Bigger Is Not Better: Inverse Scaling and Arbitration Failure in Counterfactual Visual Grounding

Fabian Grob [*1] **Sanghwan Kim** [*1 2 3] **Cordelia Schmid** [4 5] **Zeynep Akata** [1 2 3]

Abstract

Images that violate everyday semantic expectations, such as a dog with five legs, expose a fundamental limitation of multimodal large language models (MLLMs): they may rely more on learned priors than on the visual evidence itself. Although prior work has observed such failures, we still lack a unified account of when they become more severe, why they arise, and how they can be mitigated. This work addresses all three questions. Across six MLLM families and three counterfactual visual grounding benchmarks, we find an inverse scaling trend: as models grow larger, they increasingly rely on language priors over visual evidence (*when*). Linear probing reveals that visual information is faithfully encoded throughout the LLM backbone, yet the model fails to act on it, reflecting an *arbitration failure* rather than a perceptual one. Further analysis shows that language priors are already committed in the early layers before visual signals begin to integrate in the middle of the network, suggesting that the arbitration imbalance must be corrected at the very first layers the LLM processes (*why*). We propose *Visual Arbitration for MLLMs* (`VisArb`), where a sparse autoencoder (SAE) is inserted between the vision backbone and the MLP connector, replacing dense feature maps with sparse, concept-aligned representations that are immediately legible to the arbitration mechanism from the first LLM layer onward (*how*). Our `VisArb` substantially improves counterfactual conflict recognition and reduces language-prior dominance while preserving performance on standard VQA benchmarks.

[*]Equal contribution [1]Technical University of Munich, Munich, Germany [2]Helmholtz Munich, Munich, Germany [3]Munich Center for Machine Learning (MCML), Munich, Germany [4]Inria, École normale supérieure, CNRS, PSL Research University, Paris, France [5]Google Research. Correspondence to: Fabian Grob <fabian.grob@tum.de>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

A central promise of MLLMs is that language generation is grounded in visual evidence (Liu et al., 2023; Alayrac et al., 2022; Li et al., 2023). But grounding is most clearly tested when visual evidence contradicts semantic expectation (Vo et al., 2025; Luo et al., 2024; Lee et al., 2025). In such cases, a dog with five legs may still be described as an ordinary dog, revealing how learned priors can override perception. Recent benchmarks, including VAB (Vo et al., 2025), ViLP (Luo et al., 2024), and VLind-Bench (Lee et al., 2025), have formalized this problem as *counterfactual visual grounding*: the ability to follow contradictory visual evidence rather than defaulting to a learned linguistic prior. However, these benchmarks primarily reveal the symptom. They do not yet explain when the problem intensifies, why it emerges, or how it can be systematically mitigated. We address all three questions.

We begin with the most basic question: can scaling up model size resolve this failure? We evaluate 22 checkpoints from six MLLM families, namely Gemma 3 (Gemma Team, 2025), InternVL3 (Zhu et al., 2025), LLaVA-1.5 (Liu et al., 2024), LLaVA-OV (Li et al., 2024), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a), on all three benchmarks. Strikingly, language-prior bias becomes *worse* with scale across the majority of benchmarks and architectures, aligning with observations of inverse scaling in language models where stronger priors amplify undesirable behaviors rather than correct them (Lin et al., 2022; McKenzie et al., 2023; Wei et al., 2023). This rules out scale as a remedy and calls for a deeper structural diagnosis.

We next ask why this failure occurs. Linear probing shows that original and counterfactual images remain nearly perfectly separable across LLM layers, even when the model answers incorrectly. The issue is therefore not perceptual but *arbitrational*: the visual difference is represented, but not used. Chain-of-embedding analysis (Long et al., 2025) further shows that fine-grained visual evidence becomes meaningfully integrated only in the middle layers. By contrast, concept-vector steering (Ferrando et al., 2026; Panickssery et al., 2023; Zou et al., 2023) with GemmaScope2 (Google DeepMind, 2025) indicates that the language prior is established earlier: interventions are most effective in the

early layers and become increasingly marginal thereafter. Together, these results suggest a timing mismatch between prior formation and visual integration: the model commits to the expected concept before the counterfactual evidence can effectively revise the prediction.

This mechanistic understanding directly motivates our approach. Since the language prior is committed too early for inference-time corrections to be effective, the remedy must intervene before the LLM ever processes a token. Therefore, we propose Visual Arbitration for MLLMs (`VisArb`), which inserts a sparse autoencoder (SAE) (Cunningham et al., 2023; Pach et al., 2025; Lim et al., 2024; Zaigrajew et al., 2025) between the vision backbone and the LLM projector, replacing dense visual feature maps with sparse, concept-aligned representations. By making visual evidence maximally legible to the arbitration mechanism from the first LLM layer onward, the model is equipped to resist language-prior dominance before it can take hold. Applied to LLaVA-1.5 (Liu et al., 2024), this approach substantially improves visual conflict recognition and language-prior override on VLind-Bench while preserving general VQA capability.

Contributions. (*When*) We document consistent inverse scaling of language-prior bias across six MLLM families, showing that scale alone cannot resolve counterfactual grounding failures. (*Why*) We establish an encoding-arbitration dissociation through linear probing and mechanistic layer analysis, showing that visual information is faithfully encoded yet overridden by the language prior before visual integration occurs. (*How*) We confirm the causal role of early-layer SAE latents via concept-vector steering, and propose `VisArb`, a training-time approach that improves counterfactual grounding without degrading general VQA performance.[1]

2. Related Work

Counterfactual visual grounding benchmarks. VLMs Are Biased (VAB) (Vo et al., 2025) tests counting in seven visual categories where linguistic priors are strong, such as the number of legs on an animal or pieces on a chess board. ViLP (Luo et al., 2024) presents triplets including image, common-knowledge answer, and perceptual answer, which measures how often models privilege prior knowledge over visual evidence. VLind-Bench (Lee et al., 2025) takes a more controlled approach, using a four-stage pipeline to isolate pure language-prior reliance from confounds such as general knowledge limitations or visual-perception difficulty. Together, these benchmarks define the counterfactual visual grounding regime studied in this work.

[1]Code is available at https://anonymous.4open.science/r/vlm-mechanistic-analysis-FFF4/README.md.

Mechanistic interpretability for VLMs. Long et al. (Long et al., 2025) introduce chain-of-embedding analysis (CoE) and identify a visual integration point (VIP) at which multimodal hidden states diverge from text-only ones. We adopt this framework to localize where visual signals enter the residual stream in counterfactual conditions. Zhou et al. (Zhou et al., 2025) (EVA) measure layer-wise Jensen-Shannon divergence between visual and text-only forward passes to identify image-sensitive layers. We extend this to a counterfactual-discriminative variant that contrasts original and counterfactual image passes to reveal where grounding diverges. Nooralahzadeh et al. (Nooralahzadeh et al., 2026) study visual-linguistic conflicts and show that activation patching and full-sequence steering outperform last-token interventions. Ferrando et al. (Ferrando et al., 2026) use sparse autoencoders to identify entity-specific latent features in language models, a methodology we applied to the visual grounding setting.

Sparse autoencoders. Cunningham et al. (Cunningham et al., 2023) demonstrate that SAEs recover interpretable, monosemantic features in language activations, which established the foundation for feature-level mechanistic analysis. On the vision side, Zaigrajew et al. (Zaigrajew et al., 2025) apply hierarchical SAEs to CLIP (Radford et al., 2021), Joseph et al. (Joseph et al., 2025) use SAE-identified features to steer CLIP's ViT, and Pach et al. (Pach et al., 2025) show that BatchTopK SAEs on CLIP ViT improve concept linearity. Venhoff et al. (Venhoff et al., 2025) use GemmaScope SAEs as frozen probes to track how visual features align with language representations across LLM layers, finding convergence in middle-to-later layers. While prior work applies SAEs to analyze vision encoders or LLMs in isolation, we are the first to use SAE structure as a training-time bottleneck in the vision-to-language interface, directly improving multimodal grounding under counterfactual conditions.

3. Motivation: Inverse Scaling of Counterfactual Visual Grounding

A natural hypothesis is that scale may resolve language-prior bias: larger models, trained on more data, may learn to weight visual evidence more reliably when it conflicts with prior expectations. To test this, we evaluate 22 checkpoints from six model families, namely Gemma 3 (4B, 12B, 27B), InternVL3 (1B-78B, 6 sizes), LLaVA-1.5 (7B, 13B), LLaVA-OV (0.5B, 7B, 72B), Qwen2.5-VL (3B-72B, 4 sizes), and Qwen3-VL (2B-32B, 4 sizes), on three counterfactual visual grounding benchmarks: VAB (Vo et al., 2025), ViLP (Luo et al., 2024), and VLind-Bench (Lee et al., 2025). Since VLind-Bench does not provide unified accuracy and bias metrics, we construct VLind-OE, an open-ended variant using its counterfactual images paired with LLM-generated

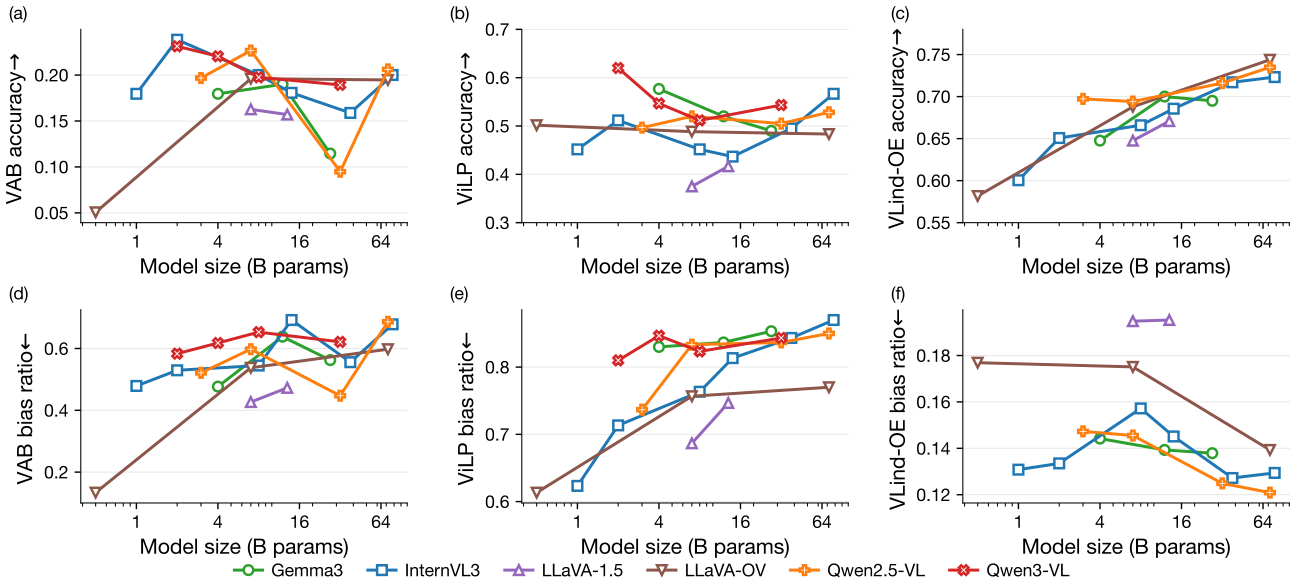

Figure 1. **Inverse scaling of visual grounding across VAB, ViLP, and VLind-OE.** *Row 1* (accuracy, ↑): *(a)* VAB; *(b)* ViLP; *(c)* VLind-OE. *Row 2* (bias ratio, ↓): *(d)* VAB; *(e)* ViLP; *(f)* VLind-OE. On VAB and ViLP, bias ratio increases consistently with scale while overall accuracy declines. On VLind-OE, accuracy improves with scale yet bias ratio does not reliably decrease.

questions, enabling direct comparison with VAB and ViLP. Full construction details are in Appendix E and benchmark descriptions are in Section 5.1.

