SEEING OR KNOWING? VISUAL CONTEXT SENSITIVITY IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How reliably can we expect a multimodal large language model (MLLM) to use or disregard an image that conflicts with prior knowledge? We introduce the WHATIFVIS benchmark, adapting and extending counterfactual datasets across five coarse-grained conflict types, such as COLOR and SIZE. Each example holds the image and question fixed while changing only the instruction to use or ignore the image. Across six MLLMs, mean paired accuracy is only 14.1%: success requires ranking the intended answer above the alternative under both instructions. Yet this failure is ambiguous: visual evidence may be missing from the model's representations (**H1**), or remain available but be mishandled (**H2**). Reconstruction models trained without counterfactual examples recover coarse attributes from the final-layer image tokens of three frozen MLLMs, arguing against **H1** for these attributes. LoRA adaptation of the language backbone on one conflict type, with the vision encoder frozen, raises mean paired accuracy to 58.75%, with gains extending to four held-out types. Activation patching identifies layer windows where interventions change whether answers follow the image or prior knowledge. A learned rank-1 intervention transfers from each adapted model to its original, adapter-free counterpart, reaching 34.2% mean paired accuracy without verbal intent instructions. Matched text statements make the same task-relevant facts easier to follow when instructed, providing a comparison that also reflects differences in evidence extraction. MLLMs can retain visual evidence yet fail to control its use. WHATIFVIS exposes this gap and provides a testbed for developing more robust multimodal models. [1]

1 INTRODUCTION

Multimodal large language models (MLLMs) answer visual questions by combining visual evidence with parametric knowledge. When the two conflict, models may report what they know rather than what they see (Golovanevsky et al., 2025; Vo et al., 2026): for example, answering "four" for a five-legged horse despite an instruction to use the image (Figure 1, left).[2] The reverse failure also occurs: a model may rely on visual information when instructed to ignore the image and answer from its parametric knowledge (Figure 1, right). Both failures reflect unreliable control over when to use visual evidence and when to disregard it. This control matters when, for example, an agent encounters images in retrieved documents or files: a request may concern either what an image depicts or general knowledge that an incidental image should not override.

Failures to answer from an image can have different causes. Under **H1** (*missing visual information*), the task-relevant information is not encoded or is lost during visual encoding or propagation through the language backbone. Under **H2** (*mishandling of available information*), the information is encoded in the language backbone, but the model fails to use or suppress it as instructed. End-to-end answer accuracy alone cannot distinguish these possible causes. Prior work examines complementary aspects of this problem. Representation analyses examine what information is encoded (Fu et al., 2025; Pan et al., 2025), visual-counterfactual benchmarks measure whether answers follow the image or parametric knowledge (Golovanevsky et al., 2025; Vo et al., 2026), and text-only work studies

[1] Our code and benchmark will be made publicly available.

[2] This echoes an account of human perception by Hermann von Helmholtz, a nineteenth-century German physician and physicist: sensory impressions are interpreted through unconscious inference informed by past experience (von Helmholtz, 1867).

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.

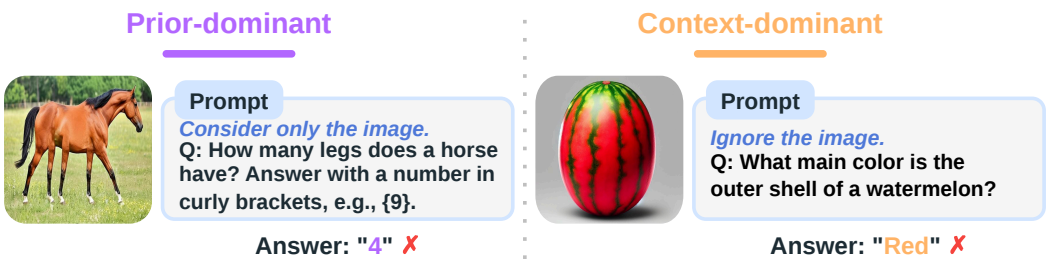


Figure 1: Failure to follow instructions. Left: `Gemma-4-E2B-IT` answers "4" for a five-legged horse despite instructions to use the image. Right: it answers "red" for the watermelon rind despite being instructed to ignore the image and answer from parametric knowledge ("green").

controllable context use when the evidence is explicitly stated (Minder et al., 2025). Visual control requires *connecting* these questions: before asking whether a model can control its use of visual evidence, we must establish that the evidence survives.

We examine **H1** through image reconstruction and **H2** through behavioral evaluation and interventions. We introduce WHATIFVIS, containing 3,125 examples across five coarse-grained conflict types: SPATIAL-TEMPORAL, COLOR, COUNT, SIZE, and WEIGHT.[3] Each example holds the image and question fixed while changing only the instruction to use or ignore the image. We adapt the MetaQueries framework (Pan et al., 2025) to reconstruct images solely from the final-layer image-token representations of frozen MLLMs, then evaluate whether the reconstructions preserve attributes that conflict with prior knowledge (Figure 4). We then extend the paired context-control framework of Minder et al. (2025) to visual inputs, testing whether models can be trained to follow these instructions and whether interventions learned in the adapted models transfer to the original models. Our main metric, *paired accuracy*, requires correct answer rankings under both instructions on the same example (Equation (6)). By default, we use probability version, unless otherwise specified.

Our experiments give three connected findings: **(i)** Coarse counterfactual attributes remain recoverable in three MLLM families: Qwen2.5-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026), and Gemma-4 (Gemma Team, 2026). This argues against complete loss of the reconstructed attributes as an explanation for failures on these visual conflicts (Section 4). **(ii)** Across six MLLMs, mean paired accuracy on WHATIFVIS is only 14.1%. Supervised fine-tuning (SFT) of the language backbone using LoRA (Hu et al., 2022) on one conflict type raises it to 58.75%, with gains also extending to the four conflict types excluded from training (Section 5). **(iii)** Building on work that uses activation interventions to interpret and steer models (Wang et al., 2023; Stoehr et al., 2024), we use activation patching to identify layer regions where interventions change whether adapted models follow the image or their parametric knowledge. A learned rank-1 intervention transfers to the original models with the adapters removed, reaching 34.2% paired accuracy without written intent instructions (Section 5.3). These results show that improving how the language backbone uses retained visual evidence can substantially increase controllability. A single learned direction per model captures part of this improvement, although control remains less reliable than with matched textual evidence. These findings show that WHATIFVIS, combined with reconstruction, exposes a gap between retaining visual evidence and reliably controlling its use, providing a testbed for developing more robust MLLMs.

2 BACKGROUND

Multimodal language models. A common MLLM architecture connects a pretrained large language model (LLM) to a **vision encoder** $\mathcal{E}$ through a **projector** $\mathcal{P}$ (Liu et al., 2023; Bai et al., 2025). Given an image I, the encoder produces patch features $\mathcal{E}(\mathrm{I}) = (\mathbf{v}_1, \ldots, \mathbf{v}_K) \in \mathbb{R}^{K \times d_v}$. The projector maps these features into the language model's D-dimensional embedding space, potentially merging or pooling features, to produce **visual tokens** $\mathcal{P}(\mathcal{E}(\mathrm{I})) = (\mathbf{z}_1, \ldots, \mathbf{z}_M) \in \mathbb{R}^{M \times D}$. The resulting MLLM p predicts each text token $x_t \in \Sigma$ from the visual tokens and preceding text $x_{<t}$:

$$p(x_t \mid \mathcal{P}(\mathcal{E}(\mathrm{I})),\, x_{<t}) = \mathrm{softmax}\big(\mathbf{W}\ \mathrm{Norm}(\boldsymbol{h}^L_{t-1})\big)_{x_t}, \tag{1}$$

[3]We use *coarse-grained* to refer to the object-level attributes and relations tested here, as distinct from fine textures or small visual details. This is an operational distinction rather than a sharp perceptual boundary.

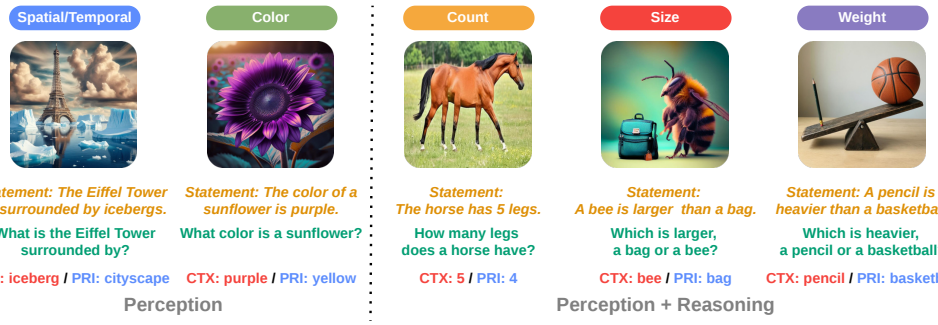


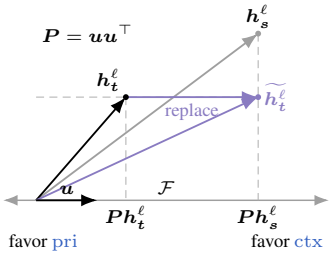
Figure 2: **Illustration of the five WHATIFVIS tasks.** Each image is paired with a text statement describing the same task-relevant fact. Contextual answers (CTX) follow the image; prior answers (PRI) follow parametric knowledge.

where Σ is the token vocabulary, $\boldsymbol{h}_{t-1}^{L} \in \mathbb{R}^{D}$ is the residual-stream state at the end of the conditioning prefix after the final layer L, Norm is the final normalization, and $\mathbf{W} \in \mathbb{R}^{|\Sigma| \times D}$ is the unembedding matrix. Encoder and projector designs vary: pretrained vision encoders include CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023; Tschannen et al., 2025), while connectors include MLPs and resamplers (Liu et al., 2023; Li et al., 2023b; Alayrac et al., 2022). We study Qwen2.5-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026), and the Gemma-4 models (Gemma Team, 2026).

Subspace interventions. In a text-only setting, Minder et al. (2025) identify a learned direction in the residual stream along which interventions change whether a model follows supplied context or its parametric knowledge. Their procedure first uses *activation patching*: replacing activations in a target forward pass with activations from a source forward pass to measure their effect on the answer. At a selected layer, building on ideas from distributed alignment search (DAS) (Geiger et al., 2024), it then learns a one-dimensional intervention at a selected layer. Let $\boldsymbol{h}_{\boldsymbol{s}}^{\ell}, \boldsymbol{h}_{\boldsymbol{t}}^{\ell} \in \mathbb{R}^{D}$ denote the residual-stream states at the last prompt token after layer ℓ for source and target prompts $\boldsymbol{s}$ and $\boldsymbol{t}$. For learning context control, the prompts share the same evidence and question but give opposite instructions about which information source to use. A unit vector $\boldsymbol{u} \in \mathbb{R}^{D}$ defines the subspace $\mathcal{F} = \mathrm{span}\{\boldsymbol{u}\}$ and its rank-1 orthogonal projection $\boldsymbol{P} = \boldsymbol{u}\boldsymbol{u}^{\top}$:

$$\boldsymbol{h}_{\boldsymbol{t}}^{\ell} = (\mathbf{I}_D - \boldsymbol{P})\boldsymbol{h}_{\boldsymbol{t}}^{\ell} + \boldsymbol{P}\boldsymbol{h}_{\boldsymbol{t}}^{\ell}, \qquad (2a) \qquad \widetilde{\boldsymbol{h}_{\boldsymbol{t}}^{\ell}} = (\mathbf{I}_D - \boldsymbol{P})\boldsymbol{h}_{\boldsymbol{t}}^{\ell} + \boldsymbol{P}\boldsymbol{h}_{\boldsymbol{s}}^{\ell}. \qquad (2b)$$

Here $\mathbf{I}_D$ is the identity matrix. Equation (2a) separates the target state into components orthogonal and parallel to $\boldsymbol{u}$; Equation (2b) replaces only the parallel component with that from the source state. The orthogonal component is unchanged by the intervention. The direction $\boldsymbol{u}$ is learned by minimizing the cross-entropy loss for the source answer when the remaining layers process the patched target state, with the model parameters frozen.



Static steering. After learning $\boldsymbol{u}$, we can replace the source-dependent component in Equation (2b) with a fixed value, so steering requires no written intent instruction in the prompt:

$$\widetilde{\boldsymbol{h}_{\boldsymbol{t}}^{\ell}} = (\mathbf{I}_D - \boldsymbol{P})\boldsymbol{h}_{\boldsymbol{t}}^{\ell} + \boldsymbol{u}\, c(w). \qquad (3)$$

Here $w \in \{\mathrm{ctx}, \mathrm{pri}\}$ selects the desired behavior: ctx means use the supplied context, and pri means use parametric knowledge. The function c assigns a model-specific scalar to each intent. The vector $\boldsymbol{u}$ defines the intervention direction, while $c(w)$ sets the target coordinate along that direction. We tune the visual steering scalars on the SPATIAL-TEMPORAL validation set (Section 5.3).

3 THE WHATIFVIS BENCHMARK

We introduce the What If Visual Counterfactual benchmark (WHATIFVIS), adapting existing counterfactual datasets into paired instruction-following tasks and extending their coverage with newly

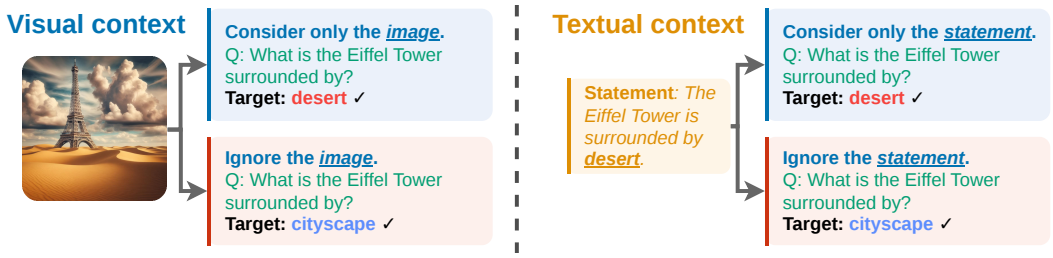


Figure 3: **Evaluation pipeline for controllable multimodal context sensitivity.** We evaluate image (left) and text (right) contexts separately, instructing the model to use the context ("Consider only the image/statement") or parametric knowledge ("Ignore the image/statement").

generated images. We use images from VLind-Bench (Lee et al., 2025), Visual CounterFact (Golovanevsky et al., 2025), VLMs-are-Biased (Vo et al., 2026), ROME (Zhou et al., 2023a), and ViLP (Luo et al., 2025), and generate additional WEIGHT and SPATIAL-TEMPORAL examples with Qwen-Image (Wu et al., 2025). Sources, splits, and annotation procedures are described in Appendix C. With 3,125 counterfactual examples, WHATIFVIS comprises five conflict types grouped into two families (Figure 2): Perception tasks (SPATIAL-TEMPORAL, COLOR) test directly observable attributes; Perception + Reasoning tasks (COUNT, SIZE, WEIGHT) require counting or interpreting relative size and weight. Using the ground-truth annotations from the source benchmarks, we formulate natural-language questions with distinct answers depending on whether the model uses the image or its parametric knowledge. We also pair each image with a text statement describing the same counterfactual, enabling the modality comparison in Section 6.

