinct object mentions, and the number of distinct object mentions that are grounded in or hallucinated using the image ground-truth annotations. Table 14 reports the average of these quantities across evaluation examples.

Across all three backbones, MA-SAE produces shorter responses than LLM-as-Judge. However, this reduction in length is not accompanied by a comparable reduction in grounded object content. For LLaVA-1.5-7B, the average response length decreases from 82.8 to 45.6 words, while the number of grounded objects remains essentially unchanged (2.87 vs. 2.88) and hallucinated objects decrease from 0.36 to 0.27 per response. Similarly, for LLaVA-1.5-13B, grounded objects increase from 2.88 to 2.97 while hallucinated objects decrease from 0.28 to 0.24. For InternVL3-8B, grounded objects remain nearly unchanged (4.03 vs. 4.04), whereas hallucinated objects decrease from 1.03 to 0.73. These results show that MA-SAE favors more concise responses, but the reduction is concentrated in unsupported content rather than grounded object mentions.

Method	Words	Distinct objects	Grounded objects	Hallucinated objects
LLaVA-1.5-7B				
LLM-as-Judge	82.8	3.23	2.87	0.36
MA-SAE	45.6	3.15	2.88	0.27
LLaVA-1.5-13B				
LLM-as-Judge	74.2	3.16	2.88	0.28
MA-SAE	52.0	3.21	2.97	0.24
InternVL3-8B				
LLM-as-Judge	211.6	5.06	4.03	1.03
MA-SAE	166.4	4.77	4.04	0.73

Table 14: Response length and object coverage under DPO. Values are averaged per response. MA-SAE produces shorter responses than LLM-as-Judge while preserving a similar number of grounded object mentions and reducing hallucinated object mentions across all three backbones.