Figure 1 reveals a striking pattern. On VAB and ViLP, task accuracy remains flat or decreases with scale while bias ratio rises consistently across all families, indicating that larger models are *more*, not less, susceptible to language priors. On VAB, bias ratio increases by 0.20 points from InternVL3-1B to InternVL3-78B and by 0.17 points from Qwen2.5-VL-3B to Qwen2.5-VL-72B, with similar trends across Qwen3-VL and Gemma 3. Accuracy correspondingly drops by up to 0.13 points, and no family makes meaningful progress on the task itself. ViLP bias ratio follows the same pattern, rising from roughly 0.62-0.74 at the small end to 0.84-0.87 at the large end across most families.

VLind-OE presents a more nuanced picture where accuracy improves with scale while bias ratio remains elevated or non-monotonic. This divergence reflects a fundamental difference in what each benchmark demands perceptually. VAB and ViLP probe object- and attribute-level counterfactuals, such as a horse with five legs or a drone with eight propellers, which require close attention to fine-grained visual details. VLind-Bench instead targets scene-level commonsense counterfactuals, such as swans in a desert, where the individual objects are perceptually salient and easier to combine into a correct response even from a coarse visual reading. Yet even in this more relaxed setting, bias ratio does not reliably decrease with scale, and no configuration achieves consistent improvement on both metrics simultaneously. Full numerical results are in Appendix D.

The consistency of this inverse trend across model families and architectures suggests that the effect is unlikely to originate from a single training dataset or implementation choice. Instead, it points to a shared failure mode that is not resolved by scale alone.

4. Mechanistic Analysis and Visual Arbitration for MLLMs (`VisArb`)

In this section, we first investigate why counterfactual grounding fails through three complementary analyses: establishing whether the failure is perceptual or arbitrational, localizing the visual integration point within the LLM backbone, and identifying which latent features mediate the arbitration decision (Section 4.1). This mechanistic understanding directly motivates `VisArb`, which replaces dense visual features with sparse, concept-aligned representations so that the LLM's arbitration mechanism receives clearer visual evidence from the first layer onward (Section 4.2).

4.1. Mechanistic Analysis

Linear Probing. We first ask whether the failure stems from the model not encoding counterfactual visual information at all. We train logistic regression probes on the last-token hidden state $\mathbf{h}_l \in \mathbb{R}^d$ at layer l to predict a visual attribute label $y \in \mathcal{Y}$, using 5-fold cross-validation with pair-level group assignment to prevent leakage between original and counterfactual pairs. $\mathcal{Y}$ is a multi-class attribute (e.g., leg count, stripe count) for VAB while it is binary factual versus counterfactual classification for VLind. We probe Gemma 3 4B

at 14 layers across the full backbone. High probe accuracy on incorrectly answered samples would indicate that the failure is not perceptual but *arbitrational*: visual evidence is encoded yet discarded by a downstream mechanism. Full probe design is in Appendix A.

Chain-of-Embedding Analysis. The next question is at which layer visual information actively enters the residual stream. We apply Chain-of-Embedding (CoE) analysis (Long et al., 2025), which identifies the *visual integration point* (VIP) by comparing hidden-state trajectories under a real-image pass against a zeroed-pixel (blind) pass. Formally, let $\mathbf{h}_l^{\text{vis}}$ and $\mathbf{h}_l^{\text{blind}}$ denote the last-token hidden states at layer l under the real and blind conditions respectively. We compute the layer-wise cosine distance:

$$d_l = 1 - \frac{\mathbf{h}_l^{\text{vis}} \cdot \mathbf{h}_l^{\text{blind}}}{\|\mathbf{h}_l^{\text{vis}}\| \, \|\mathbf{h}_l^{\text{blind}}\|}, \tag{1}$$

separately for vision-dependent samples D_{VT}, whose model output changes when the image is zeroed, and language-prior samples D_{T}, whose output is unchanged regardless of image content. The layer at which d_l begins to grow marks the point where visual information first competes with the language prior, which identifies the window in which grounding could in principle occur.

Concept Vector Steering. We then identify which latent features mediate the arbitration and confirm their causal role via concept-vector steering. For each SAE latent j at layer l, we compute a separation score following Ferrando et al. (Ferrando et al., 2026):

$$s_j^{(l)} = f_{\text{vis},j}^{(l)} - f_{\text{blind},j}^{(l)}, \tag{2}$$

where $f_{\text{vis},j}^{(l)}$ and $f_{\text{blind},j}^{(l)}$ are activation frequencies under the real-image and blind conditions respectively, using GemmaScope 2 16K SAEs (Google DeepMind, 2025). An asymmetric noise filter excludes latents firing on more than 2% of blind samples. To confirm causality, we steer the top-ranked latent per layer by adding $\alpha \cdot \mathbf{d}_j^{(l)}$ to the residual stream, where $\mathbf{d}_j^{(l)}$ is the SAE decoder direction and α controls the magnitude and sign of the intervention. Positive α amplifies visual reliance while negative α suppresses it. If these latents genuinely mediate the arbitration decision, performance changes should be layer-dependent and sign-consistent. Full details are in Appendix B.

4.2. `VisArb` Architecture

The mechanistic analyses above suggest that the arbitration imbalance originates before meaningful visual integration occurs, so any effective remedy must intervene at the vision-to-language interface itself rather than at inference time. Thus, we propose `VisArb`, which inserts a sparse autoencoder (SAE) between the vision backbone and the MLP connector, replacing dense feature maps with sparse, concept-aligned activations (Figure 2).

Sparse Autoencoders. An SAE (Cunningham et al., 2023; Bricken et al., 2023) maps a dense activation $\mathbf{x} \in \mathbb{R}^d$ to a sparse latent vector $\mathbf{z} \in \mathbb{R}^m$ ($m \gg d$) via:

$$\mathbf{z} = \text{TopK}(\text{ReLU}(W_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{enc}})), \tag{3}$$

retaining only the K largest activations. The decoder reconstructs the input as $\hat{\mathbf{x}} = W_{\text{dec}}\,\mathbf{z} + \mathbf{b}_{\text{dec}}$. Each non-zero entry z_j activates a single monosemantic concept direction, and because only $K \ll m$ latents fire per input, the representation is interpretable by construction. This sparsity is what makes SAE features better suited to the LLM's arbitration mechanism than dense visual features.

Standard MLLM Pipeline. A standard MLLM (Figure 2(a)) passes dense patch-level features directly to the LLM. Formally, let $\mathbf{V} = f_{\text{vis}}(I) \in \mathbb{R}^{N \times d_v}$ denote the feature matrix produced by a frozen vision encoder for image I, where N is the number of patch tokens and d_v is the vision hidden dimension. A learned MLP connector g_{proj} projects these into the LLM embedding space:

$$\mathbf{H} = g_{\text{proj}}(\mathbf{V}) \in \mathbb{R}^{N \times d_l}, \tag{4}$$

where d_l is the LLM hidden dimension. The LLM then processes the concatenation $[\mathbf{H}; \mathbf{E}_{\text{text}}]$ of visual and text token embeddings to generate the output.

`VisArb` Encode-Only. In the encode-only variant (Figure 2(b)), we insert the frozen SAE encoder between the vision encoder and the connector. For each patch i, the dense visual feature $\mathbf{v}_i \in \mathbb{R}^{d_v}$ is replaced by a sparse concept vector:

$$\mathbf{z}_i = \text{TopK}(\text{ReLU}(W_{\text{enc}}(\mathbf{v}_i - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{enc}})) \in \mathbb{R}^m. \tag{5}$$

A new connector $g'_{\text{proj}} : \mathbb{R}^m \to \mathbb{R}^{d_l}$ projects these sparse activations directly into the LLM embedding space, $\mathbf{H} = g'_{\text{proj}}(\mathbf{Z})$. This maximally exposes the LLM to individual concept activations.

`VisArb` Encode-Decode. In the encode-decode variant (Figure 2(c)), the sparse activations are decoded back to vision space before projection, which yields a softer bottleneck:

$$\hat{\mathbf{v}}_i = W_{\text{dec}}\,\mathbf{z}_i + \mathbf{b}_{\text{dec}} \in \mathbb{R}^{d_v}, \qquad \mathbf{H} = g'_{\text{proj}}(\hat{\mathbf{V}}), \tag{6}$$

where g'_{proj} operates on the reconstructed vision space. This variant allows the connector to operate on a familiar feature space. It still benefits from the sparsity-induced denoising of the SAE round-trip.

Training. We use LLaVA-1.5 (Liu et al., 2024) as our testbed, as its simple MLP connector and open-source codebase make the bottleneck insertion straightforward, and its

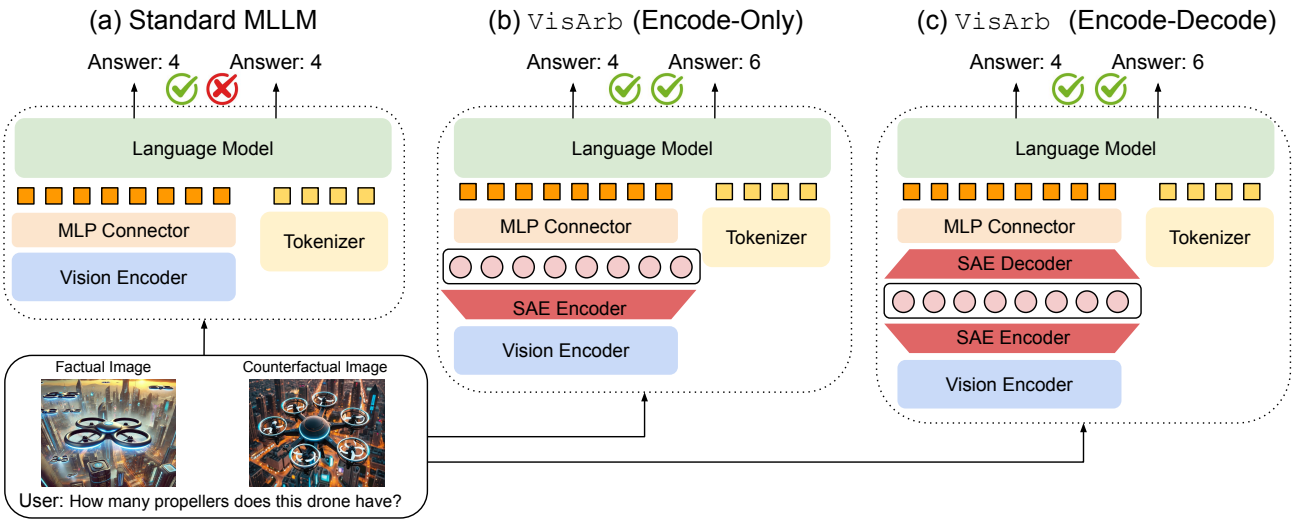

Figure 2. (a) Standard MLLMs pass dense visual feature maps through an MLP connector into the language model. (b) `VisArb` (Encode-only) replaces dense vision features with sparse, concept-aligned SAE activations before the connector, creating a stronger visual arbitration bottleneck. (c) `VisArb` (Encode-decode) decodes the sparse activations back into visual space before the connector, creating a softer bottleneck that preserves compatibility with the original visual representation.

two available sizes (7B, 13B) allow us to examine how grounding improvement scales with model capacity. We insert a frozen BatchTopK SAE ($m = 8192$, $K = 20$) trained on CLIP ViT-L/14-336 layer-22 activations (Pach et al., 2025). All parameters of the vision encoder f_{vis} and the SAE ($W_{\text{enc}}, W_{\text{dec}}, \mathbf{b}_{\text{enc}}, \mathbf{b}_{\text{dec}}$) are frozen; only the new connector g'_{proj} and the LLM weights are updated via standard instruction tuning with cross-entropy loss. We evaluate SAEs trained on two corpora (ImageNet (Deng et al., 2009) and CC3M (Changpinyo et al., 2021)+LAION (Schuhmann et al., 2022)) to assess the effect of training data coverage on downstream grounding. Full training details are provided in Appendix F.

5. Experimental Results

We first describe the benchmarks used for our analysis and evaluation (Section 5.1), then present mechanistic evidence establishing the nature and localization of the arbitration failure (Section 5.2), and finally report the grounding improvements achieved by `VisArb` (Section 5.3).

5.1. Benchmark Details

VAB (Vo et al., 2025) presents counterfactual images across seven visual domains that deliberately violate language priors, paired with direct visual questions requiring image reading rather than prior recall. We report *accuracy* (↑) and *bias ratio* (↓), and construct **VAB-pairs**: 528 matched factual/counterfactual image pairs across the Animals, Logos, and Flags domains for mechanistic analysis.