Paired evaluation. Following Minder et al. (2025), we pair prompts containing the same image and question but opposite instructions: use the visual evidence or rely on parametric knowledge (Figure 3). Holding the image and question fixed lets us compare the model's internal representations under the two instructions. We measure paired accuracy as the percentage of examples for which the model ranks the image-consistent answer higher when instructed to use the image and the prior-consistent answer higher when instructed to ignore it. Unless otherwise specified, we use probability version (PairProb)(Equation (6)).

4 IMAGE RECONSTRUCTION FROM MLLMS

Prior work has highlighted limitations in *fine-grained* visual perception, including difficulty resolving visual details and small objects (Tong et al., 2024b; Wu & Xie, 2024). This leaves the question open at the other end of the spectrum: what about *coarse-grained* content, such as object count and color?

MLLM evaluation suites (Liu et al., 2024; Yue et al., 2024; Li et al., 2026; 2024c) measure task *performance*: a low score is consistent with either missing (**H1**) or unused (**H2**) information, and prior knowledge can supply correct answers without the image. Testing **H1** therefore requires measuring *what visual content is present in the backbone's representation*, independently of the model's answers. Representational probing can provide such evidence: Pang et al. (2026) find linearly decodable quantity information in visual backbones even when VLMs fail to enumerate the corresponding synthetic Go-board configurations. We evaluate the representations *generatively*, attempting to reconstruct images from the LLMs' backbone representations rather than simply scoring their answers (Figure 4).

Probing via Reconstruction. We reconstruct images *solely from final-layer image tokens* of frozen MLLMs. We adapt the framework of MetaQuery (Pan et al., 2025) *without learnable query tokens*, pairing a frozen pretrained MLLM with a trainable six-layer bidirectional transformer connector and the SANA image decoder (Xie et al., 2025) (details in Appendix D; pipeline in Figure 10). We jointly train the connector and fine-tune SANA on ImageNet-1K (Deng et al., 2009) images only; counterfactual benchmark images are excluded from this training. During training and reconstruction, image-specific information reaches the connector and decoder only through these representations, without image-specific textual prompts, counterfactual hints, attribute labels, or separate reference-image latents. The decoder's pretrained prior can supply generic rendering knowledge.



Figure 4: **The counterfactual attribute survives to the final layer and is visible in the reconstruction.** Input images (top, GT) and reconstructions (bottom) are shown. Species identity, pose, structure, and dominant colors are largely preserved, including the counterfactual attribute (e.g. the anomalous limb count), while high-frequency texture softens.

Table 1: **Human evaluation of reconstructions.** On animal-subset of VLM-are-Biased, most preserve the counterfactual count. CF = counterfactual. Real = real-world image attribute.

Backbone	CF(%)	Real(%)
Gemma-4-E2B-IT	**97.8**	2.2
Qwen2.5-VL-3B	**93.4**	6.6
Qwen3.5-4B	**97.8**	2.2
Average	**96.3**	3.7

Table 2: **Reconstructions favor the counterfactual reference (Animal-subset)** across three models and metrics. Ref. = reference; cf = counterfactual; real = paired original real-world image.

Backbone	Ref.	mIoU↑	PSNR↑	SSIM↑	LPIPS↓
Gemma-4-E2B-IT	real	0.83	15.4	0.43	0.33
	cf	**0.94**	**17.9**	**0.52**	**0.23**
Qwen2.5-VL-3B	real	0.78	15.2	0.44	0.35
	cf	**0.86**	**16.2**	**0.48**	**0.28**
Qwen3.5-4B	real	0.79	15.4	0.43	0.33
	cf	**0.88**	**16.6**	**0.47**	**0.27**

We evaluate reconstructions on the *animal-subset* of VLMs-are-Biased (Vo et al., 2026), comparing them with both the counterfactual input and its paired original natural real-world image. [4] Three MLLMs (Qwen2.5-VL-3B (Bai et al., 2025), Qwen3.5-4B (Qwen Team, 2026), and Gemma-4-E2B-IT (Gemma Team, 2026)) achieve *near-zero* performance when answering questions about this subset. [5] The connector and the decoder, trained on ImageNet-1K (Deng et al., 2009), render animals faithfully enough to expose whether the counterfactual content was retained.

Results. We do human evaluation by asking whether the *counterfactual attribute* survives in the reconstructed image. Two annotators identified the animal and reported its color and leg count in each reconstruction; their responses were scored against the ground truth (Cohen's κ=0.84; protocol in Appendix D). Reconstructions from the final-layer image tokens recover the anomalous count correctly in 96.3% of samples on average across three models, while the remaining 3.7% revert to real-world counts (Table 1). The decoder may rely on its pretrained prior for generic rendering, but without textual cues, the counterfactual signal can only originate from the input image.

Across all three models, four complementary metrics favor each reconstruction's *own* counterfactual reference over the paired original image (Table 2). Mean intersection over union (mIoU) of SAM3 masks (Carion et al., 2026) is 0.86–0.94 against the counterfactual references, indicating substantial preservation of object shape and layout. PSNR and SSIM (Wang et al., 2004) measure pixel fidelity and local structural similarity; LPIPS (Zhang et al., 2018) measures perceptual distance in learned features (details in Appendix D). Figure 4 shows some examples from COUNT, which preserved identity, structure, and counterfactual attributes, with poorer reproduction of high-frequency texture. See more results for WHATIFVIS in Table 4 and Appendix E.

[4] The animal-category subset is also part of WHATIFVIS, including 91 unique images in total. Since VLM-are-biased provides the original real-world corresponding images, we can use the counterfactual-real pairs to evaluate the reconstruction quality.

[5] We use the same prompts in the VLM-are-Biased benchmark.

Linear probing. Successful reconstruction establishes recoverability of counterfactual evidence, without establishing that the model's computation uses it. We perform linear probing on the COUNT subset of WHATIFVIS (Section 3) to test whether object count is decodable from the MLLM's text-token representations. We train a linear classifier at each layer to predict the count from the final text-prompt hidden state, following the image, a fixed counting question, and the response prefix "Count:". Figure 5 reveals a trend across models: accuracy rises sharply in early layers and remains above the shuffled-image and no-image controls. We restrict labels to $\{2, 3, 4, 5\}$ and report mean accuracy across these four classes, with a chance baseline of 25%. Details are in Appendix D.

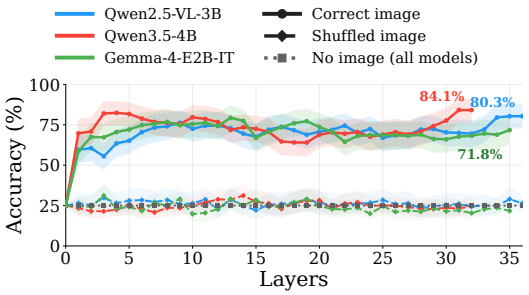


Figure 5: **Layer-wise count probing.** Five-fold macro accuracy on 179 images from COUNT. Solid: correct image; dashed: shuffled; dotted: no image. Bands: 95% bootstrap intervals. Labels: final-layer correct-image accuracy.

Takeaway: At the coarse-grained level we probe, counterfactual evidence remains recoverable from final-layer image tokens. Because these causally preceding representations cannot access the later textual instruction, the result argues against complete vision information loss and locates the remaining failure in how the following text-token computation uses the available vision evidence.

5 FINDING THE KNOB BEHIND VISUAL CONTEXT SENSITIVITY

Controllable Context Sensitivity (CCS) (Minder et al., 2025) was formulated for text-only conflicts; our controllable visual context sensitivity (CVCS) setting presents the evidence visually. Reconstruction shows that the tested coarse attributes remain recoverable (Section 4), arguing against **H1** in that setting. We now turn to **H2** and test whether the model can be steered to use or disregard visual evidence on demand.

We evaluate the released instruction-tuned models without benchmark-specific training ("vanilla" models) and adapt them via supervised fine-tuning (SFT) to test whether visual controllability is learnable and transfers across tasks. We then find intervention layers in the SFT models and test steering in the vanilla models (Sections 5.2 and 5.3).

5.1 EVALUATING MODEL BEHAVIOR

To study the image sensitivity of a model, we first evaluate how well it follows instructions to use either image context or prior knowledge, before and after SFT. We fine-tune six instruction-tuned multimodal LLMs from three model families and two size tiers each: Qwen2.5-VL-3B/7B (Bai et al., 2025), Qwen3.5-4B/9B (Qwen Team, 2026), and Gemma-4-E2B/E4B-IT (Gemma Team, 2026).

Evaluation. Each model is evaluated with explicit intent instructions ("Consider only the image to answer the question" / "Ignore the image to answer the question"); see Figure 3. All SFT, patching, and steering use SPATIAL-TEMPORAL data *only*: SFT and direction learning use its training split, and layer search and steering-value tuning use its validation split; the remaining four tasks (COLOR, SIZE, COUNT, WEIGHT) serve as out-of-distribution (OOD) test sets. We report two conjunctive paired-accuracy metrics (PairEM and PairProb) each requiring correct behavior under both intent instructions on the same example. Formal definitions and scoring details are provided in Section C, specifically Equations (4) and (6). Unless otherwise specified, accuracies are reported as percentages and differences as percentage points (pp).

Vanilla models struggle to follow both instructions. The top row of Figure 6 shows that visual context controllability is consistently low. The collapse is most pronounced on COUNT and WEIGHT, where baseline PairProb is near zero. On COLOR, however, the strongest models reach around 50%. The remaining two tasks fall in between. The bottom row shows an imbalance between the two instructions. Vanilla models achieve higher prior-accuracy than context-accuracy on SPATIAL-TEMPORAL, COUNT, and WEIGHT. In contrast, for SIZE and COLOR they achieve higher

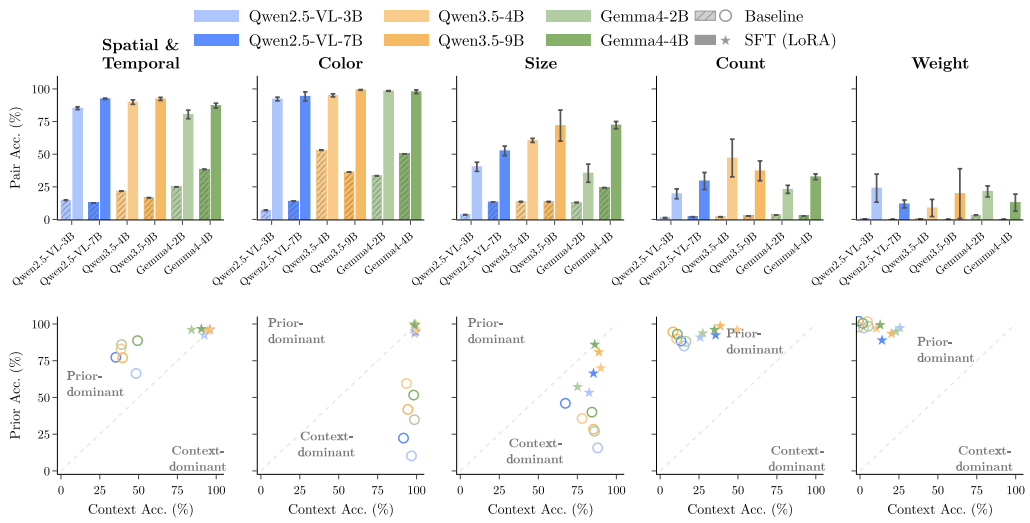


Figure 6: **SFT lifts visual controllability across every model and task in the image setting. Top row:** PairProb per task. **Bottom row:** context-accuracy (x) against prior-accuracy (y); points above the diagonal have higher prior-accuracy. Error bars are standard deviation.

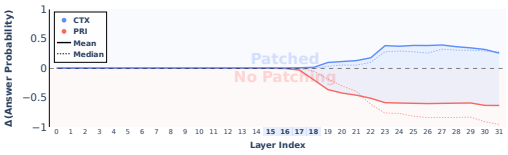
context-accuracy. PairProb exposes a brittleness that marginal accuracy hides: a model may follow the image on some examples and the prior on others, yet fail to satisfy both instructions on the *same* example. For the Eiffel Tower example in Figure 3, a pair succeeds only if the model ranks "desert" above "cityscape" when told to use the image and reverses that ranking when told to ignore it.

SFT lifts controllability and transfers across tasks. Figure 6, Figure 8, and Table 5 show that SFT lifts PairProb across all six models, raising macro PairProb from 14.1% to 58.75%. The gains are largest on the perception attributes, COLOR and SPATIAL-TEMPORAL. SFT on a *single* subset, SPATIAL-TEMPORAL, produces gains on the held-out sets, showing that the improvement transfers beyond the training task. Two tasks remain difficult after transfer, COUNT and WEIGHT, which require more reasoning after processing the visual information. For instance, when shown a pencil and a basketball on a balance scale, the model must interpret the scale's tilt, infer that the object on the lower side is heavier, and use this counterfactual visual cue to override a strong prior (Figure 2). SFT improves performance on both tasks, but their PairProb remain below those of the other tasks. The bottom row of Figure 6 shows that SFT generally brings context- and prior-accuracy closer together, although prior-accuracy sometimes decreases. See exact-match accuracy results in Appendix E.

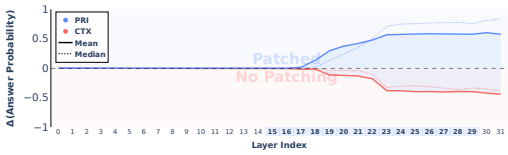
5.2 FIND INTERVENTION LAYERS FOR CONTEXT CONTROL

We analyze SFT models because they follow the two instructions more reliably than vanilla models (Tables 5 and 7). We apply activation patching to identify layers where interventions change whether an answer follows the image or prior knowledge. Activation patching adapts the causal-tracing and circuit-analysis toolkit developed for language models (Wang et al., 2023) to our multimodal setting. Following Minder et al. (2025), we replace multi-head attention outputs at the last prompt-token position in a target pass with those from a source pass carrying the opposite instruction. Prior studies of factual recall, activation steering, and grounding motivate examining this position (Yu et al., 2023; Stoehr et al., 2024; Monea et al., 2024). More details in Appendix D.

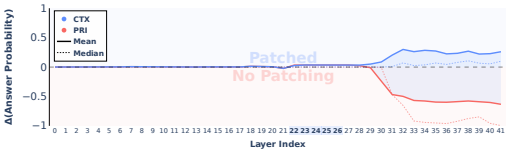
Representative patching curves for two models from different families are shown in Figure 7; the full set (all six models, both directions) is in Appendix E.3. In $\mathrm{ctx} \to \mathrm{pri}$ patching, the source instruction says to use the image and the target instruction says to ignore it; $\mathrm{pri} \to \mathrm{ctx}$ reverses these roles. In both directions, the patched readout favors the source answer over the target answer in later layers, while the unpatched readout favors the target answer. Across the six models, effective intervention windows occur in the middle and later layers, with their positions and widths varying by model and direction. This identifies layer windows where patching can change which information source the model follows; it does not establish a unique or earliest site for that computation.