ViLP (Luo et al., 2024) pairs each of 300 questions with three images: one prior-matching (image 1) and two visual-grounding images (images 2-3). We use the harder *without-fact* variant, which omits any common-knowledge statement from the prompt. We report *ViLP score* (mean accuracy on images 2-3, ↑) and *ViLP bias ratio* (accuracy on image 1, ↓).

VLind-Bench (Lee et al., 2025) applies a four-stage cascade to isolate pure language-prior reliance: (1) **CK** tests world knowledge; (2) **VP** tests visual perception; (3) **CB** adds a textual scaffold; (4) **LP** removes it. Only instances passing all three prior stages contribute to S_{LP}, ensuring that failures at LP reflect a pure inability to override the language prior. Benchmark details are available in Appendix E.

5.2. Mechanistic Analysis of Arbitration Failure

Encoding-arbitration dissociation. Linear probing confirms that the failure is not perceptual. AUC exceeds 0.99 from layer 0 on VAB despite near-floor task accuracy, and reaches 0.99 by layer 14 on VLind (Figure 3). The dissociation is sharpest for VAB Animals ($n_{\text{success}} = 0$) where the model produces the same prior-consistent answer on every failure sample, yet probe AUC remains near 1.0 throughout. Visual information is faithfully present at every layer but it is simply never acted upon during generation. These findings are consistent with related work on visual-linguistic conflict (Nooralahzadeh et al., 2026; Vo et al., 2025).

Arbitration window. CoE analysis places the visual integration point (VIP) at layers 14-19 across all four datasets (Figure 4), consistently in the middle third of the network. D_{VT} rises faster and to higher values than D_{T}, indicating

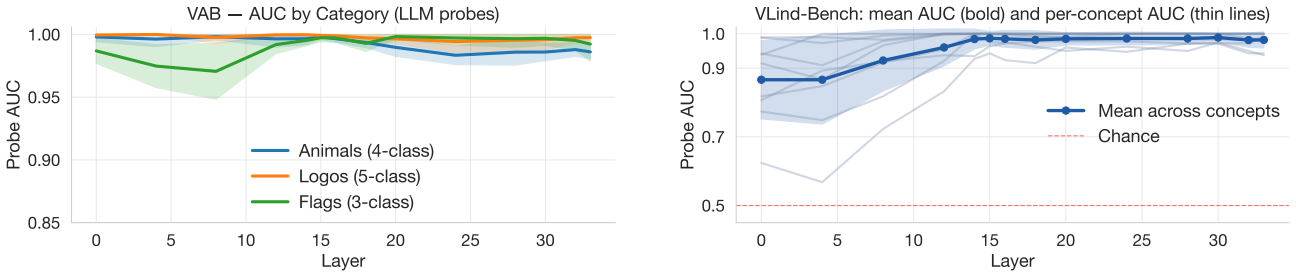

Figure 3. **Linear Probe AUC on LLM Hidden States (Gemma 3 4B).** Probes achieve near-perfect AUC at every layer across VAB and VLind-Bench, including on samples the model answers incorrectly, confirming an *encoding-arbitration dissociation*.

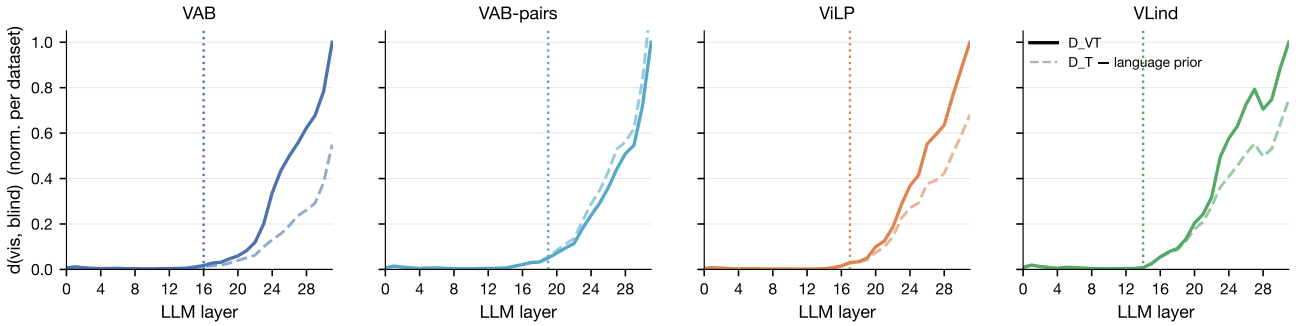

Figure 4. **Chain-of-Embedding (CoE) analysis on Gemma 3 4B.** All curves are flat in early layers and rise steeply from the visual integration point (VIP, dotted vertical line) onward into layers 14-19. Vision-dependent samples (D_{VT}, solid) consistently accumulate more visual signal than language-prior samples (D_{T}, dashed).

that vision-dependent samples accumulate stronger visual signal. Complementary EVA analysis (Zhou et al., 2025) shows that output distributions begin diverging between image conditions only from layer 22 onward on VLind and ViLP, and remain near zero on VAB-pairs throughout (Appendix C). Together, these results bound the arbitration window to layers 14-22: visual information has entered the residual stream but has not yet shaped the output distribution.

Causal role of early-layer latents. Concept-vector steering confirms that the arbitration mechanism is causally accessible in the early layers. Layers 4 and 8 produce substantially larger accuracy changes in both directions than all later layers, which remain nearly flat across the full α range (Figure 5, top). On VAB and VAB-pairs, layers 4 and 8 reduce bias ratio by 49 and 72 percentage points respectively; the effect decays sharply to below 8 pp by layer 16 and approaches zero by layer 20. VLind shows a moderate early signal (+12 pp at layer 4), while ViLP shows weak effects ($\leq$ 5 pp), suggesting its task structure requires broader interventions than single-latent steering. Crucially, the language prior is already committed at layers 4-8 well before visual integration at layers 14-19, meaning any inference-time correction arrives too late. This confirms the design target identified in Section 4.1 and motivates the training-time approach.

5.3. Visual Arbitration for MLLMs (`VisArb`) Results

The encode-only CC3M variant is the strongest configuration across both model sizes. For the 13B model, it improves VLind S_{LP} from 0.24 to 0.46 (+22 pp) and $\mathrm{Acc}_{\mathrm{LP}}$ from 0.52 to 0.64 (Figure 6 (top)). VAB bias ratio and ViLP score are essentially unchanged across all variants: the fine-grained counting demanded by VAB and the format of ViLP present obstacles that concept-level feature changes alone cannot overcome. ImageNet-trained SAEs consistently underperform CC3M+LAION counterparts, suggesting that broader concept coverage in the SAE training data matters for downstream grounding.

The VLind pipeline breakdown reveals where the improvement originates (Figure 6 (bottom)). For the 13B model, S_{VP} is already high at baseline (0.86); the bottleneck's main contribution is to S_{CB} (0.35 $\rightarrow$ 0.67, +32 pp), the stage measuring conflict recognition between visual evidence and the language prior. S_{LP} improves accordingly. For the 7B model, the bottleneck raises S_{VP} to 0.81, surfacing visual evidence that was previously inaccessible, but S_{CB} and S_{LP} remain near zero: resolving the visual-prior conflict requires model capacity that only develops at larger scale.

Performance on standard benchmarks (AI2D, MMBench, MMMU, MMStar, TextVQA) is preserved within $\pm$0.02 of baseline for all variants and both sizes (Figure 7), confirming

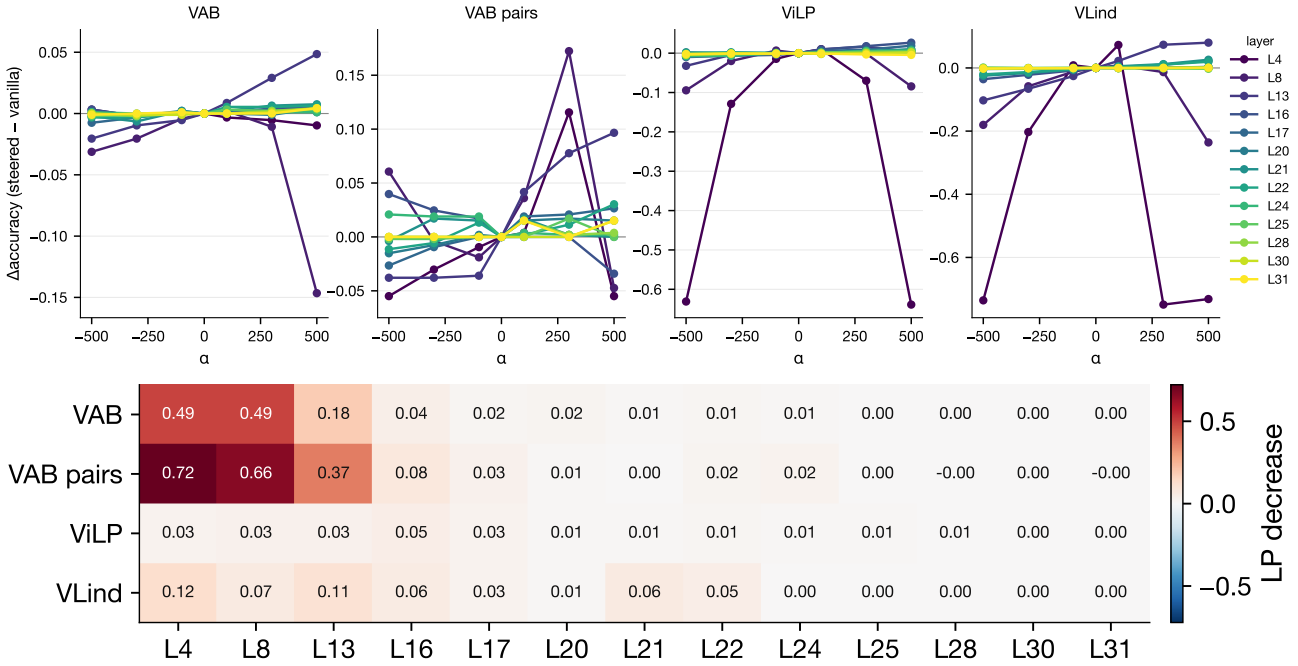

Figure 5. **Concept-vector steering on Gemma 3 4B confirms an early causal arbitration window.** *(Top)* Accuracy change (Δacc) vs. steering strength α per layer; early layers (dark purple) produce far larger changes than later layers (yellow). *(Bottom)* Best language-prior reduction per layer; the effect is concentrated in layers 4-8 and decays sharply thereafter.

that counterfactual grounding improvements do not come at the cost of general VQA capability.

Mechanistic corroboration. To confirm that behavioral improvements reflect a genuine mechanistic change, we apply CoE analysis to both the LLaVA-1.5 13B baseline and the best SAE variant (encode-only CC3M). The SAE model shows substantially stronger visual propagation across all layers and benchmarks (Figure 8). On VAB-pairs, the baseline keeps both D_{VT} and D_{T} near zero; the SAE model develops a clear gap with D_{VT} peaking at ≈ 0.20. On ViLP and VLind, $d(\mathrm{vis}, \mathrm{blind})$ is uniformly higher throughout the network. The EVA analysis shows little change between models, ruling out a shift in intermediate output predictions as the source of improvement. The gains in S_{CB} and S_{LP} therefore stem from stronger visual representation in the residual stream: the bottleneck makes visual evidence more accessible to the arbitration mechanism, and the model acts on it more reliably as a result.

Why SAE features help arbitration. The behavioral and mechanistic results raise a conceptual question: why does replacing dense visual features with sparse SAE activations make visual evidence more legible to the arbitration mechanism?

A dense visual feature $\mathbf{v}_i \in \mathbb{R}^{d_v}$ encodes a superposition of visual concepts entangled in a single continuous vector. When projected into the LLM's embedding space, this mixture must compete with text representations already structured around discrete, language-aligned concepts. The early-layer arbitration mechanism operates on precisely such concept-level representations: steering individual SAE latents at layers 4-8 produces large behavioral changes, while steering dense residual-stream directions at the same layers has negligible effect.

Sparse SAE activations disentangle this mixture. Each non-zero entry z_j activates a single monosemantic concept direction $W_{\mathrm{dec},j}$, and $\mathbf{z}_i$ is an explicit additive sum of $K = 20$ such directions. The LLM therefore receives a small set of discrete concept activations whose magnitudes directly reflect concept strength in the image – a format that is structurally closer to how the LLM represents language and more actionable at the layers where the arbitration decision is made. The CoE corroboration confirms this: the SAE model propagates visual information more strongly from the earliest layers onward, not because the vision encoder sees differently, but because the information is formatted in a way the LLM can act on sooner.

6. Conclusion and Discussion

This work addresses three questions about counterfactual visual grounding in MLLMs: when does the failure worsen, why does it arise, and how can it be mitigated? Our findings converge on a coherent picture. Evaluating 22 checkpoints

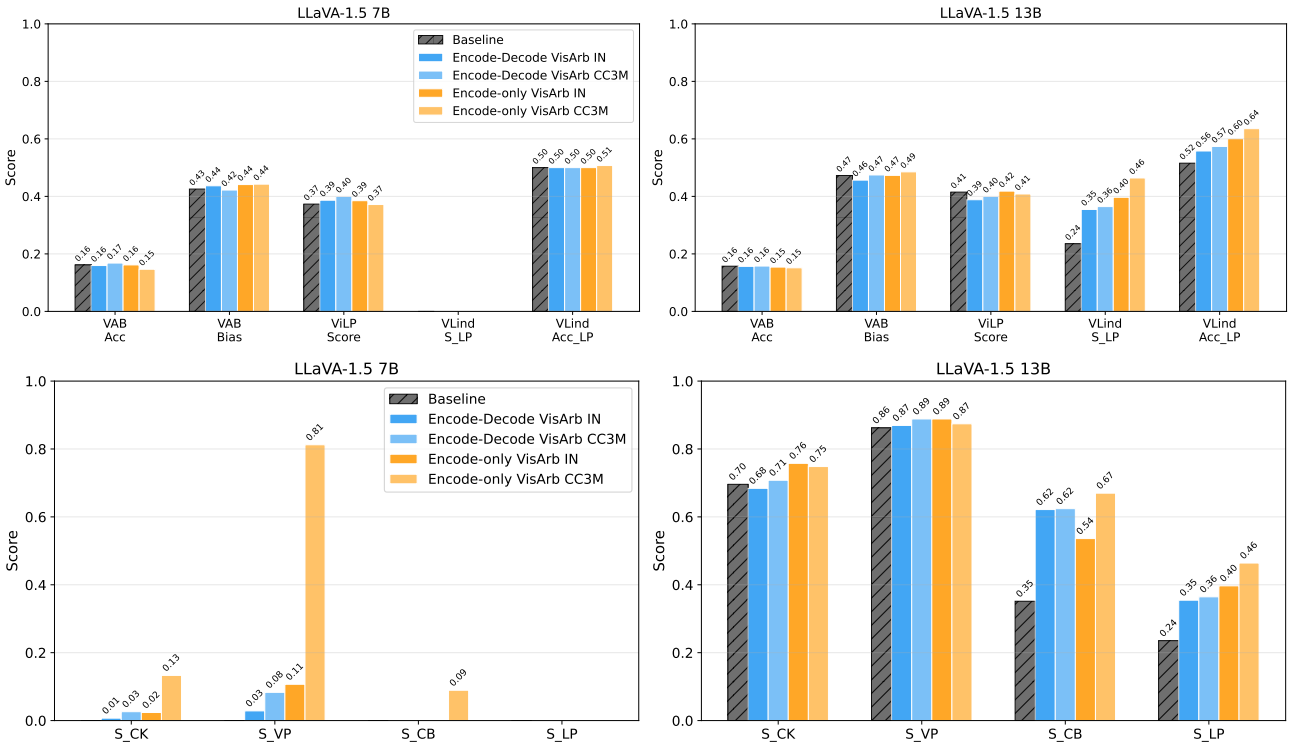

Figure 6. Results of LLaVA-1.5 with `VisArb`. **(Top) Counterfactual grounding.** Each panel shows one model size; bars compare the baseline against four SAE variants (encode-only vs. encode-decode × ImageNet vs. CC3M+LAION). The 13B encode-only CC3M variant achieves clear gains on VLind S_{LP}; VAB and ViLP metrics remain flat across all variants. **(Bottom) VLind-Bench pipeline stage scores.** For the 13B model, the main gain is at S_{CB} (conflict recognition), with S_{LP} improving accordingly. For 7B, S_{VP} improves substantially but S_{CB} and S_{LP} remain near zero.

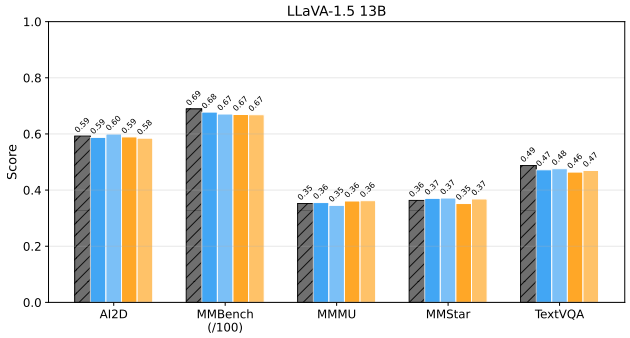

Figure 7. **Standard benchmark results for LLaVA-1.5 with `VisArb`.** All SAE variants remain within ±0.02 of the baseline on AI2D, MMBench, MMMU, MMStar, and TextVQA, for 13B (7B results in Appendix G).

across six model families, we find that language-prior bias grows consistently with scale. Mechanistic analysis reveals the underlying cause that visual information is faithfully encoded at every layer, yet the language prior commits in the early layers (4-8) before visual signals integrate in the middle of the network (layers 14-19), leaving the arbitration effectively decided before visual evidence becomes available. `VisArb` addresses this at training time by replacing dense visual feature maps with sparse, concept-aligned representations that are immediately legible to the arbitration mechanism. Applied to LLaVA-1.5 13B, this yields substantial gains in conflict recognition (S_{CB} +32 pp) and language-prior override (S_{LP} +22 pp) while preserving performance across standard VQA benchmarks.

Limitations and future directions. Our study is limited by the available counterfactual visual grounding benchmarks, which are largely machine-generated and may not reflect the ambiguity of real visual data. Some edits create borderline cases, where it is unclear whether an implausible object should retain its original label. These datasets are therefore controlled stress tests rather than deployment proxies. Medical imaging would offer a more realistic setting, where anomalies are common and deviations from learned priors can be consequential, but large-scale expert-validated benchmarks of this kind are not yet available.

Acknowledgements

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and

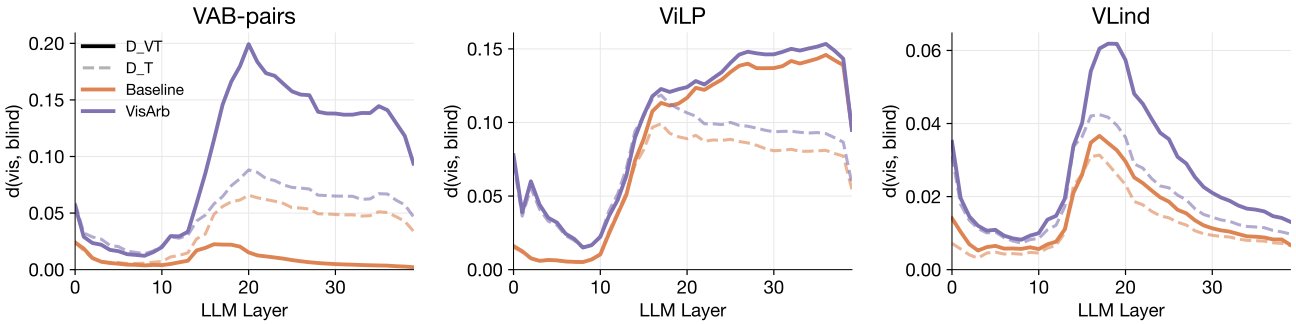

Figure 8. **CoE analysis on LLaVA-1.5 13B: `VisArb` strengthens visual propagation.** The SAE variant (purple) shows higher $d(\text{vis}, \text{blind})$ than the baseline (orange) across all benchmarks and both sample types. On VAB-pairs, the baseline keeps both curves near zero; the SAE model develops a clear gap consistent with the larger share of vision-dependent samples after the bottleneck.

Space.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gausscentre.eu) for funding this project by providing computing time on the Supercomputer JUPITER at Jülich Supercomputing Centre (JSC). This work was supported by the EuroHPC JU and the Gauss Centre for Supercomputing (GCS) through funding by the European Commission, German Federal Ministry of Research, Technology and Space (BMFTR), the Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW).

The authors gratefully acknowledge the scientific support and resources of the AI service infrastructure *LRZ AI Systems* provided by the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities (BAdW), funded by Bayerisches Staatsministerium für Wissenschaft und Kunst (StMWK).

This work was also partially funded by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation, which we thank for their generous support.

Impact Statement

Improving MLLM reliability in counterfactual grounding has direct benefits for safety-critical applications such as medical imaging and quality inspection, where models must respond to visual evidence rather than learned stereotypes. The SAE feature search and steering methodology introduced here could in principle also be used to amplify language-prior reliance. We see no direct path to harmful deployment from the present work, but flag this as a consideration for future research.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-vl technical report. *arXiv*, 2025b.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ferrando, J., Lopez-Cuena, E., Martin-Torres, P. A., Hinjos, D., Arias-Duart, A., and Garcia-Gasulla, D. Language models can explain visual features via steering, 2026. URL https://arxiv.org/abs/2603.22593.

Gemma Team, G. D. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google DeepMind. Gemma scope 2: Technical paper. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf, 2025. Technical report.

Joseph, S., Suresh, P., Goldfarb, E., Hufe, L., Gandelsman, Y., Graham, R., Bzdok, D., Samek, W., and Richards, B. A. Steering clip's vision transformer with sparse autoencoders, 2025. URL https://arxiv.org/abs/2504.08729.

Lee, K.-i., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., and Jung, K. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lim, H., Choi, J., Choo, J., and Schneider, S. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *arXiv preprint arXiv:2412.05276*, 2024.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Long, L., Oh, C., Park, S., and Li, S. Understanding language prior of lvlms by contrasting chain-of-embedding. *arXiv preprint arXiv:2509.23050*, 2025.

Luo, T., Cao, A., Lee, G., Johnson, J., and Lee, H. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*, 2024.

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

Nooralahzadeh, F., Rohanian, O., Zhang, Y., Fürst, J., and Stockinger, K. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts, 2026. URL https://arxiv.org/abs/2604.09364.

nostalgebraist. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong blog post.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.

Venhoff, C., Khakzar, A., Joseph, S., Torr, P., and Nanda, N. How visual representations map to language feature space in multimodal llms, 2025. URL https://arxiv.org/abs/2506.11976.

Vo, A., Nguyen, K.-N., Taesiri, M. R., Dang, V. T., Nguyen, A. T., and Kim, D. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Wei, J., Kim, N., Tay, Y., and Le, Q. Inverse scaling can become u-shaped. In *Proceedings of the 2023 Conference*

on Empirical Methods in Natural Language Processing, pp. 15580–15591, 2023.

Zaigrajew, V., Baniecki, H., and Biecek, P. Interpreting clip with hierarchical sparse autoencoders. *arXiv preprint arXiv:2502.20578*, 2025.

Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J. A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., and Liu, Z. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. 2024. URL https://arxiv.org/abs/2407.12772.

Zhou, H., Zhang, Z., and Chen, H. Extracting visual facts from intermediate layers for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2507.15652*, 2025.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Probe Methodology Details

Probe design. For each benchmark, we define the probe target as follows. **VAB**: binary or 3-class prediction of the ground-truth visual attribute (leg count for animals; logo count; star count bucket for flags). Training data: all samples (original and counterfactual images) labeled with the original ground truth. **VLind**: binary prediction of image condition (factual $= 1$, counterfactual $= 0$). **ViLP**: binary prediction of the correct candidate answer (at chance due to feature design; see main text).

Probes are logistic regression classifiers trained on the last-token hidden state (dimension 2560 for Gemma 3 4B). We use 5-fold cross-validation with pair-level group assignment to prevent leakage between original/counterfactual image pairs.