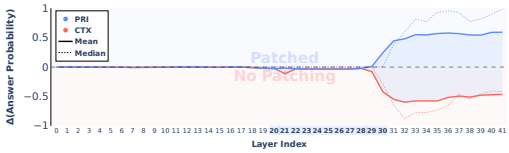
(a) Qwen3.5-9B, ctx → pri, L15–18



(b) Qwen3.5-9B, pri → ctx, L15–29



(c) Gemma-4-E4B, ctx → pri, L22–26



(d) Gemma-4-E4B, pri → ctx, L20–30

Figure 7: **Activation patching identifies layers that influence the use of visual context.** Each curve shows $\Delta(AnswerProbability) = Probability(\text{source}) - Probability(\text{target})$ for the first answer token in the layer-wise readout. Blue: patched; red: unpatched. CTX and PRI denote context- and prior-consistent answers. Solid/dotted curves summarize across-example means/medians. Positive values favor the source answer; negative values favor the target answer. The blue–red separation measures patching's effect; positive blue values above a negative red baseline indicate reversed first-token preference. Highlighted indices mark the simultaneously patched layers.

5.3 A Controllable Subspace for Visual Context Sensitivity

Having identified candidate intervention layers (Section 5.2), we learn a direction for controlling whether the model follows visual evidence or prior knowledge. Following Minder et al. (2025), we investigate whether a one-dimensional subspace can control reliance on the provided context versus parametric knowledge. Our procedure has three stages.

First, for each backbone and modality, we freeze the selected SFT model and learn only a rank-one orthogonal projection, parameterized by a unit vector $\boldsymbol{u}$ with $\mathcal{F} = \text{span}\{\boldsymbol{u}\}$, using the objective in Section 2. Training uses only the SPATIAL-TEMPORAL training split. We retain pairs whose first answer token is correctly predicted in both modes. Second, we transfer the learned projection to the vanilla model, and independently sweep the $c(\text{ctx})$ and $c(\text{pri})$ scalar values using SPATIAL-TEMPORAL validation set. Third, we fix the selected values and evaluate them on the five held-out test sets. During inference, $c(w)$ sets the projected coordinate through Equation (3): $c(\text{ctx})$ steers toward the provided context, whereas $c(\text{pri})$ steers toward prior knowledge. The intervention is applied once at the final response-prefix position, and the prompt contains no explicit intent instruction.

Steering without intent instructions improves controllability. Subspace steering lifts mean PairProb from 14.1% to 34.2%, averaged across the six models (Figure 8, Table 5). Steering thus *outperforms the instructed vanilla baseline*, although it remains below SFT (58.75%). A single rank-one direction therefore recovers part of the SFT improvement without intent instructions, paralleling the direction previously shown to control reliance on textual evidence in text-only LLMs (Minder et al., 2025). What's more, we adapted PvP steering method (Golovanevsky et al., 2025) on WHATIFVIS, and also evaluate on the other four general multimodal understanding benchmarks including MME Fu et al. (2026), MMBench Liu et al. (2024), SEED Li et al. (2023a), MMMU Yue et al. (2024). [6] See more details in Table 5 and Table 6 in Appendix E.

Takeaway: Vanilla MLLMs rarely answer correctly under both use-image and ignore-image instructions for the same image and question, and their context sensitivity varies sharply across tasks, such as COLOR versus WEIGHT. As in text-only LLMs, a rank-one direction shifts reliance on visual evidence versus prior knowledge. Learned at layers identified through activation patching in SFT models, this direction transfers to their vanilla counterparts and raises mean PairProb without explicit intent instructions.

[6] We use LMMs-Eval Zhang et al. (2025) to evaluate the models on multimodal understanding tasks.

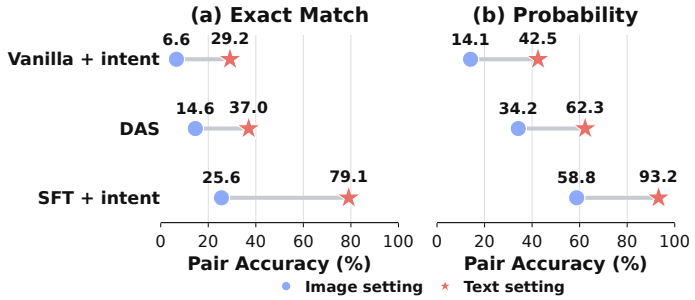


Figure 8: Pair accuracy (%) averaged over six models and the test splits of the five tasks, in the image and text settings. DAS steers the vanilla model without intent instructions.

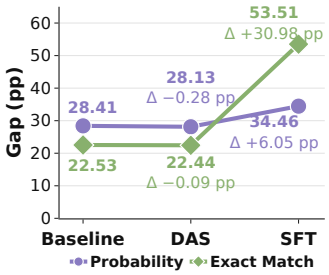


Figure 9: DAS leaves the gap nearly unchanged, whereas SFT widens it.

6 Comparison with Textual Context

We repeat the baseline, SFT, and subspace evaluations using WhatIfVis's matched text statements, such as "The horse has 5 legs" (Figure 2). Although the two settings convey the same counterfactual fact, they do not make it equally accessible. Text states the relevant fact directly, whereas an image requires the model to extract and integrate visual evidence. We therefore treat text as a reference for controllability when the counterfactual is explicitly lexicalized, rather than as a pure modality ablation. The cross-modal comparison is restricted to vanilla, SFT, and DAS. See more results in Appendix F.

The per-task text–image gap tracks visual difficulty. According to Figure 8, for matched counterfactual content, text exceeds image by 28.41 pp in PairProb (42.51% versus 14.10%) and by 22.53 pp in PairEM (29.15% versus 6.62%). According to Figure 27, for PairProb, adaptation nearly closes the gap on Color. In contrast, Count and Weight remain strongly modality-dependent. Overall, visually demanding tasks remain bottlenecked by extracting and using evidence from pixels, even when the same information is handled successfully in text. See PairEM results in Figure 26.

The modality gap persists under subspace steering and widens under SFT. From Figure 9, DAS steering leaves the text–image gap essentially unchanged: 28.13 pp in PairProb and 22.44 pp in PairEM, only 0.28 and 0.09 pp smaller than the corresponding baseline gaps. SFT instead increases the gaps to 34.46 and 53.51 pp, respectively—6.05 and 30.98 pp wider than at baseline. Thus, neither intervention meaningfully closes the cross-modal disparity: DAS steering shifts the two modalities almost in parallel, whereas SFT benefits the text setting disproportionately.

7 Conclusions

MLLMs need to follow visual evidence that contradicts familiar knowledge and disregard it when the task calls for prior knowledge. In this work, we ask: when an MLLM fails on such a conflict, is the fault that it does not see the evidence, or that it sees it and does not use it? Our study begins by separating the two possibilities, **H1** and **H2**. For the coarse attributes we study, counterfactual content remains recoverable from frozen models' final-layer image-token representations, directing attention to how the model acts on that evidence.

We then analyze this downstream behavior using the CVCS framework and find it uneven. Models follow visual context more reliably for directly observable attributes, while Count and Weight remain difficult even after SFT. Building on this, activation patching identifies layer windows where interventions change whether answers follow the image or prior knowledge. A single learned vector transfers from each SFT model to the original model and improves controllability without explicit intent instructions, recovering part of the SFT improvement. Moreover, our paired image and text design exposes an asymmetry: the same task-relevant facts are easier to follow when stated in words.

Separating what MLLMs retain from how they use it points to the language backbone as a target for improvement, and our adaptation and steering results show that such improvement is possible. WhatIfVis provides a testbed for measuring this gap and progress toward more robust MLLMs.

AI USE STATEMENT

We used Qwen-Image-Edit-2511 to generate part of WHATIFVIS, with human verification as described in Appendix C. We also used AI assistants to help draft and edit the manuscript, check relevant literature and mathematical exposition, discuss interpretations of results and methodological limitations, and prepare LaTeX figures. We didn't use generative AI assisntant to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology. And transcribe recordings of research material are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Human evaluation consisted of checking benchmark examples and judging reconstructed images (Appendices C and D). The two annotators are also co-authors of the paper. We identify no specific ethical concerns arising from this study.

REPRODUCIBILITY STATEMENT

Appendix C describes the benchmark sources, curation, splits, and scoring definitions, and Appendix G gives image and text prompt examples. Appendix D describes reconstruction, linear probing, SFT, finding the important layers, subspace interventions. Sections 5.2 and 5.3 describe activation patching and the learned intervention; Appendices E.3, E.4, and F provide additional results. We will publicly release the source code and the WHATIFVIS benchmark.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=EbMuimAbPbs.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pat-*

tern Recognition, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=qQb1JLrwol.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, November 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, December 2023. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458/.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. URL https://arxiv.org/abs/1312.6114.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-bench: Measuring language priors in large vision-language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, April 2025. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.231. URL https://aclanthology.org/2025.findings-naacl.231/.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? A vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024a. doi: 10.1162/tacl_a_00698. URL https://aclanthology.org/2024.tacl-1.68/.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. RAVENEA: A benchmark for multimodal retrieval-augmented visual culture understanding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4zAbkxQ23i.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. Understanding retrieval robustness for retrieval-augmented image captioning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9285–9299, August 2024b. doi: 10.18653/v1/2024.acl-long.503. URL https://aclanthology.org/2024.acl-long.503/.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19077–19095, November 2024c. doi: 10.18653/v1/2024.emnlp-main.1063. URL https://aclanthology.org/2024.emnlp-main.1063/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, Viktar Atliha, Tony Ng, Xiao Han, Chuyan Zhu, Chenyang Zhang, Ding Liu, Juan-Manuel Perez-Rua, Sen He, Jürgen Schmidhuber, Wenhu Chen, Ping Luo, Wei Liu, Tao Xiang, Jonas Schult, and Yuren Cong. TUNA: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025. URL https://arxiv.org/abs/2512.02014.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. *arXiv preprint arXiv:2604.24763*, 2026. URL https://arxiv.org/abs/2604.24763.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, November 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. Controllable context sensitivity and the knob behind it. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Igm9bbkzHC.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the Matrix? locating and detecting language model grounding with Fakepedia. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6828–6844, August 2024. doi: 10.18653/v1/2024.acl-long.369. URL https://aclanthology.org/2024.acl-long.369/.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.642. URL https://aclanthology.org/2026.acl-long.642/.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025. URL https://arxiv.org/abs/2504.06256.

Xingzhou Pang, Yifan Hou, Junling Wang, and Mrinmaya Sachan. Unveiling the visual counting bottleneck in vision-language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.30170.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1285. URL https://aclanthology.org/2025.findings-emnlp.1285/.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, November 2024. doi: 10.18653/v1/2024.findings-emnlp.479. URL https://aclanthology.org/2024.findings-emnlp.479/.

Hao Tang, Chen-Wei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. UniLiP: Adapting CLIP for unified multimodal understanding, generation and editing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6tx4BGjwJP.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024a. doi: 10.52202/079017-2771. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. MetaMorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17001–17012, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Tong_MetaMorph_Multimodal_Understanding_and_Generation_via_Instruction_Tuning_ICCV_2025_paper.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DG4S2OlGQA.

Hermann von Helmholtz. *Handbuch der physiologischen Optik*. Leopold Voss, 1867. URL https://archive.org/details/handbuchderphysi00helm.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. URL https://ece.uwaterloo.ca/~z70wang/research/ssim/.

Hongyang Wei, Baixin Xu, Hongbo Liu, Size Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, Chuanxin Tang, Zidong Wang, Yichen Wei, Liang Hu, Boyi Jiang, Wei Li, Ying He, Yang Liu, Xuchen Song, Yangguang Li, and Yahui Zhou. Skywork UniPic 2.0: Building Kontext model with online RL for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025. URL https://arxiv.org/abs/2509.04548.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun

Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9924–9959, December 2023. doi: 10.18653/v1/2023.emnlp-main.615. URL https://aclanthology.org/2023.emnlp-main.615/.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-eval: Reality check on the evaluation of large multimodal models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.51. URL https://aclanthology.org/2025.findings-naacl.51/.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. URL https://arxiv.org/abs/2510.11690.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. ROME: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10185–10197, December 2023a. doi: 10.18653/v1/2023.findings-emnlp.683. URL https://aclanthology.org/2023.findings-emnlp.683/.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pp. 14544–14556, December 2023b. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

APPENDIX

This appendix provides supplementary methodology, analyses, and results supporting the main paper:

- §A situates our reconstruction diagnostic, benchmark, and context-control intervention relative to prior work.
- §B discusses the scope and limitations of our study and directions for future work.
- §C describes the composition of WHATIFVIS, its task formulation, scoring rules, data splits, and human curation.
- §D details reconstruction, linear probing, supervised fine-tuning, activation-patching localization, and subspace interventions.
- §E reports additional reconstruction, fine-tuning, activation-patching, and steering results in the image setting.
- §G provides the image- and text-context prompt templates used in our experiments.
- §F presents the full textual-context results and additional image–text comparisons.

A RELATED WORK

Visual information in MLLMs. Multimodal large language models (MLLMs) are typically pretrained via next-token prediction over heterogeneous modalities, including images, text, and videos. Early MLLMs, BLIP (Li et al., 2022), Flamingo (Alayrac et al., 2022) introduce cross-attention mechanisms to enable interactions between visual and linguistic representations. Modern MLLMs predominantly follow the paradigm established by LLaVA (Liu et al., 2023), where visual tokens extracted by pretrained vision encoders, e.g., SigLIP (Zhai et al., 2023), are projected into the token space of large language models (LLMs). Prior studies have revealed that current MLLMs are fundamentally bottlenecked by the representational capacity of their vision encoders (Tong et al., 2024b; Wu & Xie, 2024). Consequently, combining multiple vision backbones or task-specialized visual models has emerged as an effective strategy, as different encoders provide complementary semantic and structural cues that improve downstream multimodal reasoning (Tong et al., 2024a; Liu et al., 2025). Recently, inspired by the platonic representation hypothesis (Huh et al., 2024; Li et al., 2024a), some work attempts to jointly enhance visual understanding and generation capabilities within a unified vision encoder or MLLMs (Liu et al., 2026; Zheng et al., 2025; Tong et al., 2025). Unlike prior work focusing primarily on fine-grained representation analysis, our work studies whether VLMs can reliably capture and reason over simple coarse-grained visual semantics through specific question-based evaluation. Moreover, rather than analyzing representations solely at the vision encoder level (Fu et al., 2025), we evaluate the quality of visual information by pixel-level reconstruction after its propagation through both the projector and the LLM backbone.

Reconstruction and generation. Prior reconstruction-based approaches such as UniLiP (Tang et al., 2026) and UniPic 2.0 (Wei et al., 2025) aim to improve generation and editing: UniLiP adapts CLIP for reconstruction and uses learnable queries alongside multimodal hidden states. Both UniLiP and UniPic 2.0 additionally supplies reference-image latents from a variational autoencoder (VAE) (Kingma & Welling, 2013) as the decoder input to improve generation quality. In our work, we reconstruct images *solely from final-layer image tokens* of frozen MLLMs. We adapt the framework of MetaQuery (Pan et al., 2025) *without learnable query tokens*.