Limitations of the VLind probe. The factual-vs-counterfactual task conflates image-level difference with concept-level encoding. The two images are visually distinct (different photograph of the same entity in different conditions), so CLIP alone would likely achieve high AUC. The result confirms the model distinguishes the images but does not directly confirm it encodes the semantic concept that the two images differ on.

ViLP probe limitation. The feature vector used for ViLP was $[\mathbf{h}_{\text{layer}}; \mathbf{e}(\text{first token of answer})]$, where $\mathbf{h}_{\text{layer}}$ is the hidden state (identical for all three candidates within a question). This gives the probe no interaction term between the hidden state and the answer candidate, so it cannot distinguish candidates. A bilinear probe or direct use of model logits would be appropriate for ViLP.

B. SAE Feature Search Details

B.1. Full Methodology

We use GemmaScope 2 16K IT SAEs (Google DeepMind, 2025) at medium L0 level, loaded for the residual stream (`resid_post`) of each target layer. The SAE has $d_{\text{SAE}} = 16{,}384$ latents and a BatchTopK sparsity constraint; a latent is considered *active* for a given input if its post-TopK activation value is strictly positive.

For each dataset, we run two forward passes per sample through Gemma 3 4B: a *vis* pass with the real image and a *blind* pass with zeroed pixel values. Hidden states are extracted at the last input token position (the position immediately before generation begins), which integrates context from all preceding tokens and is the standard probe location for next-token prediction behavior. Sample sizes are: VAB (928), VAB-pairs (528, i.e. 264 factual/counterfactual image pairs pooled), ViLP (600, i.e. 300 questions $\times$ 2 visual-grounding images), and VLind (302).

For each SAE latent j and each dataset, we compute activation frequencies under each condition:

$$f_{\text{vis},j} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\text{latent } j \text{ active on sample } i \text{ under vis}], \tag{7}$$

$$f_{\text{blind},j} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\text{latent } j \text{ active on sample } i \text{ under blind}], \tag{8}$$

and define the separation score $s_{\text{visual},j} = f_{\text{vis},j} - f_{\text{blind},j}$.

Asymmetric noise filter. Following Ferrando et al. (Ferrando et al., 2026), we exclude latents that fire on more than 2% of blank-image (blind) samples from the visual-reliance ranking, as such latents respond to text context rather than the image. After filtering, 16,263-16,352 latents (out of 16,384) remain eligible per dataset/layer combination. The filter is *not* applied to the language-prior ranking, because the blind condition is itself the positive signal for prior latents; filtering on it would suppress the targets.

The search is run at 13 layers spanning the full network: 4, 8, 13, 16, 17, 20, 21, 22, 24, 25, 28, 30, and 31. For the steering experiments (Section 5.2), we select the top-ranked visual-reliance latent per layer as the steering direction.

B.2. Per-Layer Top Latents and Scores

Table 1 reports the top visual-reliance latent index and separation score for each layer and dataset. Figure 9 plots the best s_{visual} across layers.

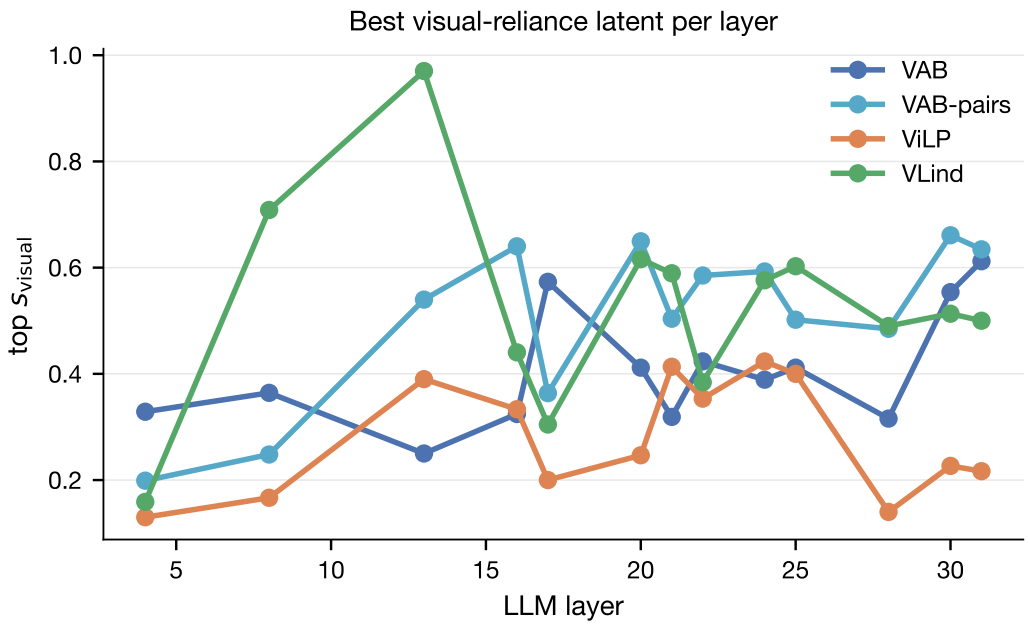

Figure 9. **Best visual-reliance separation score per layer.** Each curve shows the maximum s_{visual} across all latents (after noise filtering) at each probed layer, for each dataset. VLind peaks sharply at layer 13 ($s = 0.97$, latent 12115 activates on 97% of vis samples and 0% of blind samples) before dropping, suggesting a dataset-specific visual-semantic integration event at that layer. VAB and VAB-pairs rise more gradually toward later layers. ViLP scores are consistently lower, reflecting the coarser image-text alignment in that benchmark's triplet format.

B.3. Cross-Dataset Latent Overlap

To test whether visual-reliance representations are shared or task-specific, we compute the pairwise overlap of the top-50 visual latents (by s_{visual}) across datasets at each layer and average over layers. Results: VAB vs. VAB-pairs, 66% (expected, as both use VAB images); all other pairs are below 14%: VAB vs. ViLP, 11%; VAB vs. VLind, 13%; VAB-pairs vs. ViLP, 11%; VAB-pairs vs. VLind, 5%; ViLP vs. VLind, 3%. The near-zero overlap between ViLP and VLind, and the low overlap between VAB and either, confirms the main-text claim: there is no single universal "look-at-image" direction in GemmaScope 16K, and the latents mediating visual reliance are largely task-specific.

C. EVA Analysis

C.1. Method

We adapt the EVA framework of Zhou et al. (Zhou et al., 2025), which applies the logit-lens technique (nostalgebraist, 2020) to localize where visual information begins to shape the model's output distribution. At each layer ℓ, the last-token hidden state is projected through the language model head to produce an intermediate next-token distribution $p^\ell \in \Delta^{|V|}$. The Jensen–Shannon (JS) divergence between distributions for two input conditions then indicates the layer at which the two inputs begin to produce measurably different predictions.

We employ two complementary variants:

Standard EVA (vis $\|$ blind)**.**

$$\mathrm{JSD}^\ell_{\text{std}} = \mathrm{JS}\big(p^\ell_{\text{vis}} \,\|\, p^\ell_{\text{blind}}\big)\,, \tag{9}$$

where p^ℓ_{vis} uses the real image and p^ℓ_{blind} uses zeroed pixels. This measures where the model starts noticing that an image is present at all, as opposed to a blank input.

CF-discriminative EVA (vis $\|$ cf)**.**

$$\mathrm{JSD}^\ell_{\text{cf}} = \mathrm{JS}\big(p^\ell_{\text{vis}} \,\|\, p^\ell_{\text{cf}}\big)\,, \tag{10}$$

where p^ℓ_{cf} uses the matched counterfactual image (e.g. factual vs. counterfactual object for VAB-pairs, or prior-matching vs.

Table 1. **Top-3 visual-reliance latents per layer per dataset.** For each layer and dataset we report the GemmaScope 16K latent indices and separation scores ($s_{\text{visual}} = f_{\text{vis}} - f_{\text{blind}}$) for the three highest-ranked latents after the asymmetric noise filter. Dashes (–) indicate that no additional latent exceeds a separation score of zero. Bolded indices recur across datasets: latent 346 (layer 4, VAB/VAB-pairs), latent 454 (layer 20, VAB/VAB-pairs), latent 13898 (layer 24, VAB/VAB-pairs), latents 1733 and 14598 (layers 30-31, VAB/VAB-pairs). Score decay from rank 1 to rank 3 is notably steep for ViLP and VLind at layer 4, where only one latent clears the noise filter, and for VLind at layer 8, where the top latent (s=0.709) is five times stronger than the next (s=0.149).

VAB

	#1		#2		#3	
Layer	idx	s	idx	s	idx	s
4	**346**	0.329	11890	0.111	237	0.072
8	13659	0.364	15944	0.102	9721	0.095
13	1691	0.250	14241	0.232	3811	0.212
16	12002	0.324	9504	0.295	4931	0.258
17	11149	0.573	4082	0.361	12175	0.296
20	**454**	0.412	4395	0.318	4835	0.262
21	13072	0.319	4395	0.290	11315	0.255
22	5788	0.423	9574	0.347	1950	0.319
24	**13898**	0.389	12650	0.352	15892	0.330
25	9597	0.412	6978	0.319	4888	0.294
28	5290	0.316	5098	0.292	14814	0.278
30	**1733**	0.554	11454	0.408	9101	0.402
31	**14598**	0.612	654	0.471	9083	0.351

VAB-pairs

	#1		#2		#3	
Layer	idx	s	idx	s	idx	s
4	**346**	0.199	11890	0.093	10546	0.027
8	13983	0.248	12207	0.186	11914	0.074
13	15862	0.540	1691	0.383	14878	0.259
16	1702	0.640	9840	0.534	9504	0.379
17	13942	0.364	4082	0.309	13449	0.299
20	**454**	0.650	13497	0.438	4835	0.388
21	2059	0.504	13072	0.456	12223	0.453
22	12943	0.585	13715	0.532	5788	0.519
24	**13898**	0.593	11929	0.559	12322	0.496
25	6978	0.502	10491	0.434	4888	0.434
28	5098	0.485	15538	0.441	9729	0.426
30	**1733**	0.661	11454	0.633	8696	0.580
31	**14598**	0.634	9083	0.555	654	0.547

ViLP

	#1		#2		#3	
Layer	idx	s	idx	s	idx	s
4	11890	0.130	–		–	
8	14552	0.167	10731	0.117	3105	0.117
13	10796	0.390	12825	0.387	15538	0.350
16	9504	0.333	1201	0.277	16110	0.180
17	1204	0.200	6921	0.193	4766	0.177
20	13051	0.247	12707	0.197	11586	0.190
21	15507	0.413	16070	0.327	4210	0.307
22	2848	0.353	15310	0.240	6105	0.233
24	6837	0.423	6658	0.410	6597	0.303
25	5213	0.400	6246	0.310	13643	0.283
28	14283	0.140	4172	0.127	10059	0.120
30	7567	0.227	12268	0.120	15810	0.057
31	2771	0.217	13506	0.117	13501	0.107

VLind

	#1		#2		#3	
Layer	idx	s	idx	s	idx	s
4	9814	0.159	–		–	
8	8769	0.709	14009	0.149	7956	0.046
13	12115	0.970	13940	0.930	6438	0.742
16	9200	0.440	8021	0.371	14228	0.318
17	2597	0.305	8388	0.288	9031	0.238
20	15824	0.616	653	0.460	16280	0.361
21	12388	0.589	14443	0.563	4125	0.490
22	3455	0.384	4277	0.338	2341	0.308
24	15334	0.576	6394	0.497	15073	0.470
25	14920	0.603	5675	0.526	4864	0.507
28	13106	0.490	9875	0.417	14136	0.354
30	6655	0.513	11030	0.503	6004	0.490
31	13170	0.500	14933	0.411	5678	0.374

visual-grounding image for ViLP/VLind). This is a strictly harder probe than the standard variant: both inputs contain a real image, so divergence isolates the layer at which the model begins to distinguish *between* the two semantically different images, rather than merely detecting the presence of any visual content. Near-zero $\text{JSD}^{\ell}_{\text{cf}}$ at all layers would indicate the model's intermediate predictions are fully image-agnostic on the relevant dimension.