Context sensitivity. Multimodal large language models (MLLMs) fundamentally rely on balancing two competing information sources: the immediate visual context provided by the input and the parametric prior knowledge acquired during large-scale pretraining. Although extensive pretraining equips LLMs with strong capabilities in factual recall and sequence memorization (Zhou et al., 2023b; Carlini et al., 2023; Karamolegkou et al., 2023), it can also induce substantial knowledge conflicts when in-context evidence contradicts the parametric prior, a failure mode first characterized for text-only question answering (Longpre et al., 2021) and later found to drive cross-modal conflicts and hallucinations when pretrained linguistic priors override visual evidence (Tong et al., 2024b; Li et al., 2024b; Vo et al., 2026). VLind-Bench (Lee et al., 2025) tests basic perceptual and knowledge capabilities before measuring language priors. To mitigate this issue, recent work has begun to control this competition directly. Pixels-versus-Priors (PvP) (Golovanevsky et al., 2025) traces layer-wise predictions on Visual CounterFact and constructs steering vectors by averaging activation differences between image-focused and prior-focused prompts for color and size counterfactuals. Ortu et al. (2026) instead

identify factual and counterfactual attention heads and rescale their attention to image and text tokens. Beyond visual–prior conflicts, SteerVLM (Sivakumar et al., 2025) learns a nonlinear, token-dependent module for general multimodal behavior control and hallucination mitigation. Our setting differs in both intervention and evaluation. Building on the causal-subspace formulation of Minder et al. (2025), we use activation patching and DAS (Geiger et al., 2024) to learn a single rank-1 direction, which is transferred from an adapted model to its original adapter-free counterpart. We hold the image and question fixed, vary only whether the model should use or disregard the image, and require both intents to succeed on the same example. We further evaluate instruction-free control across five conflict types, four held out from intervention learning, and compare matched visual and textual contexts.

B Limitations and Future Work

We focus on why MLLMs fail on coarse visual attributes, such as object identity, shape, count, size, color, and weight. On the architecture front, while we show that the visual evidence survives by reconstructing images from three architecturally distinct families, a standard attention, a linear-attention hybrid and a per-layer-embedding design, our scope is not exhaustive; for instance, other designs such as mixture-of-experts routing remain to be probed. Regarding data, our benchmark forces the model to answer the question directly and excludes intermediate reasoning traces; attributes that require a further inference step, such as counting and weight, therefore remain a bottleneck for reliable control and a critical area for future exploration. As for cross-modal comparison, our text condition states the counterfactual outright and serves as a reference for instruction following without visual extraction; the performance gap between modalities on the same underlying fact remains a critical area for future exploration. In the long run, we believe the capability to decide when to trust the image and to *integrate* it with prior knowledge will drive the convergence of seeing and knowing, fostering multimodal reasoning grounded in what a model actually sees.

C Benchmark Composition and Curation

WhatIfVis contains 3,125 examples across five conflict types (Table 3 and Figure 2). Images are repurposed and extended from VLind-Bench (Lee et al., 2025), Pixels-versus-Priors (Golovanevsky et al., 2025), the counting subset of VLMs-are-Biased (Vo et al., 2026) , the size subset of ROME (Zhou et al., 2023a), ViLP (Luo et al., 2025), and generated images from Qwen-Image[7] (Wu et al., 2025). Each example carries a `question`, `image`, `context_answer`, `prior_answer`, `response_template`, and a declarative `statement` for the textual condition.

Consider an MLLM p over an alphabet Σ. Let $\mathcal{Q}$ be a set of questions and $\mathcal{C}$ a set of evidence, where each piece of evidence $c \in \mathcal{C}$ is either a counterfactual image $\mathbf{I}$ (**visual context**) or a one-sentence statement s (**textual context**) asserting the same content. For a question q (e.g., *What is the Eiffel Tower surrounded by?*) and evidence c (e.g., an image showing *the Eiffel Tower surrounded by desert*), let $a(q, \varepsilon)$ be the context-independent answer (*cityscape*) and $a(q, c)$ the context-dependent one (*desert*). An *intent* $w \in \{\mathrm{ctx}, \mathrm{pri}\}$ selects whether to follow the evidence (ctx, target $a(q, c)$) or the prior (pri, target $a(q, \varepsilon)$), realized by a formatting function F that maps (q, c, w) to a prompt. On a training split $\mathcal{S}_{\mathrm{trn}}$, models see both $F(q, c, \mathrm{pri}) \cdot a(q, \varepsilon)$ and $F(q, c, \mathrm{ctx}) \cdot a(q, c)$ ($\cdot$ = concatenation); a disjoint $\mathcal{S}_{\mathrm{tst}}$ is held out for testing. Examples of both modalities are shown in Figure 3.

Exact-match scorer. Let $\mathrm{Canon}(\cdot)$ denote the canonicalization used by our evaluator: Case-folding, replacement of non-letter/non-digit characters by spaces, and whitespace collapsing. For intent $w \in \{\mathrm{ctx}, \mathrm{pri}\}$, define the generated answer fragment as

$$\widehat{\boldsymbol{a}}_w \stackrel{\text{def}}{=} \mathrm{Decode}_{\pi,\xi}[p(\cdot \mid F(q, c, w))],$$

where π is the model-family generation policy and ξ is the fixed evaluation seed. An example is correct only when the complete normalized generation matches the corresponding gold answer under both intents:

$$\mathrm{PairEM}(p, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{(q,c)\in\mathcal{S}} \mathbb{1}\{\mathrm{Canon}(\widehat{\boldsymbol{a}}_{\mathrm{ctx}}) = \mathrm{Canon}(a(q, c))\} \cdot \mathbb{1}\{\mathrm{Canon}(\widehat{\boldsymbol{a}}_{\mathrm{pri}}) = \mathrm{Canon}(a(q, \varepsilon))\}. \quad (4)$$

[7] https://huggingface.co/Qwen/Qwen-Image-Edit-2511

Table 3: **WHATIFVIS composition.** The benchmark spans five conflict types across two families, incorporating data from VLindBench (Lee et al., 2025), ViLP (Luo et al., 2025), Pixel-vs-Priors (Golovanevsky et al., 2025), VLMsAreBiased (Vo et al., 2026), ROME (Zhou et al., 2023a), and a Generated set (Wu et al., 2025). Perception tasks require reading directly observable attributes; reasoning tasks additionally require inference beyond direct visual features. Counts report context–prior pairs in each split; only Spatial–Temporal contributes training examples.

Family	Task	Description	Sources	Train	Val	Test	Total
Perception	SPATIAL-TEMPORAL	Identify what surrounds, feeds, or characterizes an entity.	VLindBench, ViLP, Generated	1,136	102	300	1,538
	COLOR	Report the color of an object shown in a non-canonical hue.	Pixel-vs-Priors, ViLP, VLindBench	0	67	215	282
Reasoning	COUNT	Count parts or instances when the depicted number is anomalous.	VLMsAreBiased, ViLP	0	68	146	214
	SIZE	Judge relative size when the visual cue inverts the usual ordering.	ROME, Pixel-vs-Priors, VLindBench	0	473	300	773
	WEIGHT	Judge relative weight from a balance-scale image with counterfactuals.	VLindBench, Generated	0	68	250	318
Total				**1,136**	**778**	**1,211**	**3,125**

Probability scorer. For a prompt $\boldsymbol{x}_w = F(q, c, w)$ and a candidate answer whose tokenization is $\boldsymbol{y} = (y_1, \ldots, y_{T_{\boldsymbol{y}}})$, we compute the length-normalized score

$$s_p(\boldsymbol{y} \mid \boldsymbol{x}_w) \stackrel{\text{def}}{=} \frac{1}{T_{\boldsymbol{y}}} \sum_{t=1}^{T_{\boldsymbol{y}}} \log p(y_t \mid \boldsymbol{x}_w, y_{<t}) \,. \tag{5}$$

Each candidate is thus evaluated along its own gold-token history, and s_p is the log of its per-token geometric-mean probability. Using this score, the probability-based paired accuracy is

$$\begin{aligned} \mathrm{PairProb}(p, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{(q,c)\in\mathcal{S}} & \mathbb{1}\{s_p(a(q,c) \mid \boldsymbol{x}_{\mathrm{ctx}}) > s_p(a(q,\varepsilon) \mid \boldsymbol{x}_{\mathrm{ctx}})\} \\ & \cdot \mathbb{1}\{s_p(a(q,\varepsilon) \mid \boldsymbol{x}_{\mathrm{pri}}) > s_p(a(q,c) \mid \boldsymbol{x}_{\mathrm{pri}})\} \,. \end{aligned} \tag{6}$$

Unlike unnormalized sequence log-likelihood, the mean reduction explicitly accounts for unequal candidate-token lengths. This scorer bypasses free-form decoding but still requires the model to rank the intended complete answer above the alternative under both prompts for the same example.

Gold answers and scoring protocol. Each example inherits its `context_answer` (matching the image) and `prior_answer` (matching world knowledge) directly from the source dataset's annotations. Answers are constrained to a per-category `response_template` (e.g. "The color is {}"), so there will be less impact of the response style.

Splits and source diversity. Examples are divided into train/val/test with source diversity deliberately maintained *across* the train/test boundary to limit leakage. In particular, the SPATIAL-TEMPORAL split, the only split used for SFT, patching, and steering, draws its *test* set primarily from ViLP (Luo et al., 2025) and VLind-Bench (Lee et al., 2025), whereas its training set is dominated by VLind-Bench (Lee et al., 2025) and generated images. The original image and their edited variants are in the same split. In-domain test images come from a different sub-distribution than training to avoid train-test leakage. The four held-out tasks (COLOR, SIZE, COUNT, WEIGHT) are never used for training and constitute the out-of-distribution evaluation. Per-task counts and sources are in Table 3.

Curation pipeline and human verification. The Generated subset (part of WEIGHT; part of SPATIAL-TEMPORAL) supplies categories with sparse existing coverage, produced with the Qwen-Image-Edit-2511 diffusion image-editing model (Wu et al., 2025). Every generated image is human-verified for counterfactual correctness: two annotators independently rate how well each image

follows its target counterfactual prompt on the VLind-Bench (Lee et al., 2025) faithfulness scale (0-3), and we only retain images that *both* annotators assign the top rating ("3"). Images failing this double-top-rating criterion are discarded.

For the other subsets, each image, along with its associated question and answer, was independently reviewed by two annotators. They assessed whether the question could be answered both with and without the image, yielding different answers. We retained only those images for which both annotators agreed that the question was answerable in both settings. The matched text statements are from the original datasets.

D IMPLEMENTATION

Reconstruction. Following MetaQuery (Pan et al., 2025), we use Qwen2.5-VL-3B, Qwen3.5-4B, and Gemma-4-E2B-IT as frozen MLLM backbones (Figure 10). We resize the images to retain 256 final-layer image tokens and condition SANA-0.6B through a six-layer bidirectional transformer connector. SANA is initialized from its publicly released pretrained checkpoint; the connector is initialized from scratch. We jointly train the connector and fine-tune SANA on ImageNet-1K with a global batch size of 512 and an initial learning rate of 4e-4. The learning rate is warmed up for 1,000 steps, then decays to 1e-5 following a cosine schedule, in total 10,000 steps. Counterfactual benchmark images are excluded from reconstruction training. During both training and reconstruction, the connector receives image-token states without image-specific captions or attribute annotations. The MLLM input may include a generic system prompt before the image.

Independence from textual intent. In our experiments, image tokens precede the question and intent instruction in the causal sequence. Consequently, the final-layer image-token representations of MLLMs may attend to the image tokens and preceding system prompt prefix, but cannot attend to the subsequent question or instruction. The representations supplied to the reconstruction probe therefore cannot encode whether the model was later asked to follow or ignore the image. This makes reconstruction a test of whether visual evidence survives in the backbone before the following text-token computation selects how that evidence should be used.

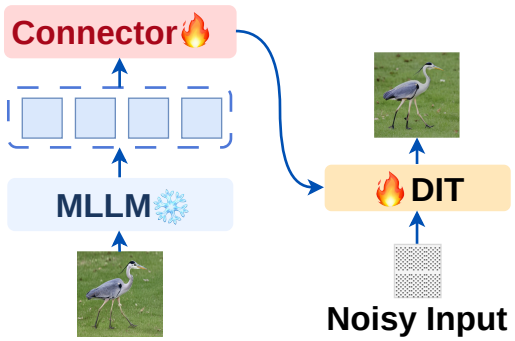


Figure 10: Reconstruction probe adapted from MetaQuery. A connector maps the frozen MLLM's final-layer image-token states to conditioning inputs for SANA, without learnable query tokens.

Reconstruction evaluation. For the animal reconstructions, two annotators identified the animal and reported its color and leg count. If they answer the question wrongly, we assume the image is not correct. Then we count how many correct images in total at last. Two annotators are co-authors in the paper. Responses were compared with the ground-truth attributes; inter-annotator agreement was Cohen's κ=0.84. Table 1 reports recovery of the counterfactual in animal-subset in VLMs-Are-Biased.

We also compare reconstructions with the counterfactual inputs and their paired original natural real-world images using four metrics. Mean intersection over union (mIoU) measures overlap between SAM3 object masks (Carion et al., 2026). Peak signal-to-noise ratio (PSNR) measures fidelity through squared pixel error, and the structural similarity index (SSIM) compares local luminance, contrast, and structure (Wang et al., 2004). Learned perceptual image patch similarity (LPIPS) measures distance between learned image features (Zhang et al., 2018). Higher mIoU, PSNR, and SSIM and lower LPIPS indicate closer agreement with the reference.

Layer-wise linear probing. The COUNT validation and test files contain 214 rows in total. We retain rows whose context label belongs to $\mathcal{Y} = \{2, 3, 4, 5\}$, and remove the duplicated images. This produces 179 unique images, with 15, 27, 33, and 104 images for labels 2, 3, 4, and 5, respectively. To remove object-name and question-template shortcuts, every example uses the identical user text

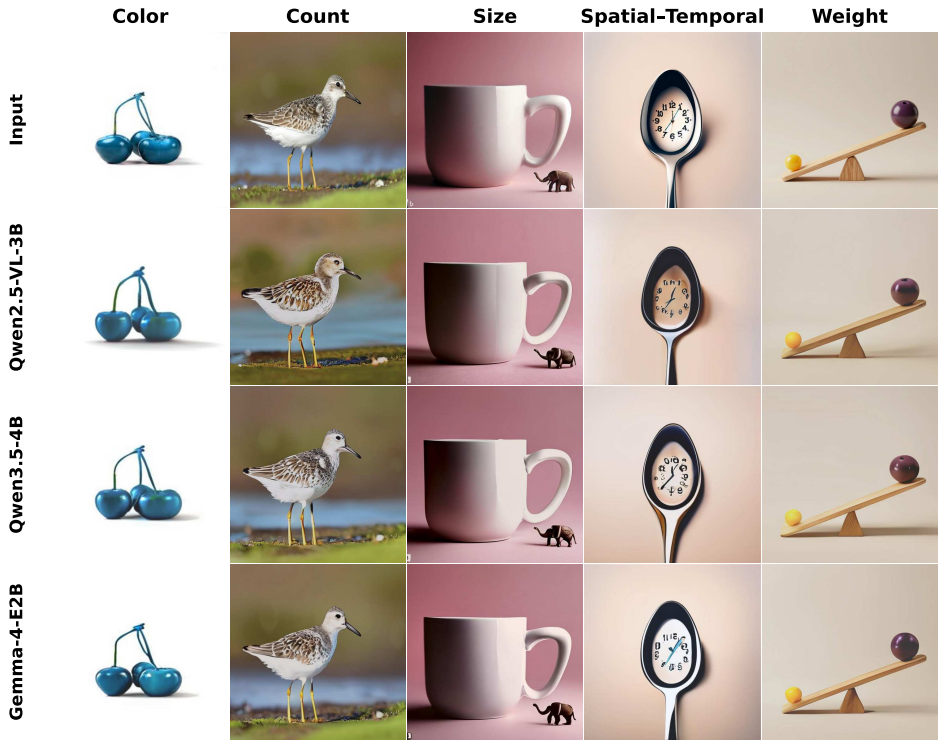


Figure 11: Qualitative reconstruction results across five attributes in WHATIFVIS.