C.2. EVA Results with Gemma 3 4B

Figures 10 and 11 reveal a key dissociation. The standard variant (vis vs. blind) shows that Gemma 3 4B does eventually react to the presence of an image in ViLP and VLind, starting around layer 22. The CF-discriminative variant (vis vs. cf) shows that this reaction is *not* driven by distinguishing the specific counterfactual content from the factual: VAB-pairs, where the counterfactual manipulation is a subtle change in object attributes, shows near-zero divergence at every layer. The model's intermediate output distribution treats the factual and counterfactual counting images as identical throughout all 34 layers, despite the model encoding the images differently (probe AUC ≈ 1.0; Section 5.2).

For ViLP and VLind, both variants show the same sharp rise from layer 22, suggesting the CF-discriminative signal in these tasks is dominated by the presence of different visual scenes rather than a subtle feature-level distinction. Together with the

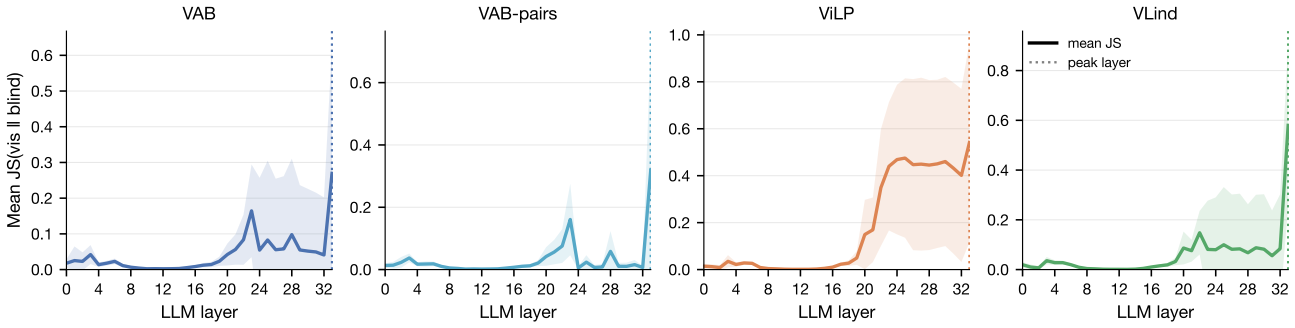

Figure 10. **Standard EVA** (vis || blind) **on Gemma 3 4B.** Mean layer-wise JS divergence between visual and blind (zeroed-pixel) forward passes across four datasets. VAB and VAB-pairs remain very low throughout (peak ≈ 0.25 at layer 33), while ViLP rises steeply from layer 22 (peak ≈ 0.5) and VLind rises sharply from layer 22 to plateau around 0.35-0.4. The late rise for all datasets indicates that the model's intermediate output distribution remains insensitive to visual input through most of the network.

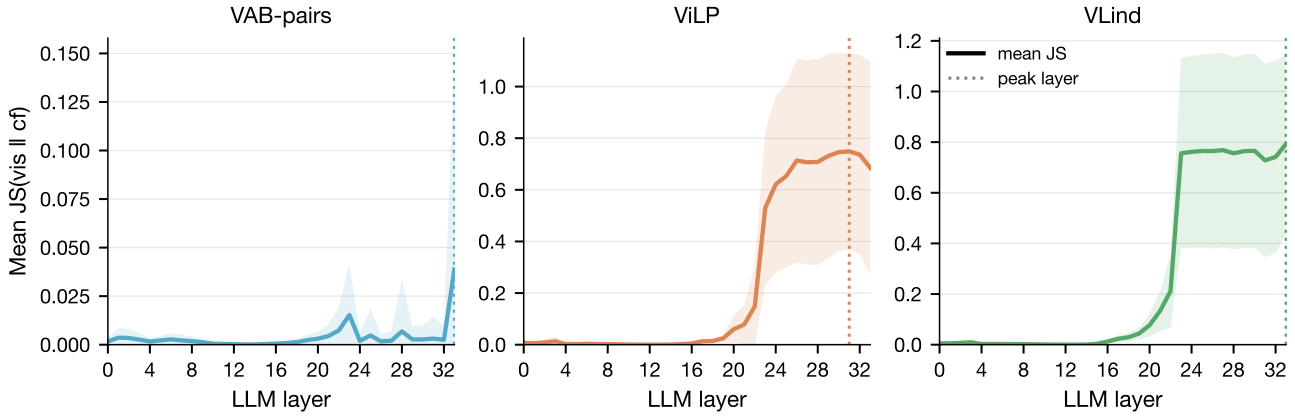

Figure 11. **CF-discriminative EVA** (vis || cf) **on Gemma 3 4B.** Mean layer-wise JS divergence between original and counterfactual image forward passes. VAB-pairs remains near zero throughout (peak ≈ 0.04 at layer 33), indicating the model's intermediate predictions are essentially identical for the factual and counterfactual counting images at all layers. ViLP and VLind show the same sharp rise from layer 22 as in the standard variant, with peaks at layers 31 and 33 respectively.

CoE analysis (Section 5.2), this defines the *arbitration window*: visual information enters the residual stream at VIPs in layers 14-19 (CoE) but does not shape the output distribution until layer 22 (EVA), a gap of roughly 5-8 layers in which the language prior overrides the signal.

C.3. EVA Results with LLaVA-1.5 13B: Baseline vs. SAE Bottleneck

Comparing Figures 12 and 13 for LLaVA corroborates the main-text interpretation. The standard vis-vs-blind variant shows the SAE model elevates JS divergence for VAB-pairs (consistent with CoE showing stronger visual propagation), but the CF-discriminative variant remains near zero for VAB-pairs for both models: even though the SAE model represents visual information more strongly, the bottleneck is not sufficient to make the model's intermediate predictions sensitive to the specific counterfactual attribute manipulation in the counting task. On VLind, where the SAE model achieves a meaningful S_{LP} improvement ($0.24 \rightarrow 0.46$), the CF-discriminative curve is modestly elevated in the early-to-mid layers (14-20), suggesting the bottleneck helps the model begin distinguishing factual from counterfactual images somewhat earlier, consistent with the S_{CB} improvement (Section 5.3). Neither variant shows a change for ViLP, consistent with the absence of behavioral improvement on that benchmark.

C.4. Counterfactual (CF) Chain-of-Embedding Analysis

The standard CoE analysis (Section 5.2) computes $d(\mathrm{vis}, \mathrm{blind})$, measuring how strongly visual information diverges from a blank-pixel baseline. A complementary counterfactual variant replaces the blank input with the matched counterfactual

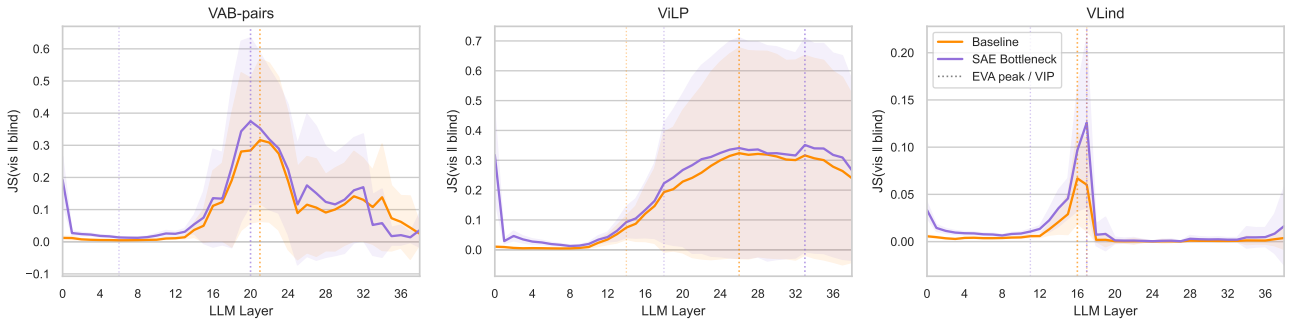

Figure 12. **Standard EVA** (vis || blind) **for LLaVA-1.5 13B baseline vs. SAE enc-only CC3M.** Orange: baseline; purple: SAE model. Dotted lines mark the EVA peak / VIP. On VAB-pairs, the SAE model reaches a slightly higher peak ($\sim$0.35 vs. $\sim$0.30) and an earlier peak layer. On ViLP, both curves overlap closely at a plateau of $\sim$0.28-0.30. On VLind, the SAE model shows a sharper, more localized spike at layer 16 ($\sim$0.13 vs. $\sim$0.05), after which both return toward zero. The elevated VAB-pairs peak is consistent with the CoE finding that the SAE model propagates visual information more strongly; the effect on ViLP is negligible.

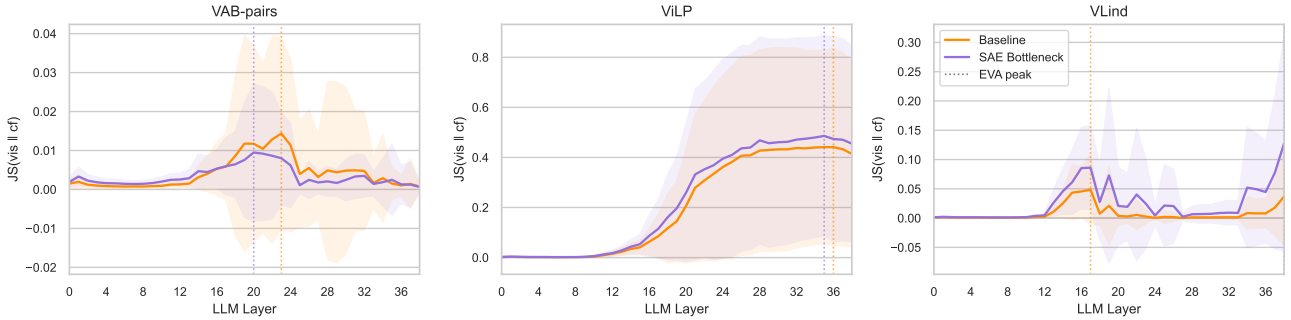

Figure 13. **CF-discriminative EVA** (vis || cf) **for LLaVA-1.5 13B baseline vs. SAE enc-only CC3M.** Orange: baseline; purple: SAE model. On VAB-pairs, both models produce near-zero divergence with high variance, and the SAE model provides no meaningful CF-discriminative signal (the model's output distribution does not change between factual and counterfactual images for this dataset even after the bottleneck intervention). On ViLP, both curves are nearly identical ($\sim$0.4-0.5 plateau from layer 20 onward). On VLind, the SAE model shows modestly higher divergence in layers 14-20 compared to the baseline, consistent with its improved S_{CB} score; both return to near-zero after layer 20.

image:

$$d^{\ell}_{\mathrm{cf}} = 1 - \cos\big(\mathbf{h}^{\ell}_{\mathrm{vis}},\, \mathbf{h}^{\ell}_{\mathrm{cf}}\big)\,, \tag{11}$$

where $\mathbf{h}^{\ell}_{\mathrm{vis}}$ and $\mathbf{h}^{\ell}_{\mathrm{cf}}$ are the last-token hidden states for the factual and counterfactual image, respectively. This measures at which layer the residual stream begins to differentiate between two semantically different real images, rather than merely detecting image presence. Because both inputs contain a real image, d^{ℓ}_{cf} is a strictly harder probe than d^{ℓ}_{std} and its values are typically smaller.

The near-zero d^{ℓ}_{cf} for VAB-pairs at all layers directly corroborates the CF-discriminative EVA result (Figure 11): the failure to distinguish factual from counterfactual counting images is not an artifact of the logit-lens projection used in EVA, but is visible in the raw residual-stream geometry as well. For ViLP and VLind, the residual stream does develop some sensitivity to the counterfactual distinction after the VIP, but this does not translate into behavioral change for the language-prior samples, reinforcing that the arbitration failure occurs downstream of visual encoding.

D. Full Scaling Results

Table 2 reports all evaluated checkpoints and all metrics. ViLP scores use the *without-fact* variant throughout (see Section 3). VLind-Bench reports the full four-stage cascade: S_{CK}, S_{VP}, S_{CB}, and S_{LP}.

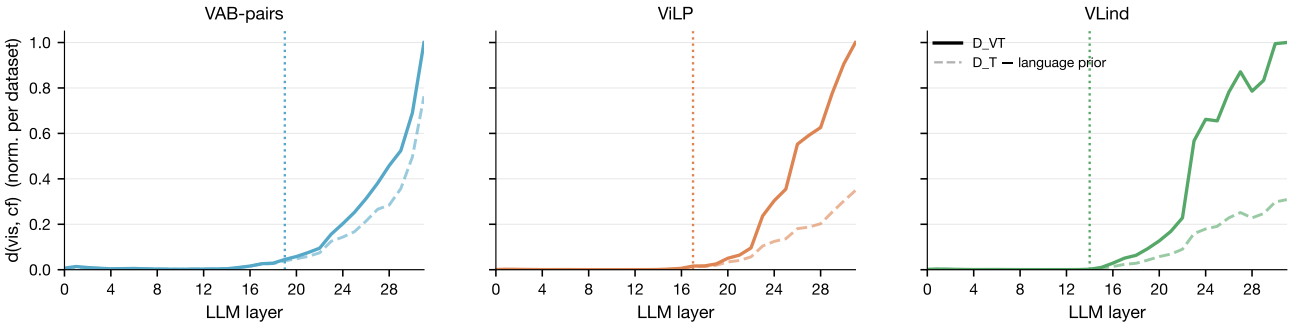

Figure 14. **CF CoE: $d(\text{vis}, \text{cf})$ in Gemma 3 4B.** Layer-wise cosine distance between factual and counterfactual image hidden states, split by vision-dependent (D_{VT}, solid) and language-prior (D_{T}, dashed) samples. VAB-pairs remains near zero throughout all layers, indicating the residual stream does not differentiate between the two counting images at any depth. ViLP and VLind show a gradual rise from the VIP onward, but at lower magnitude than the standard CoE, reflecting the harder nature of distinguishing two real images. The D_{VT} advantage over D_{T} is present for ViLP and VLind, consistent with the behavioral split.

E. Benchmark Details

VLMs Are Biased (VAB). VAB (Vo et al., 2025) contains approximately 930 counterfactual images spanning seven visual domains: animals, logos, flags, chess boards, optical illusions, color charts, and dice. Each image deliberately violates the language prior (e.g., an Adidas logo with four stripes instead of three, an animal with an atypical number of legs). For each domain, a model that reads the image correctly answers based on the visible counterfactual content; a model relying on its prior gives the memorized answer. We report two metrics: *accuracy* (fraction of visually grounded responses, higher is better) and *bias ratio* (fraction of responses matching the language prior, lower is better).

From VAB we additionally construct **VAB-pairs**: matched factual/counterfactual image pairs for the Animals, Logos, and Flags domains (528 pairs total), used to compare hidden-state representations directly in the probe and CoE analyses. VAB-pairs are not evaluated for task accuracy; they serve purely as a controlled contrastive stimulus set.

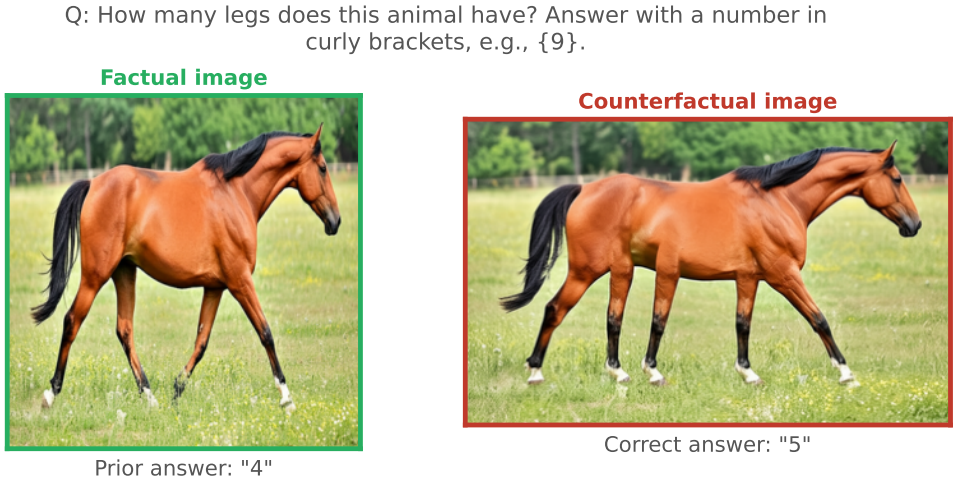

Figure 15. **VAB example: factual/counterfactual animal pair.** Left: the factual image (standard number of legs); right: the matched counterfactual image (atypical number of legs). A visually grounded model answers based on the counterfactual image content; a model relying on its language prior gives the memorized answer corresponding to the factual image. Both images are included in VAB-pairs and used for representation analysis.

ViLP. ViLP (Luo et al., 2024) contains 300 questions paired with 900 images in triplets: one image matching the linguistic prior and two images requiring genuine visual grounding. We use the *without-fact* variant throughout (referred to as ViLP), which withholds the textual statement of the common-knowledge fact so the model must rely solely on the image. We

Table 2. **Full scaling results.** All evaluated checkpoints across six model families. ViLP scores use the without-fact variant. All values multiplied by 100. VAB accuracy and ViLP score: higher is better (↑). VAB bias ratio and ViLP bias ratio: lower is better (↓). VLind-Bench: S_* are cascade-filtered pass rates; acc_{CB} and acc_{LP} are direct stage accuracies (all instances, no cascade conditioning). † LLaVA-1.5-7B answers at chance on all VLind stages; cascade scores shown as 0.

		VAB		ViLP (w/o fact)		VLind-Bench						VLind-OE		
						cascade (S_*)				direct acc				
Family	Scale	acc ↑	bias ↓	score ↑	prior ↓	S_{CK} ↑	S_{VP} ↑	S_{CB} ↑	S_{LP} ↑	acc_{CB} ↑	acc_{LP} ↑	acc ↑	bias ↓	other ↓
Gemma 3	4B	18.0	47.7	57.7	83.0	71.0	89.6	73.9	60.9	86.2	74.1	64.8	14.4	20.8
	12B	19.0	63.8	52.0	83.7	90.3	91.2	80.0	82.7	88.2	87.3	70.0	13.9	16.1
	27B	11.5	56.3	49.0	85.3	92.4	90.5	86.9	82.5	92.9	90.2	69.5	13.8	16.7
InternVL3	1B	18.0	47.9	45.2	62.3	53.4	87.2	72.0	68.0	80.4	71.7	60.1	13.1	26.9
	2B	23.9	53.0	51.2	71.3	75.8	89.3	66.5	63.5	80.2	69.4	65.1	13.4	21.6
	8B	20.0	54.5	45.2	76.3	91.2	90.5	58.9	74.9	72.6	72.1	66.6	15.7	17.7
	14B	18.1	69.2	43.7	81.3	93.6	90.7	72.8	75.2	81.1	74.4	68.6	14.5	16.9
	38B	15.9	55.6	49.5	84.3	95.5	92.9	80.9	75.7	89.2	82.7	71.7	12.7	15.6
	78B	20.0	67.8	56.7	87.0	96.9	93.6	82.8	78.9	91.0	87.2	72.3	12.9	14.7
LLaVA-1.5	7B	16.3	42.7	37.5	68.7	0.0	0.0	0.0	0.0†	50.0	50.0	64.8	19.5	15.8
	13B	15.7	47.3	41.7	74.7	69.4	86.0	35.6	24.0	65.2	51.5	67.1	19.5	13.4
LLaVA-OV	0.5B	5.1	13.4	50.2	61.3	39.4	92.4	54.2	59.8	70.8	66.9	58.2	17.7	24.1
	7B	19.6	53.8	48.8	75.7	90.3	93.6	66.8	67.7	80.2	71.2	68.8	17.5	13.7
	72B	19.5	59.8	48.3	77.0	96.7	93.4	78.4	85.2	89.3	92.2	74.4	13.9	11.7
Qwen2.5-VL	3B	19.7	52.1	49.7	73.7	82.7	90.0	75.9	66.1	86.0	73.3	69.7	14.7	15.5
	7B	22.7	59.8	52.0	83.3	81.0	91.2	56.9	76.2	75.8	73.8	69.4	14.5	16.0
	32B	9.5	44.7	50.5	83.7	96.2	90.7	77.8	76.0	88.5	84.8	71.6	12.5	15.9
	72B	20.6	68.6	52.8	85.0	94.3	90.3	96.2	90.5	97.9	95.5	73.5	12.1	14.4
Qwen3-VL	2B	23.1	58.3	62.0	81.0	71.7	91.0	64.2	73.6	76.0	74.0	–	–	–
	4B	22.1	61.8	54.7	84.7	86.7	89.6	57.5	72.9	70.9	70.4	–	–	–
	8B	19.8	65.3	51.2	82.3	91.9	89.8	60.0	74.9	74.6	72.9	–	–	–
	32B	18.9	62.1	54.3	84.3	93.8	89.3	80.3	77.0	89.2	86.1	74.0	12.9	13.1

report *ViLP score* (`vilp_score`: mean accuracy on the two visual-grounding images) and *ViLP bias ratio* (`vilp_prior`: accuracy on the prior-matching image, i.e. the fraction of responses driven by language priors).

VLind-Bench. VLind-Bench (Lee et al., 2025) isolates language-prior reliance through a four-stage filtering cascade applied per concept instance:

1. **Commonsense Knowledge (CK):** the model is asked the factual knowledge question without any image. Only instances where the model answers correctly proceed.

2. **Visual Perception (VP):** the model is shown the counterfactual image and asked to identify its content. Only instances where the model correctly perceives the counterfactual attribute proceed.

3. **Commonsense Bias (CB):** the model is shown the counterfactual image with a contextual scaffold (a sentence framing the counterfactual scenario). Only instances where the model answers correctly with scaffolding proceed.

4. **Language Prior (LP):** the model is shown the counterfactual image with no scaffolding and asked the same question. S_{LP} is the pass rate on this final stage.

By construction, S_{LP} is computed only on instances where the model has demonstrated world knowledge (CK), visual perception (VP), and scaffold-aided reasoning (CB). A failure at LP therefore cannot be attributed to knowledge gaps or perceptual difficulties; it is a pure failure to override the language prior when facing contradictory visual evidence. The benchmark covers nine concept categories with both factual and counterfactual image variants per instance.

VLind-OE. VLind-OE converts VLind-Bench's true/false cascade into an open-ended format to enable direct comparison with VAB and ViLP via accuracy and bias ratio metrics.

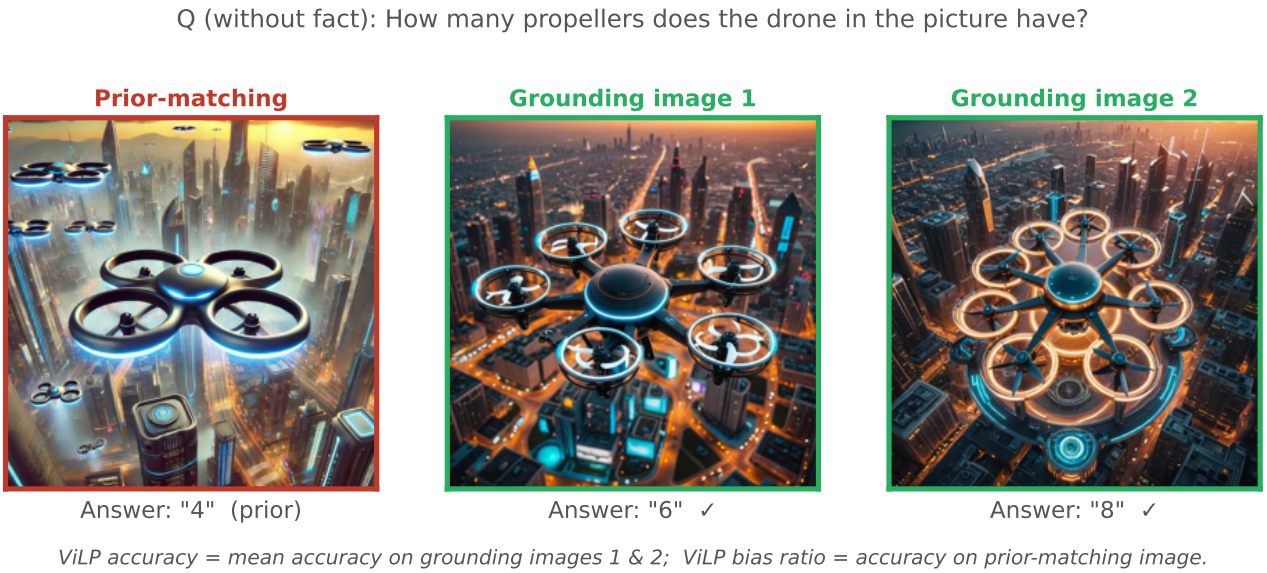

Figure 16. **ViLP example: image triplet.** Each question is paired with three images. Left: the prior-matching image (the scene that corresponds to the expected common-knowledge answer). Center and right: the two visual-grounding images, which require the model to override its prior. In the without-fact variant, no textual fact is provided; the model must read the image to answer correctly. We evaluate accuracy on the two grounding images and bias ratio on the prior-matching image.