"How many target objects are visible in the image?", followed by the same answer-format instruction. The source statement, intent instruction, and answer tokens are excluded from the prompt.

For each frozen MLLM and each layer, we take the representation at the final text prompt position following the fixed response prefix "Count:". The prompt contains no intent instruction or answer token. We use seed-42 label-stratified five-fold cross-validation, with each unique image held out exactly once. Within every fold, features are standardized using the training partition only. At each hidden-state index we fit a multinomial logistic linear probe with an intercept, balanced class weights, fixed L_2 regularization ($C = 1$), the LBFGS solver, tolerance 10^{-4}, and at most 5,000 iterations.

Let $\widehat{y}_i^{(\ell)}$ be the out-of-fold prediction for example i at layer index ℓ. We report pooled out-of-fold macro accuracy

$$\mathrm{Acc}_{\mathrm{macro}}(\ell) = \frac{1}{|\mathcal{Y}|} \sum_{k \in \mathcal{Y}} \frac{1}{n_k} \sum_{i: y_i = k} \mathbb{1}\left[\widehat{y}_i^{(\ell)} = k\right]. \tag{7}$$

At every layer, 95% percentile intervals use 2,000 label-stratified image bootstrap replicates. We evaluate correctly paired images, no image, and a seed-42 label-agnostic image derangement performed independently within each fold. The probes are refitted independently for each control—not reused from the correct-image condition. At the final hidden state, correct-image macro accuracy is 80.31%, 84.12%, and 71.76% for Qwen2.5-VL-3B, Qwen3.5-4B, and Gemma-4-E2B-IT, respectively; no-image accuracy is 25.00% for all three, and shuffled-image accuracy is 26.35%, 24.78%, and 21.71%. Figure 5 reports all layer results rather than selecting a best layer.

SFT. We use *only* the SPATIAL-TEMPORAL training set (1,136 samples) to adapt six instruction-tuned MLLMs, comprising two size tiers from each of three families: Qwen2.5-VL-3B/7B (Bai et al., 2025), Qwen3.5-4B/9B (Qwen Team, 2026), and Gemma-4-E2B/E4B-IT (Gemma Team, 2026). We fine-tune these models using standard next-token prediction loss on the training data. We apply LoRA (Hu et al., 2022) to the attention projection matrices in all layers, using a global batch size of 32, the AdamW optimizer, and a learning rate of 2×10^{-4}. We use $(r, \alpha) = (16, 32)$; hence the

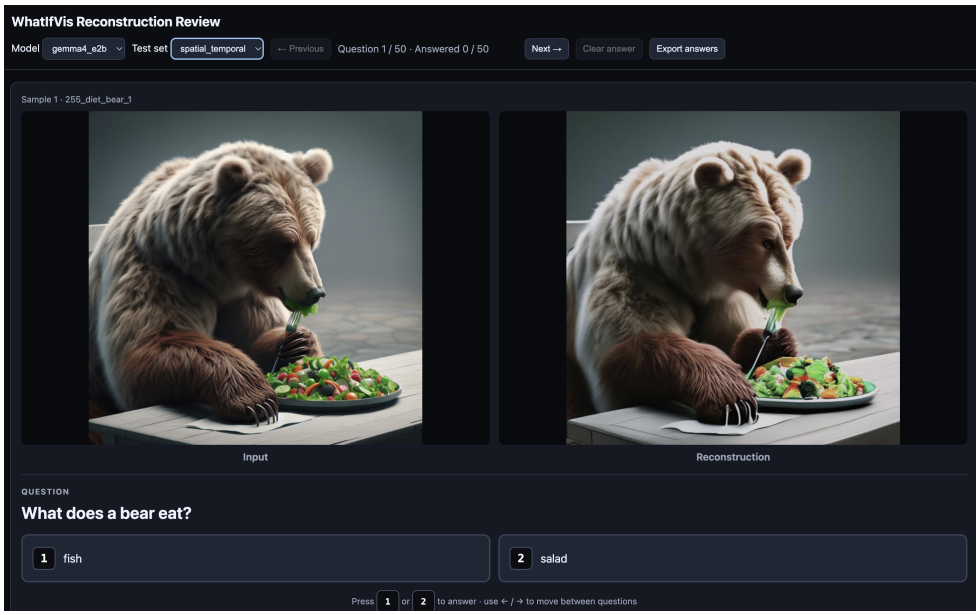


Figure 12: UI for the human evaluation.

standard LoRA multiplier is $\alpha/r = 2$ for every model. LoRA dropout is 0 and the bias is frozen every reported model is trained for one epoch.

Find important layers. Before layer search, we filtered the 102 paired examples in the Spatial-Temporal validation split separately for each SFT checkpoint and modality. A pair was retained only if both instructed prompts preferred their respective complete answers: the context-following prompt had to assign a higher score to the context answer than to the prior answer, and the prior-following prompt had to show the reverse preference. We additionally required the two answers to begin with different token IDs. We patched the output of the attention output-projection module at the final prompt position, replacing all head slices and therefore the complete output vector.

Subspace interventions. For each backbone and modality, we trained a rank-one orthogonal DAS projection at a single validation-selected transformer layer using only the Spatial-Temporal training split. The backbone was frozen, and only the projection was optimized with full-answer teacher forcing in both context-to-prior and prior-to-context directions. We used at most 1,024 eligible training pairs, AdamW with a learning rate of (10^{-3}) and zero weight decay, one epoch, global batch size=16, a maximum sequence length of 512. At inference, the learned projection was transferred to the corresponding adapter-free vanilla model without an intent instruction and applied at the final response-prefix token as the absolute coordinate replacement ($h' = (I - uu^\top)h + c_w u$), with separate constants (c_{ctx}) and (c_{prior}). These constants were tuned using coarse-to-fine grid search over the Spatial-Temporal validation sets and selected using the full-answer pair accuracy (exact match); test examples were never used for selection. We retained configurations with the highest pair accuracy (exact match).

E Extra Results

In this section, we reports additional reconstruction, fine-tuning, activation-patching, and steering results in the image setting.

E.1 Reconstruction Results

Table 1 and 2 reports the counterfactual-attribute recovery for the animal-subset, the only one attribute equipped with matched natural pairs. Here we quantify how faithfully the reconstruction preserves the other coarse attributes, extending the qualitative evidence of Figure 4. For the all subsets in WhatIfVis, we sample 50 samples each and measure pixel-level agreement between

Table 4: **Reconstruction fidelity on WHATIFVIS.** Gemma-4-E2B achieves the strongest pixel-level fidelity across all five attributes. Counterfactual accuracy is measured on 50 sampled reconstructions per attribute; SSIM, PSNR, and LPIPS are computed over every unique image in each test set. CF=counterfactual

			Reconstruction fidelity		
Attribute	**Backbone**	**CF accuracy (%)**	**SSIM↑**	**PSNR↑**	**LPIPS↓**
SPATIAL-TEMPORAL	Qwen2.5-VL-3B	84	0.407	13.53	0.376
	Qwen3.5-4B	88	0.389	13.29	0.375
	Gemma-4-E2B	90	**0.458**	**15.09**	**0.305**
COUNT	Qwen2.5-VL-3B	92	0.485	15.21	0.308
	Qwen3.5-4B	94	0.486	15.69	0.278
	Gemma-4-E2B	96	**0.528**	**16.96**	**0.236**
COLOR	Qwen2.5-VL-3B	94	0.733	15.83	0.241
	Qwen3.5-4B	98	0.765	18.36	0.163
	Gemma-4-E2B	98	**0.795**	**20.21**	**0.133**
SIZE	Qwen2.5-VL-3B	92	0.672	14.56	0.285
	Qwen3.5-4B	96	0.686	15.84	0.237
	Gemma-4-E2B	98	**0.739**	**18.03**	**0.188**
WEIGHT	Qwen2.5-VL-3B	92	0.814	19.04	0.158
	Qwen3.5-4B	92	0.802	18.69	0.161
	Gemma-4-E2B	94	**0.833**	**20.77**	**0.125**

each reconstruction and its counterfactual reference by doing the human evaluation asking them the same questions that send to MLLMs Figure 12. If the annotator answer incorrectly, we assume the image isn't generated correctly. Then we count the number of the correct images in total to get the accuracy. We also measure the reconstruction quality along three complementary axes for all the images in each test sets: SSIM and PSNR for global structure and intensity, and LPIPS for perceptual similarity. See the results in Table 4.

Table 4 shows that on every attribute the coarse structure is recovered, with moderate SSIM/PSNR and low LPIPS. Gemma-4-E2B has the highest SSIM and PSNR and the lowest LPIPS on every attribute, and the highest or tied-highest counterfactual accuracy. Fidelity is highest on WEIGHT and COLOR, and SSIM and LPIPS are worst on SPATIAL-TEMPORAL and COUNT. Reconstruction fidelity therefore does not track controllability: WEIGHT, one of the hardest tasks for the models, has the most faithful reconstructions. Crucially, even the least faithful cases remain well within a range that preserves object identity and layout, confirming that the presence of counterfactual evidence is not specific to COUNT.

E.2 SUPERVISED FINE-TUNING RESULTS

The main text reports the probability scorer throughout; here we show the results under the stricter exact-match scorer (PairEM), which requires the decoded answer under the model-family generation policy to match the gold answer for *both* intents. Figure 13 and Table 5(a) show the per-task exact-match pattern for the SFT lift: the largest gains fall on SIZE, SPATIAL-TEMPORAL, and COLOR and the smallest on COUNT and WEIGHT, as under the probability scorer. From Figure 8, absolute values are lower, since exact-match penalizes formatting mismatches that the probability scorer ignores, but the ordering is preserved, the same vanilla $<$ steer $<$ SFT ordering as under the probability scorer.

E.3 INTERVENTION-LAYER LOCALIZATION

Section 5.2 identifies layer windows for context control with cross-direction activation patching and reports representative curves for two families in Figure 7. Here we give the remaining image-setting curves for Qwen3.5-4B (Figure 14), Gemma-4-E2B-IT (Figure 15), and Qwen2.5-VL-3B/7B (Figures 16 and 17). For each SFT model we run both patching directions on the image channel, $\mathrm{ctx} \to \mathrm{pri}$ (source is the context-intent pass, target is the prior-intent pass) and $\mathrm{pri} \to \mathrm{ctx}$, replacing

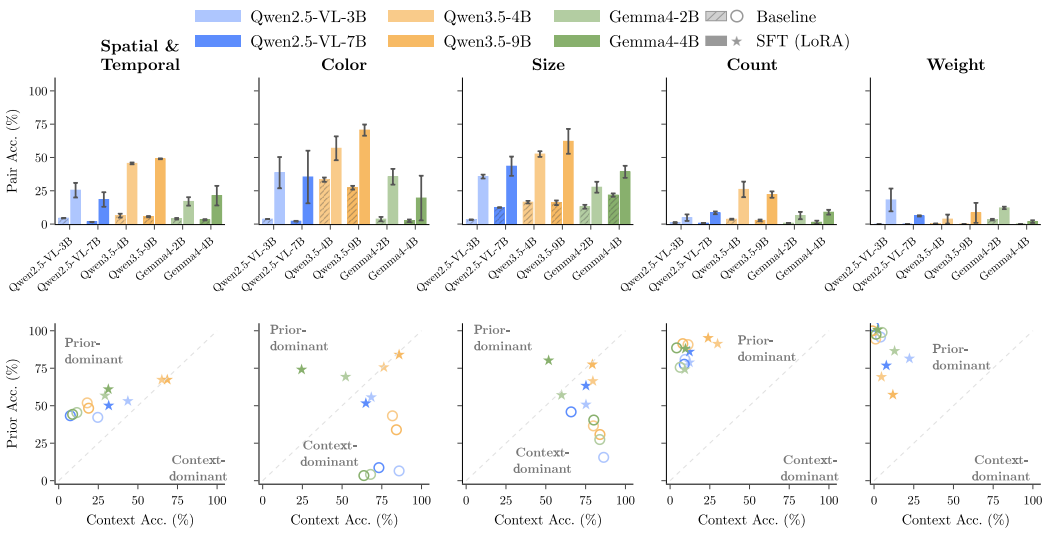
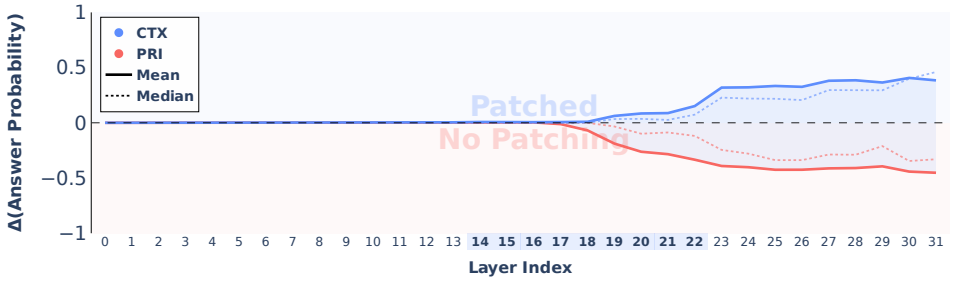
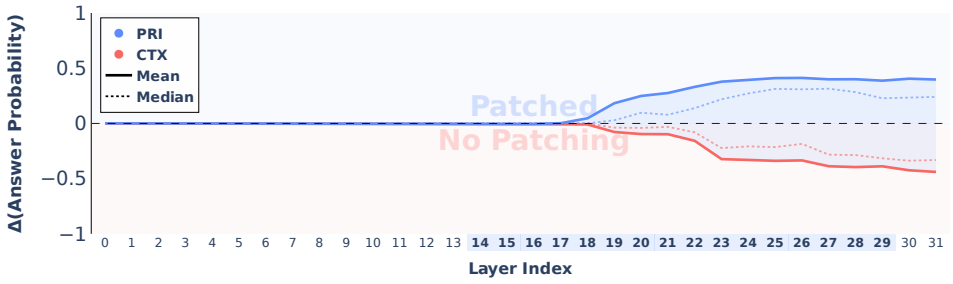


Figure 13: **The SFT lift reproduces under exact-match at lower absolute values in image setting.** The qualitative pattern, with large gains on the perceivable attributes and marginal gains on the reasoning-heavy ones, matches the PairProb (Figure 6).

the multi-head attention outputs at the last-token position of the target pass with those from the source pass. A binary range search (threshold 0.8) then reports the contiguous band of layers whose substitution flips the decision.

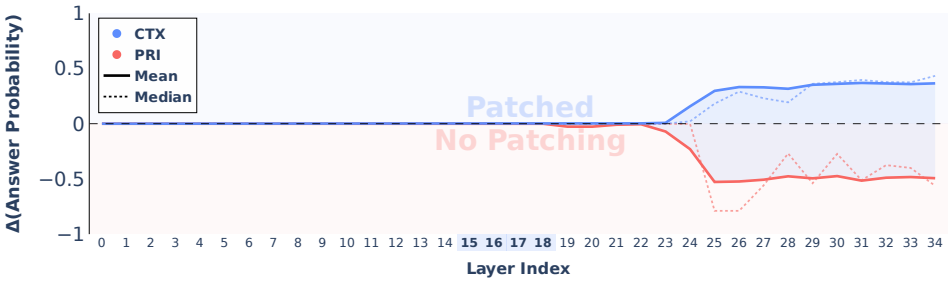


(a) ctx → pri.

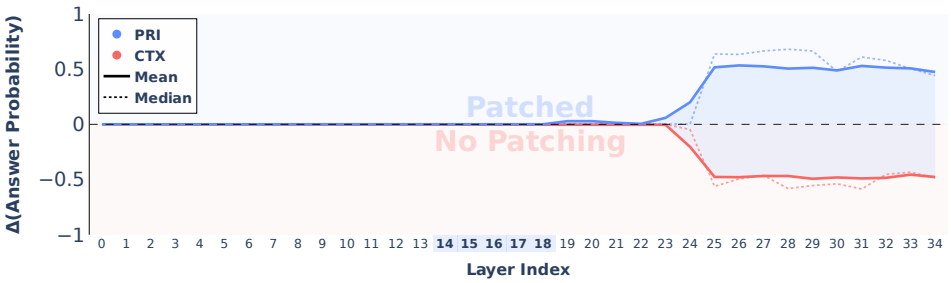


(b) pri → ctx.

Figure 14: Cross-direction patching curves for Qwen3.5-4B in the image setting.

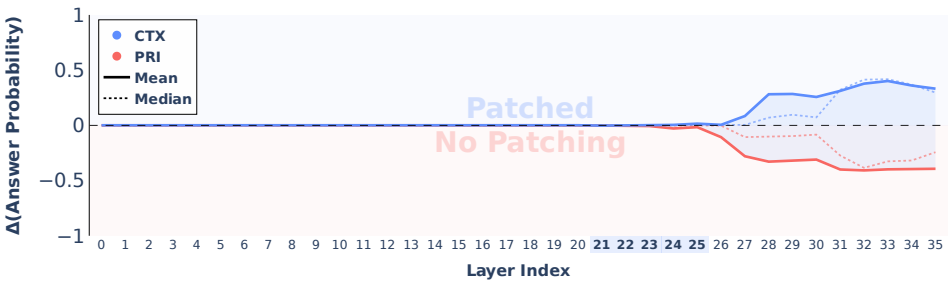


(a) ctx → pri.

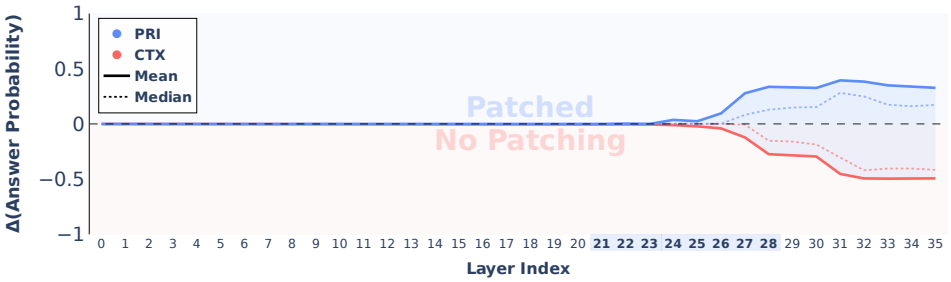


(b) pri → ctx.

Figure 15: Cross-direction patching curves for Gemma-4-E2B-IT in the image setting.

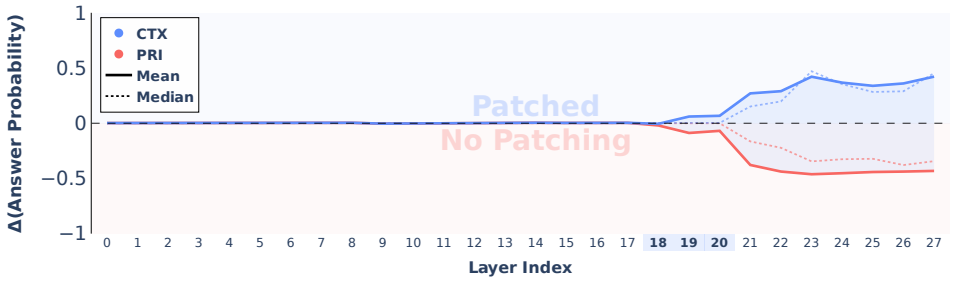


(a) ctx → pri.

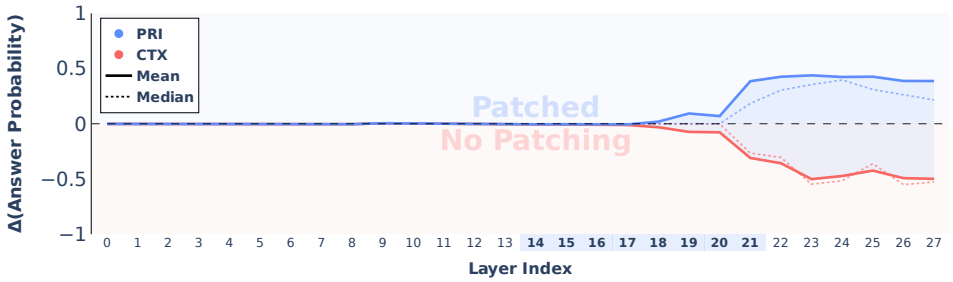


(b) pri → ctx.

Figure 16: Cross-direction patching curves for Qwen2.5-VL-3B in the image setting.



(a) ctx → pri.



(b) pri → ctx.

Figure 17: Cross-direction patching curves for Qwen2.5-VL-7B in the image setting.

Table 5: Pair accuracy (%) on the test splits of the five tasks in the image setting. Baseline (vanilla) and SFT use the intent instruction; PvP and DAS steer the vanilla model *without* it. Avg. is the unweighted mean over the five test sets. S&T denotes SPATIAL-TEMPORAL. We choose the best SFT and baseline checkpoints from three random-seed runs. The best result is shown in **bold**.

(a) PairEM

Model	Method	Color	Count	Size	S&T	Weight	Avg.
Qwen2.5-3B	Baseline	3.72	1.37	3.33	4.33	0.00	2.55
	PvP	**33.95**	**8.22**	13.00	14.67	4.00	14.77
	DAS	28.37	0.68	13.00	**26.33**	7.60	15.20
	SFT	30.70	5.48	**37.33**	25.33	**20.00**	**23.77**
Qwen2.5-7B	Baseline	1.86	0.68	12.33	1.67	0.00	3.31
	PvP	4.65	4.11	8.00	9.67	**15.60**	8.41
	DAS	**32.09**	0.68	15.00	12.00	11.20	14.20
	SFT	23.72	**9.59**	**40.67**	**14.33**	5.60	**18.78**
Qwen3.5-4B	Baseline	33.49	3.42	17.00	7.00	0.00	12.18
	PvP	32.09	4.11	5.67	3.67	0.00	9.11
	DAS	**52.09**	4.79	15.33	12.33	1.20	17.15
	SFT	49.77	**32.19**	**51.67**	**45.33**	**7.60**	**37.31**
Qwen3.5-9B	Baseline	28.84	2.05	18.00	5.67	0.00	10.91
	PvP	34.42	4.11	1.67	4.67	1.60	9.29
	DAS	54.42	3.42	25.67	13.33	2.80	19.93
	SFT	**67.44**	**19.86**	**53.00**	**49.33**	**16.80**	**41.29**
Gemma-E2B	Baseline	4.65	0.68	14.00	4.33	3.20	5.37
	PvP	0.93	**4.79**	27.00	10.00	24.00	13.34
	DAS	0.00	0.00	26.67	4.67	**27.20**	11.71
	SFT	**32.09**	3.42	**32.33**	**16.67**	13.20	**19.54**
Gemma-E4B	Baseline	3.26	0.68	20.33	2.67	0.00	5.39
	PvP	6.98	5.48	11.00	2.00	0.40	5.17
	DAS	**11.16**	0.68	28.67	3.67	**1.60**	9.16
	SFT	3.26	**10.96**	**34.00**	**14.33**	0.80	**12.67**

(b) PairProb

Model	Method	Color	Count	Size	S&T	Weight	Avg.
Qwen2.5-3B	Baseline	6.98	1.37	3.67	14.67	0.40	5.42
	PvP	63.72	14.38	13.33	53.33	3.60	29.67
	DAS	89.77	2.74	17.00	76.33	7.60	38.69
	SFT	**93.49**	**15.75**	**40.67**	**85.67**	**27.20**	**52.56**
Qwen2.5-7B	Baseline	13.95	2.05	13.33	12.67	0.00	8.40
	PvP	66.98	**35.62**	9.33	54.00	**17.20**	36.63
	DAS	83.72	2.05	15.67	54.67	12.40	33.70
	SFT	**97.67**	33.56	**51.33**	**93.00**	13.20	**57.75**
Qwen3.5-4B	Baseline	53.02	2.05	13.67	21.67	0.40	18.16
	PvP	60.93	6.16	3.33	18.33	0.00	17.75
	DAS	88.37	4.79	17.00	43.67	0.40	30.85
	SFT	**93.95**	**63.70**	**62.33**	**88.00**	**16.40**	**64.88**
Qwen3.5-9B	Baseline	36.28	2.74	13.67	16.67	0.00	13.87
	PvP	59.53	5.48	7.33	27.33	1.20	20.18
	DAS	73.02	4.11	24.33	54.00	2.00	31.49
	SFT	**99.53**	**30.14**	**61.67**	**93.33**	**41.20**	**65.17**
Gemma-E2B	Baseline	33.49	3.42	13.00	25.00	3.20	15.62
	PvP	75.35	17.12	26.33	47.33	23.20	37.87
	DAS	43.26	0.00	26.67	54.00	**29.60**	30.70
	SFT	**98.60**	**19.86**	**43.00**	**76.67**	25.60	**52.75**
Gemma-E4B	Baseline	50.23	2.74	24.33	38.33	0.00	23.13
	PvP	88.37	15.75	14.67	37.00	0.00	31.16
	DAS	95.35	0.00	35.67	63.67	4.00	39.74
	SFT	**96.28**	**34.25**	**72.33**	**85.33**	**8.80**	**59.40**

E.4 STEERING VECTOR RESULTS

Section 5.3 reports that a single learned vector, applied with no intent instruction, raises macro PairProb. In Table 5, we provide all the results in both PairEM and PairProb across five held-out test sets. We also adapt the PvP Golovanevsky et al. (2025) to the WHATIFVIS. To explore the

Table 6: Comparison of the baseline, PvP, and our DAS steering method on four multimodal benchmarks. MME-P is reported as a raw score; all other metrics are percentages. Baselines are shown in gray. DAS denotes DAS context steering with the intervention layer and model-specific coefficient shown in parentheses. PvP uses the selected `pvp-sft-to-base` direction. The final two rows report the mean performance drop, Baseline − Method, across the six models; lower is better. Best results are **bolded**, and second-best results are underlined.

Model	Method	MME-P ↑	MMBench ↑	SEED-Image ↑	MMMU-val ↑
Qwen2.5-VL-3B	Baseline	**1542.58**	**78.35**	**74.97**	**47.22**
	PvP (L24, $c = 1$)	1508.48	75.95	74.52	45.78
	DAS (L25, $c = -68$)	1364.99	66.32	70.75	43.22
Qwen2.5-VL-7B	Baseline	**1674.68**	**83.93**	**77.36**	**51.89**
	PvP (L18, $c = 2$)	1099.14	28.78	65.36	38.11
	DAS (L20, $c = -154$)	1630.46	82.56	76.42	50.11
Qwen3.5-4B	Baseline	**1702.76**	**78.69**	**77.71**	**52.11**
	PvP (L13, $c = 4$)	1668.31	74.14	73.49	43.22
	DAS (L18, $c = -8$)	1614.24	**78.69**	73.12	49.00
Qwen3.5-9B	Baseline	**1715.57**	84.28	**77.96**	48.33
	PvP (L15, $c = 4$)	1650.27	**86.00**	77.29	**52.78**
	DAS (L18, $c = 31.5$)	1608.00	84.62	76.05	52.33
Gemma-4-E2B	Baseline	**801.16**	61.51	**67.39**	**42.11**
	PvP (L18, $c = 2$)	334.19	36.34	61.10	35.44
	DAS (L16, $c = 39$)	398.45	**64.18**	26.47	41.00
Gemma-4-E4B	Baseline	**1123.20**	**70.70**	**73.22**	**46.44**
	PvP (L27, $c = 1$)	1033.73	68.81	71.69	46.33
	DAS (L26, $c = -80$)	1110.68	57.47	65.19	43.89
Average $\Delta \downarrow$	PvP	210.97	14.57	**4.19**	4.41
	DAS	**138.86**	**3.94**	10.10	**1.43**

impact of the steering vector, we evaluate the vanilla models with the steering vector on the other four general multimodal understanding benchmarks including MME Fu et al. (2026), MMBench Liu et al. (2024), SEED Li et al. (2023a), MMMU Yue et al. (2024).

F Comparison with Textual Context

Section 6 presents the modality comparison; here we give more results in text setting, a comparison of intervention layers across the image and text conditions. Every counterfactual image in WhatIfVis is paired with a one-sentence statement describing the same task-relevant fact. We repeat the baseline, SFT, and steering evaluations with these statements in place of images. Because the statement gives the relevant fact explicitly, the text condition provides a reference for instruction following without visual extraction.

F.1 Text-Setting Results

In this section, we report the results in text-only settings, by using the text statements instead of the images in WhatIfVis. The results of the vanilla models and SFT models are shown in Figure 18 and Figure 19. From Figure 18 and Figure 19, the models prefer prior knowledge learned during the pretraining in Size and Weight, while more rely on the context in Color and Spatial-Temporal. Table 7 shows the PairEM and PairProb in details.

F.2 Text-Setting Intervention Layers

We identify intervention layers in the text condition with activation patching in both directions. The following figures report the two directions separately for each model so that the curves remain readable at appendix scale.