Question and answer generation. For each VLind-Bench instance, we prompt Qwen3.5-27B (temperature $= 0$, pinned revision, few-shot with six examples spanning habitat, diet, color, size, landmark, and time concepts) to produce: (1) a short neutral open-ended question answerable from the counterfactual image, (2) 4-6 expected answers matching the counterfactual content, and (3) 4-6 biased answers matching the language prior. The question must not hint that the image contradicts common sense. All answers are constrained to 1-3 words to ensure robust substring matching. Generation is fully deterministic (temperature $= 0$, fixed seed) and resumable; failed instances are retried with a stricter prompt before being discarded.

Manual review. All generated instances are manually inspected to remove ambiguous questions, cases where the expected and biased answer sets overlap, and instances where the counterfactual manipulation is too subtle to be answered reliably from the image alone. Only counterfactual images rated as high-quality by at least two human annotators in the original VLind-Bench release are included. After filtering, the dataset contains 2,233 image-question pairs across the same nine concept categories as VLind-Bench (habitat, diet, color, size, material, pattern, action, landmark, time).

Evaluation protocol. Model responses are scored as *correct* (substring match with any expected answer), *biased* (substring match with any biased answer), or *other*, after normalization (lowercase, punctuation removed, articles stripped). Expected answers are checked before biased answers; a response matching both is labelled correct. We report accuracy (= fraction correct, ↑) and bias ratio (= fraction biased, ↓), following VAB and ViLP.

Evaluation infrastructure. All models are evaluated with greedy decoding via a fork of `lmms-eval` (Zhang et al., 2024). Standard benchmarks (AI2D, MMBench, MMMU, MMStar, TextVQA) use the default `lmms-eval` evaluation protocols without modification.

F. Model and Training Details

Base model. We use LLaVA-1.5 (Liu et al., 2024) in two sizes (7B and 13B). LLaVA-1.5 pairs a CLIP (Radford et al., 2021) ViT-L/14-336 vision encoder with a two-layer MLP projector that maps the 1024-dimensional CLIP output to the LLM input space. The 7B variant uses Vicuna-7B as the language model; the 13B variant uses Vicuna-13B.

SAE architecture and pretraining. We train BatchTopK SAEs from scratch using the architecture and codebase of Pach et al. (Pach et al., 2025), applied to the post-MLP residual stream of layer 22 of CLIP ViT-L/14-336 (the second-to-last

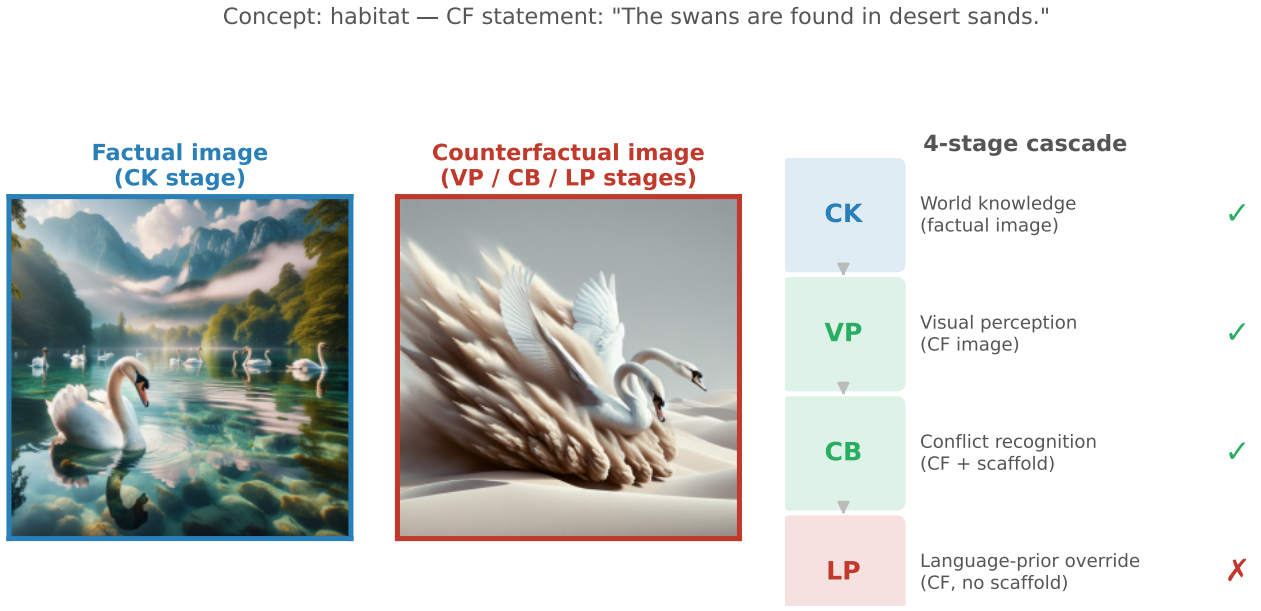

Figure 17. **VLind-Bench example: four-stage cascade.** Left: the factual image used in the CK stage (verifying world knowledge). Right: the counterfactual image used in VP, CB, and LP stages. The LP question (identical to the VP question, without the CB scaffold) is the final test of language-prior override. A model that passes CK, VP, and CB but fails LP demonstrates a pure language-prior bias: it knows the fact, can perceive the counterfactual, and can reason correctly with help, but defaults to the prior when the scaffold is removed.

transformer block). The SAE has an expansion factor of 8, yielding a dictionary of 8192 features with top-$K = 20$ active features per token. Activations are extracted from two training corpora: **ImageNet**: all 1.2M ImageNet-1K training images. **CC3M+LAION**: all Conceptual Captions 3M training images combined with a 200-shard subset of LAION-400M (approximately 2M images), providing a broader, web-scale distribution. In both cases, 2 random patch tokens are sampled per image during extraction, doubling the number of training activation vectors.

SAE training uses the BatchTopK trainer from the `dictionary_learning` library (Pach et al., 2025) with a batch size of 4096 and 100 000 gradient steps. The learning rate follows the scaling law of Pach et al. (Pach et al., 2025): $\text{lr} = e/125\sqrt{d_{\text{SAE}}}$, where $e = d_{\text{SAE}}/d_v$ is the expansion factor, giving $\approx 7.07 \times 10^{-4}$ for $e = 8$, $d_{\text{SAE}} = 8192$. An auxiliary-k loss with $\alpha = 0.03$ prevents dead latents.

Architectural variants. `VisArb` Encode-only: the 8192-dimensional sparse activation vector from the SAE encoder is fed directly into a new projector, replacing the original $1024 \to 4096$ two-layer MLP with a $8192 \to 4096 \to 4096$ network. The SAE decoder is discarded; this enforces a hard bottleneck. **`VisArb` Encode-decode**: the sparse activation vector is passed through the SAE decoder, reconstructing a 1024-dimensional CLIP-space vector, which is then processed by the original projector. This provides a softer bottleneck that preserves the CLIP reconstruction path.

Instruction-tuning pipeline. Both variants follow LLaVA-1.5's two-stage pipeline (Liu et al., 2024). *Stage 1 (projector alignment)*: only the projector is trained on 558K image-caption pairs from LLaVA-Pretrain (BLIP/LAION/CC/SBU mixture) using a learning rate of 1×10^{-3}. For the encode/decode variant the original pretrained $1024 \to 4096$ projector is reused directly; for encode-only a new $8192 \to 4096 \to 4096$ projector is trained from scratch. *Stage 2 (visual instruction tuning)*: both variants are fine-tuned for 1 epoch on the LLaVA-1.5 665K instruction-tuning mixture (`llava_v1_5_mix665k.json`), which combines LLaVA-Instruct-150K with a curated set of academic VQA and reasoning datasets. Optimization uses AdamW with learning rate 2×10^{-5}, cosine decay schedule, 3% linear warmup, and weight decay 0. The global batch size is 128 in both stages (per-device batch 4, gradient accumulation 32 for the 7B model on a single H100; per-device batch 2, gradient accumulation 16, 4 H100s for 13B). Training uses bfloat16 precision with TF32 matmuls and DeepSpeed ZeRO-3. The vision encoder (CLIP ViT-L/14-336) and the SAE weights are kept frozen throughout both stages.

Standard benchmark evaluation. We evaluate on AI2D, MMBench, MMMU, MMStar, and TextVQA using the default

`lmms-eval` protocols. Results are reported in Figure 7; all SAE variants remain within ± 0.02 of the LLaVA-1.5 baseline on all five benchmarks.

G. Standard Benchmark Results

Figure 7 in the main text reports 13B results. Figure 18 below reports the corresponding 7B results. All SAE variants remain within ± 0.02 of the LLaVA-1.5 baseline on all five benchmarks for both model sizes.

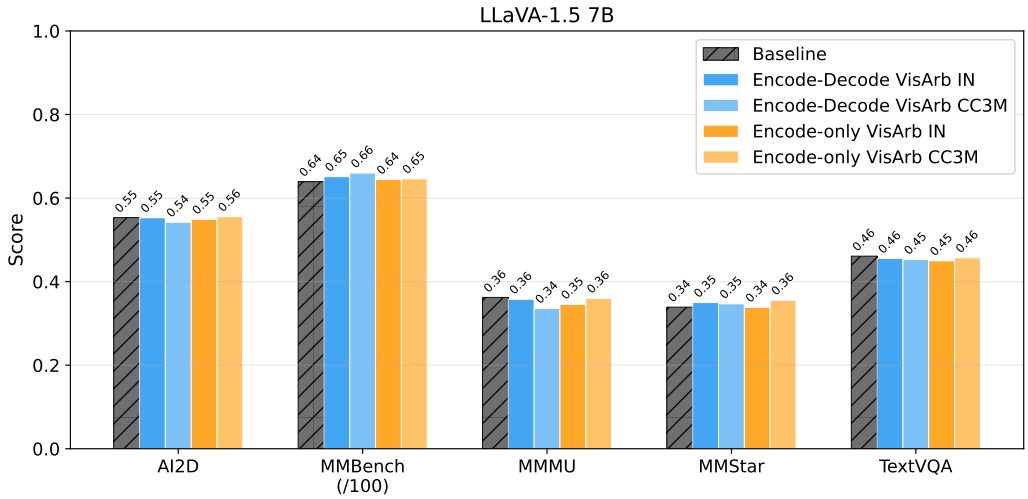

Figure 18. **Standard benchmark results for LLaVA-1.5 7B with `VisArb`.** All SAE variants remain within ± 0.02 of the baseline on AI2D, MMBench, MMMU, MMStar, and TextVQA.

Computational resources. SAE training requires approximately 6-8 hours per model on a single H100 (80 GB); we train two SAEs (ImageNet and CC3M+LAION corpora). Projector alignment (Stage 1, encode-only variant) requires approximately 6 hours on a single H100. Instruction tuning (Stage 2) requires approximately 12 hours on $4\times$H100 for the 13B variant and approximately 12 hours on a single H100 for the 7B variant; four LLaVA variants are trained in total (7B and 13B $\times$ encode-only and encode/decode). Scaling evaluations are lightweight: the full sweep across 22 checkpoints and three counterfactual benchmarks completes in under 48 GPU-hours on a single H100 per model. Mechanistic analyses (CoE, EVA, feature search, steering) each complete in under 2 GPU-hours per dataset.