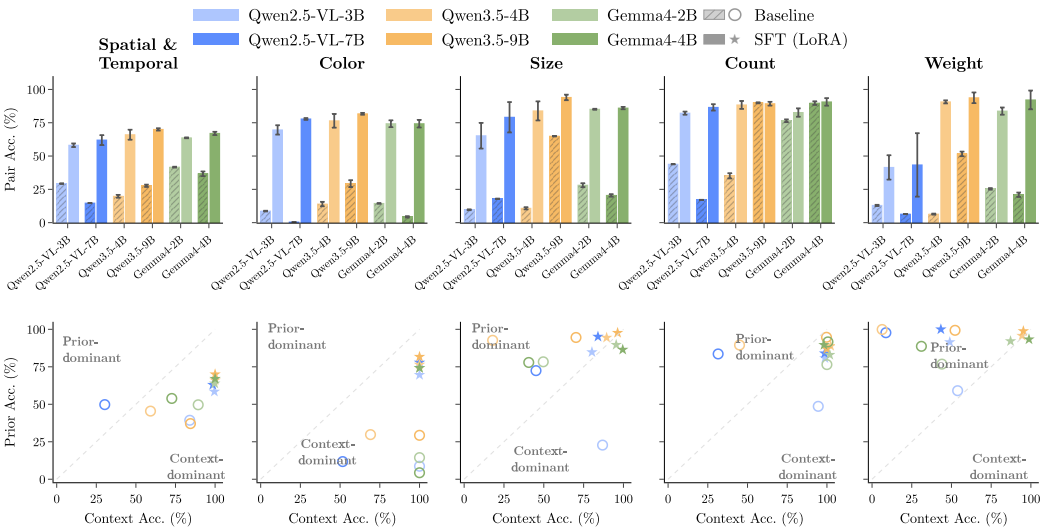


Figure 18: **The SFT improves** PairEM **in text setting. Top row:** PairEM per task. **Bottom row:** the same runs as context-accuracy (x) against prior-accuracy (y); points above the diagonal follow the prior over the text statement.

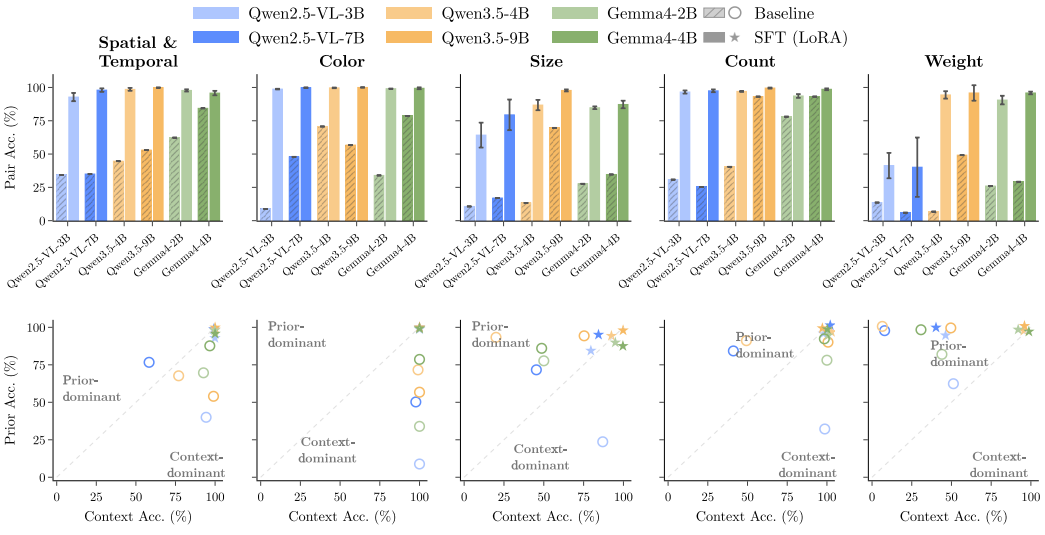
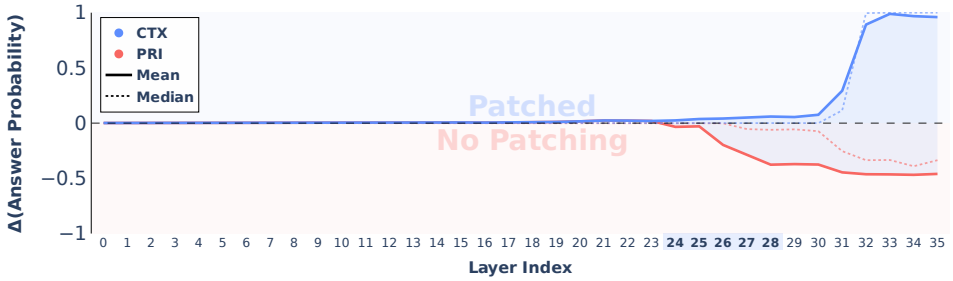
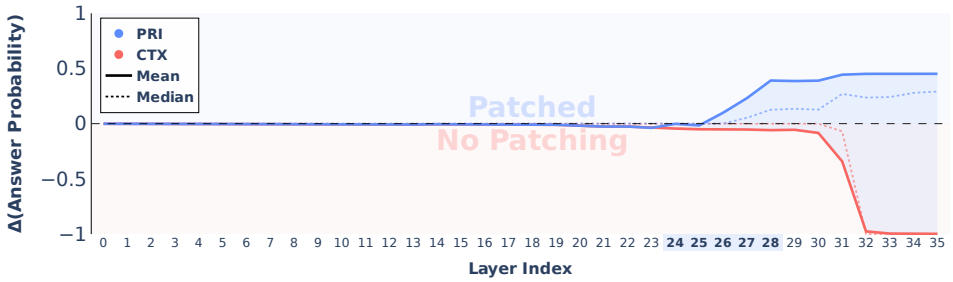


Figure 19: **SFT lifts textual controllability across every model and task in text setting. Top row:** PairProb per task. **Bottom row:** the same runs as context-accuracy (x) against prior-accuracy (y); points above the diagonal follow the prior over the text statement. SFT lifts controllability consistently.

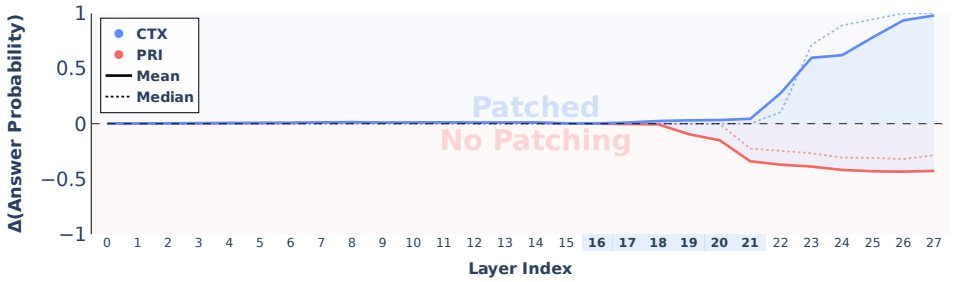


(a) ctx → pri.

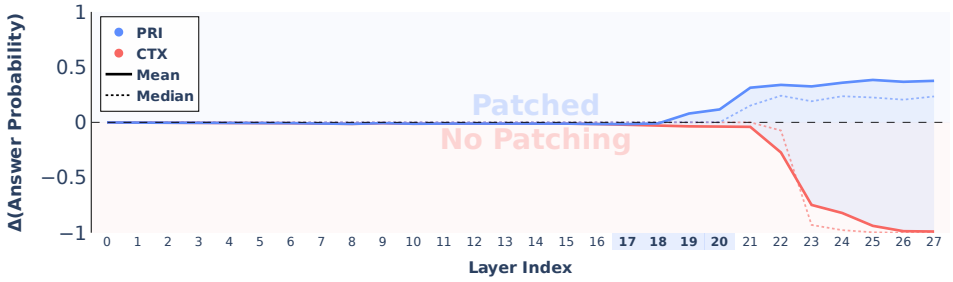


(b) pri → ctx.

Figure 20: Cross-direction patching curves for Qwen2.5-VL-3B in the text setting. Curve styles and answer labels follow Figure 7.

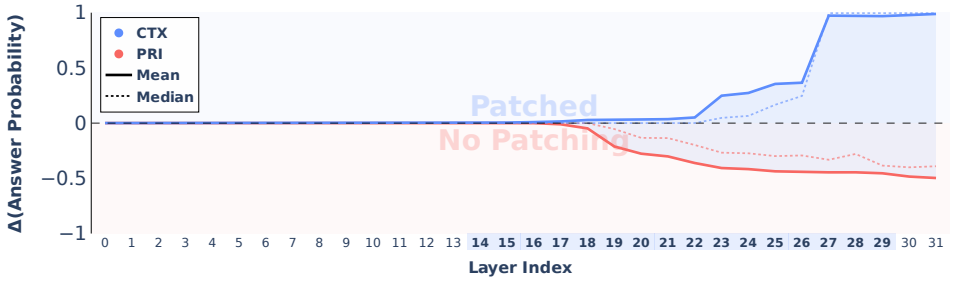


(a) ctx → pri.

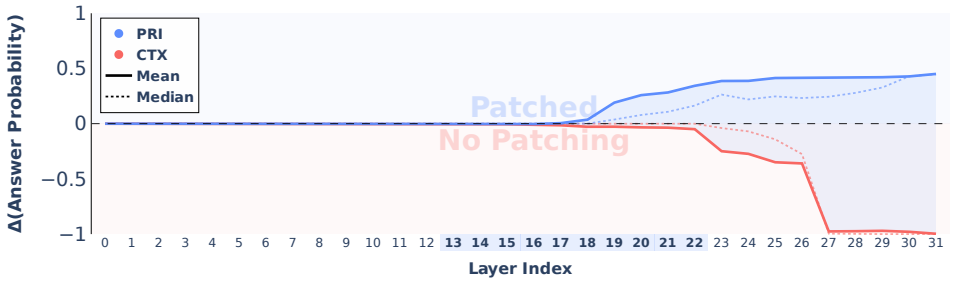


(b) pri → ctx.

Figure 21: Cross-direction patching curves for Qwen2.5-VL-7B in the text setting. Curve styles and answer labels follow Figure 7.

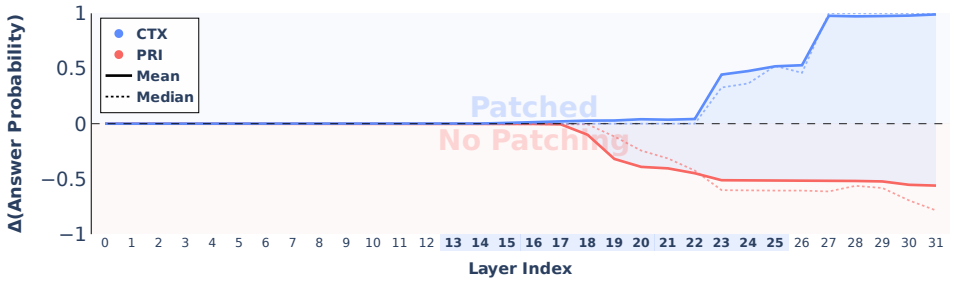


(a) ctx → pri.

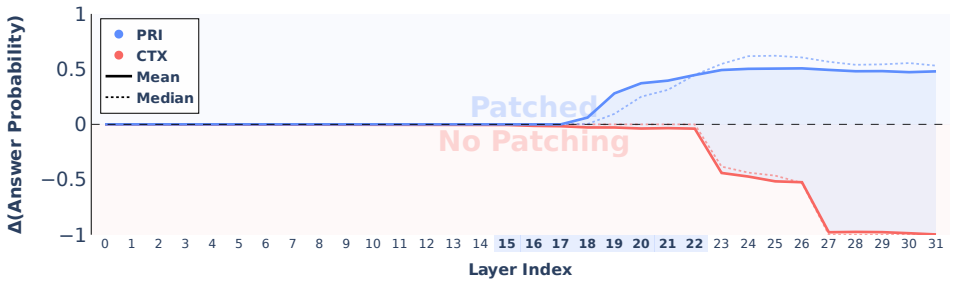


(b) pri → ctx.

Figure 22: Cross-direction patching curves for Qwen3.5-4B in the text setting. Curve styles and answer labels follow Figure 7.

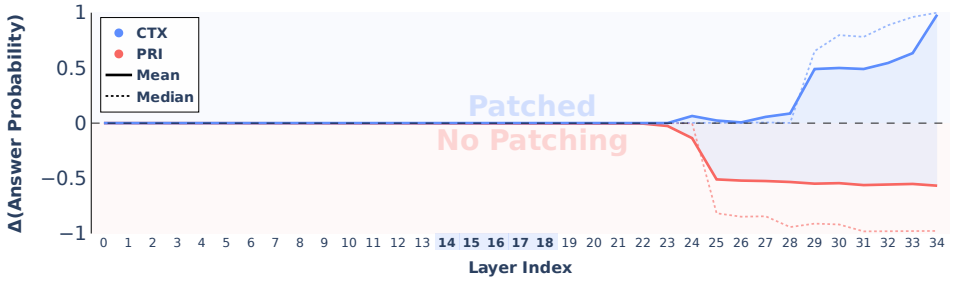


(a) ctx → pri.

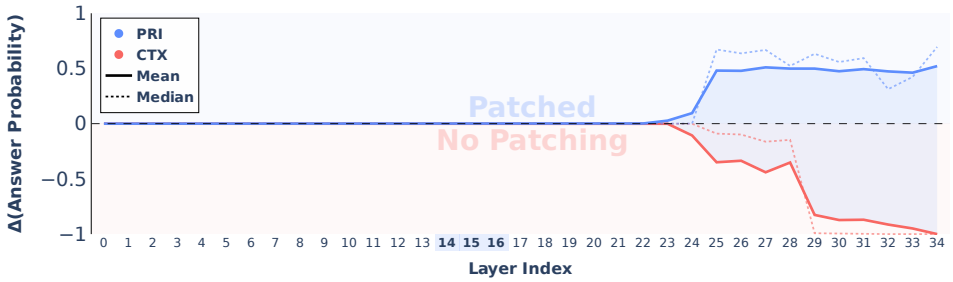


(b) pri → ctx.

Figure 23: Cross-direction patching curves for Qwen3.5-9B in the text setting. Curve styles and answer labels follow Figure 7.

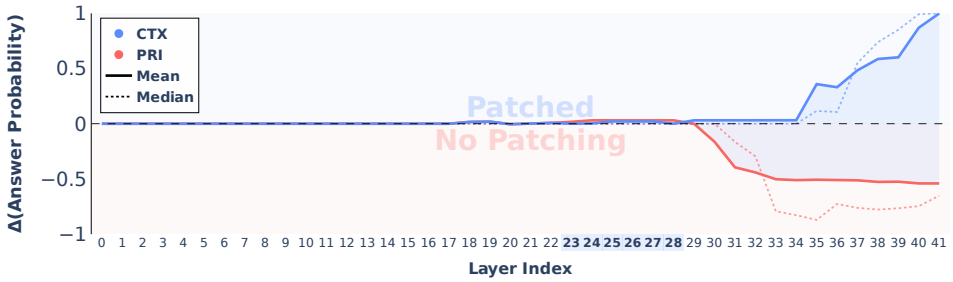


(a) ctx → pri.

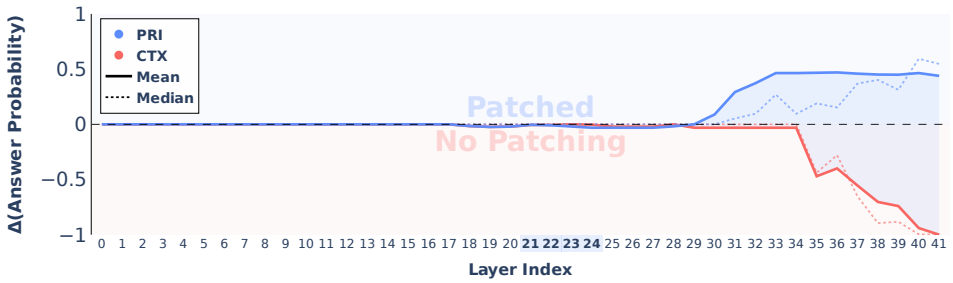


(b) pri → ctx.

Figure 24: Cross-direction patching curves for Gemma-4-E2B-IT in the text setting. Curve styles and answer labels follow Figure 7.



(a) ctx → pri.



(b) pri → ctx.

Figure 25: Cross-direction patching curves for Gemma-4-E4B-IT in the text setting. Curve styles and answer labels follow Figure 7.

Table 7: Pair accuracy (%) on the test splits of the five tasks in the **text** setting. Baseline (vanilla) and SFT use the intent instruction; DAS steers the vanilla model *without* it. Avg. is the unweighted mean over the five test sets. S&T denotes SPATIAL-TEMPORAL. We choose the best SFT and baseline checkpoints from three random-seed runs. The best result is shown in **bold**.

(a) PairEM

Model	Method	Color	Count	Size	S&T	Weight	Avg.
Qwen2.5-3B	Baseline	8.84	43.84	10.00	29.33	13.20	21.04
	DAS	66.05	21.23	17.67	**59.33**	38.00	40.46
	SFT	**70.70**	**82.88**	**75.00**	58.67	**46.40**	**66.73**
Qwen2.5-7B	Baseline	0.47	17.12	18.00	14.67	6.40	11.33
	DAS	26.51	33.56	12.00	23.67	34.80	26.11
	SFT	**78.14**	**86.30**	**89.67**	**62.67**	**65.20**	**76.39**
Qwen3.5-4B	Baseline	15.81	32.88	10.00	21.00	6.80	17.30
	DAS	40.93	85.62	27.67	28.33	22.80	41.07
	SFT	**77.67**	**89.73**	**88.67**	**67.33**	**92.40**	**83.16**
Qwen3.5-9B	Baseline	31.63	**89.73**	64.67	28.33	50.40	52.95
	DAS	21.40	83.56	62.33	25.67	47.60	48.11
	SFT	**83.26**	89.04	**92.33**	**72.33**	**96.80**	**86.75**
Gemma-E2B	Baseline	14.42	77.40	28.00	41.67	24.80	37.26
	DAS	20.93	0.00	72.33	40.00	79.60	42.57
	SFT	**75.35**	**80.14**	**87.67**	**64.33**	**86.40**	**78.78**
Gemma-E4B	Baseline	4.65	91.10	21.00	38.00	20.40	35.03
	DAS	15.81	2.74	32.00	21.00	46.80	23.67
	SFT	**72.09**	**91.78**	**87.67**	**66.67**	**94.80**	**82.60**

(b) PairProb

Model	Method	Color	Count	Size	S&T	Weight	Avg.
Qwen2.5-3B	Baseline	8.84	30.82	10.67	34.33	13.60	19.65
	DAS	**99.07**	39.04	20.67	85.00	45.20	57.80
	SFT	98.60	**97.26**	**74.00**	**95.33**	**47.20**	**82.48**
Qwen2.5-7B	Baseline	47.91	25.34	17.00	35.00	6.00	26.25
	DAS	97.21	56.85	19.00	80.67	37.60	58.27
	SFT	**100.00**	**96.58**	**89.00**	**99.67**	**60.80**	**89.21**
Qwen3.5-4B	Baseline	70.70	40.41	13.33	44.67	6.80	35.18
	DAS	97.21	86.30	32.67	71.67	23.20	62.21
	SFT	**99.53**	**97.26**	**90.33**	**99.33**	**98.00**	**96.89**
Qwen3.5-9B	Baseline	56.74	93.15	69.67	53.00	49.20	64.35
	DAS	92.56	86.99	73.33	69.33	60.80	76.60
	SFT	**100.00**	**100.00**	**97.33**	**99.67**	**100.00**	**99.40**
Gemma-E2B	Baseline	33.95	78.08	27.67	62.33	26.00	45.61
	DAS	67.91	0.00	72.67	70.00	85.20	59.15
	SFT	**99.07**	**95.89**	**85.67**	**98.67**	**94.40**	**94.74**
Gemma-E4B	Baseline	78.60	93.15	34.67	84.33	29.20	63.99
	DAS	70.23	0.00	74.67	79.67	75.20	59.95
	SFT	**99.53**	**99.32**	**89.67**	**97.33**	**96.80**	**96.53**

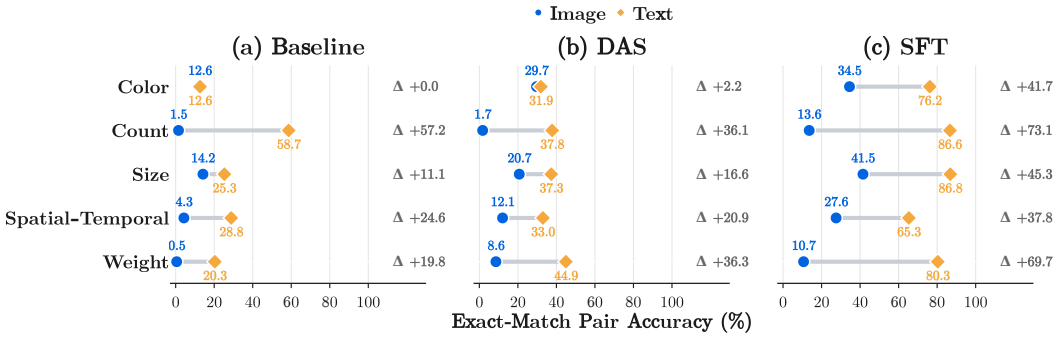


Figure 26: PairEM gap between image and text settings across five attributes.

F.3 PER-TASK IMAGE–TEXT GAPS

Figures 26 and 27 compare the image–text gap separately for the exact-match and probability scorers.

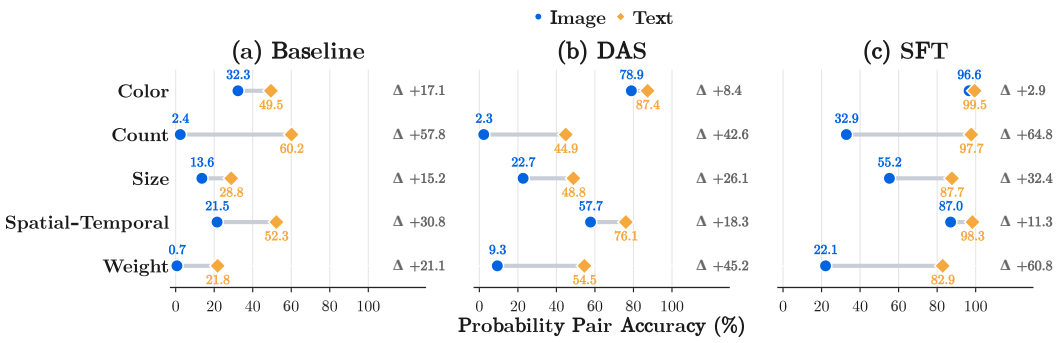


Figure 27: PairProb gap between image and text settings across five attributes.

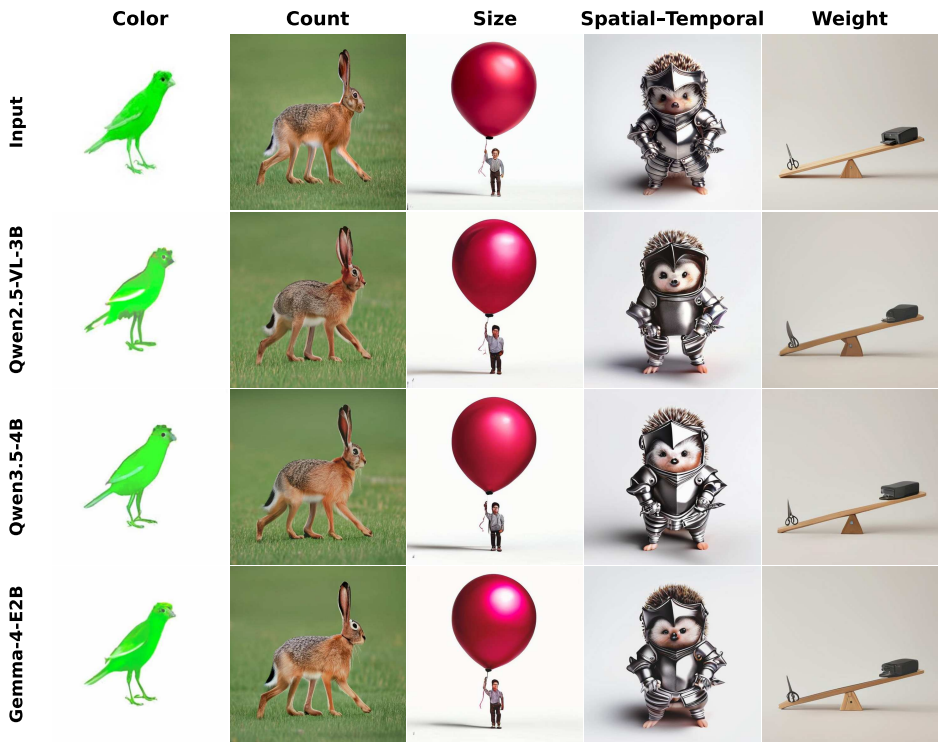


Figure 28: More qualitative reconstruction results across five attributes in WHATIFVIS (I).

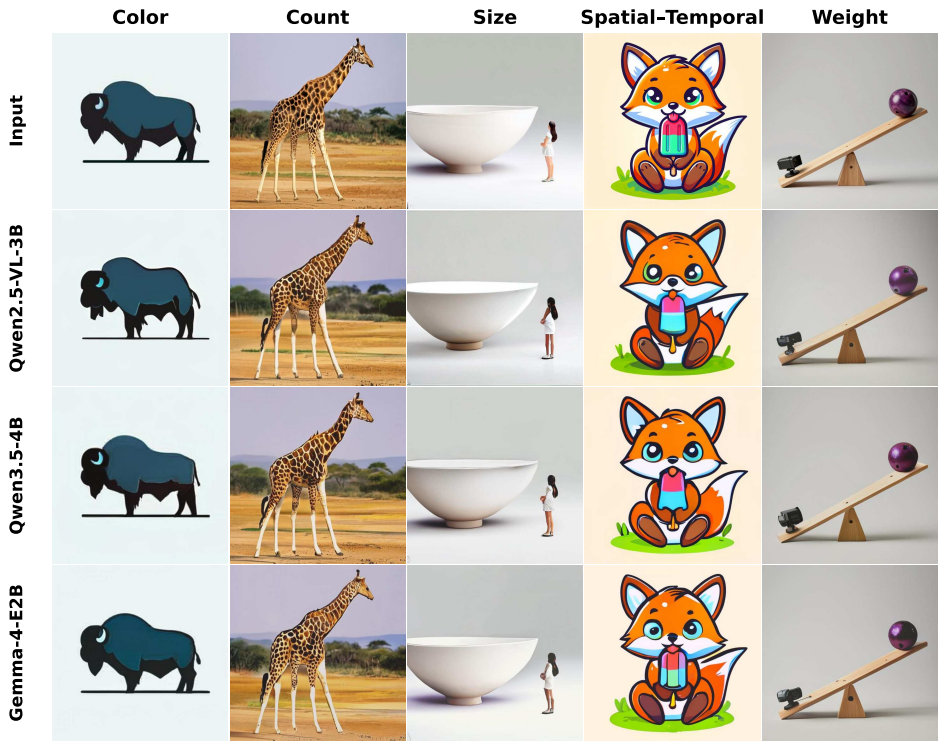


Figure 29: More qualitative reconstruction results across five attributes in WHATIFVIS (II).

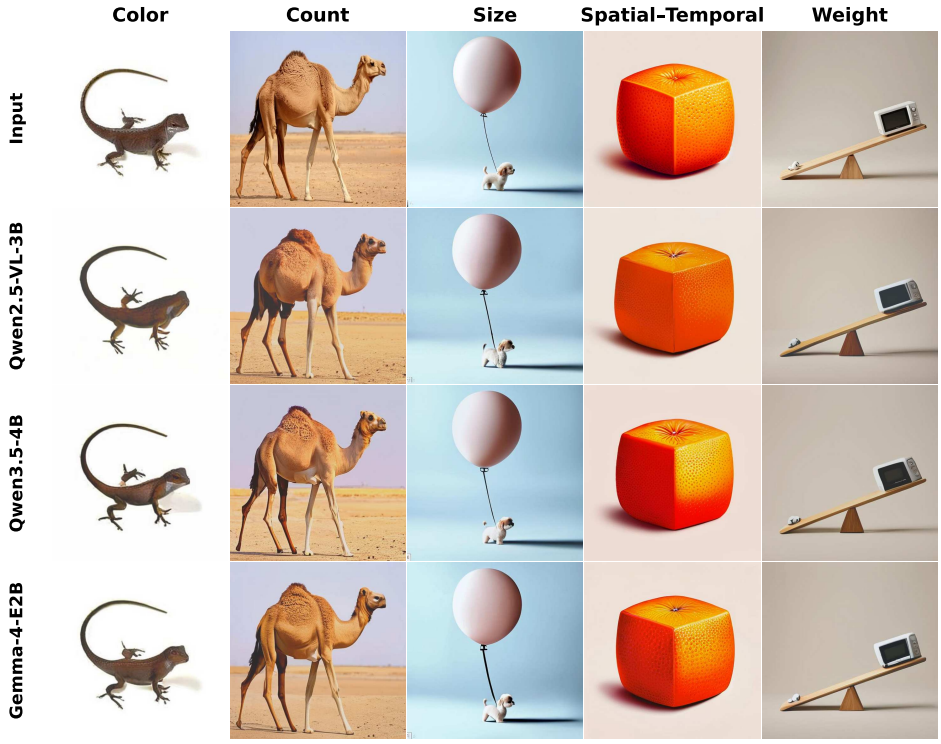


Figure 30: More qualitative reconstruction results across five attributes in WHATIFVIS (III).

G PROMPTS

Image version

System: Always answer the question in one concise sentence. When the input contains an instruction, follow the instruction exactly while staying concise.
User: <image>*N
Instruction: Consider only the image to answer the question.
Question: What is the Eiffel Tower surrounded by?
Assistant: The Eiffel Tower is surrounded by

System: Always answer the question in one concise sentence. When the input contains an instruction, follow the instruction exactly while staying concise.
User: <image>*N
Instruction: Ignore the image to answer the question.
Question: What is the Eiffel Tower surrounded by?
Assistant: The Eiffel Tower is surrounded by

Text version

System: Always answer the question in one concise sentence. When the input contains an instruction, follow the instruction exactly while staying concise.
User: Statement: The Eiffel Tower is surrounded by the iceberg.
Instruction: Consider only the statement to answer the question.
Question: What is the Eiffel Tower surrounded by?
Assistant: The Eiffel Tower is surrounded by

System: Always answer the question in one concise sentence. When the input contains an instruction, follow the instruction exactly while staying concise.
User: Statement: The Eiffel Tower is surrounded by the iceberg.
Instruction: Ignore the statement to answer the question.
Question: What is the Eiffel Tower surrounded by?
Assistant: The Eiffel Tower is surrounded by