TEST-TIME TRAINING FOR MODALITY ORDER CONSISTENCY IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We find that vision-language models are sensitive to a specific semantically irrelevant change: the order in which the image and question are presented. Across three models on three core benchmarks, with additional evaluation on three further benchmarks, image-first prompting consistently outperforms question-first prompting, revealing a repeatable modality order failure. We use this gap to design an order-consistent test-time training method. Our method substantially closes the modality-order gap across all evaluated settings. The stronger image-first branch is largely preserved, with observed improvements in several settings, hence bootstrapping both orderings toward mutual consistency. Activation patching localizes the ordering failure to a narrow mid-network region where representations diverge sharply between prompt orders. We find that the test-time training method repairs this misalignment across layers. Together, our results identify modality-order sensitivity as a circuit-level failure in VLMs and demonstrate that simple, asymmetric test-time adaptation can effectively mitigate it and even improve performance over the baseline.

1 INTRODUCTION

Vision-language models (VLMs) have advanced rapidly in recent years, achieving strong performance across a broad range of multimodal tasks, including visual question answering, image captioning, and multimodal reasoning (Bai et al., 2025; Chen et al., 2024b; Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023). These models process images and text jointly, but the precise mechanisms by which visual and linguistic information are integrated remain poorly understood. In particular, it is unclear how the relative ordering of modalities in the input affects internal computation, and whether models remain robust when the semantic content and underlying multimodal task remain unchanged.

We identify that modern VLMs are very sensitive to modality order in prompts: presenting the same image and question in different orders, that is, swapping the textual query and the image component of the prompt, produces substantially different predictions on a significant fraction of examples across multiple model families and datasets. We find that image-first prompting consistently outperforms query-first with accuracy gaps ranging from 6 to 26 percentage points. A semantically irrelevant perturbation thus induces a repeatable and structured performance asymmetry, suggesting a systematic difference in the internal computations elicited by two prompt orderings.

We propose a simple test-time training (TTT) procedure to repair this inconsistency. For each test instance, we construct both orderings and treat the stronger image-first variant as a teacher signal. We then perform a small number of gradient steps that align the question-first prediction distribution toward the image-first one, using a KL divergence test-time objective that keeps the teacher fixed. Our procedure instead uses the stronger computation as an anchor and selectively repairs the weaker one, without requiring labeled data, external supervision, or task-specific fine-tuning at test time.

After a small number of test-time gradient steps, our method substantially closes the question-first accuracy gap across all three models and datasets. More surprisingly, image-first accuracy also improves slightly in most settings, suggesting that the adaptation procedure does not merely overfit to one ordering but produces a more internally consistent model. These results hold across model families and benchmark types, indicating that the phenomenon and its repair are not artifacts of any particular architecture, dataset, or evaluation setting.

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.

To understand where the ordering failure arises and how repair works internally, we combine activation patching with a TTT ablation over contiguous layer windows, where each window restricts gradient updates to a fixed block of transformer layers. Activation patching reveals that the failure is causally concentrated in a narrow mid-network band: replacing query-first hidden states with image-first states in early layers has almost no effect, while patching within a specific mid-layer window sharply restores accuracy. Our window ablation of test-time training is consistent with this localization: adaptation targeted at mid-network layers achieves faster convergence and higher question-first accuracy than adaptation at earlier or later layers. We additionally find that the asymmetric objective is essential: symmetric alternatives that treat both orderings as equally reliable recover substantially less of the ordering gap, confirming that the directionality must be reflected in the adaptation objective.

Swapping the order of image and question should not materially change a vision-language model's answer, yet in practice it often does. We show that this semantically irrelevant prompt perturbation induces a repeatable circuit-level failure in VLMs, that the failure is causally localizable, and that a directional test-time repair can use the stronger image-first branch to effectively repair predictions produced by the weaker question-first branch.

In summary, our key contributions are :

- We identify a circuit-level failure mode in vision-language models induced by modality ordering: identical image-question pairs can produce substantially different predictions when the order of the two modalities is swapped.
- Using activation patching, we show that the ordering failure is not diffuse, but is causally localized to a narrow depth regime.
- We propose a mechanism-guided asymmetric test-time repair procedure that uses the stronger image-first variant as teacher to repair the weaker question-first variant, closing most of the ordering gap while largely preserving the stronger image-first behavior.

2 Related Work

Vision Language Models. Modern vision-language models integrate visual and linguistic information through a shared transformer backbone, either via cross-attention mechanisms (Alayrac et al., 2022; Li et al., 2023) or by projecting visual features into the language model's token space (Liu et al., 2023; Dai et al., 2023). Despite strong performance across multimodal benchmarks, the precise mechanisms by which these models fuse modalities remain poorly understood. Prior work has identified sensitivity to prompt formatting (Huang et al., 2023), and recent studies suggest that VLMs may rely disproportionately on text over visual input in certain settings (Deng et al., 2025) and that a performance gap persists between analogous visual and textual tasks across model families (Nikankin et al., 2025b). Our work contributes to this understanding by identifying modality order as a structured and repeatable source of inconsistency, and we propose a test-time training method that closes this gap without harming stronger ordering, and surprisingly, slightly improves it.

Test-time training. Our method builds on test-time training (Sun et al., 2020; Gandelsman et al., 2022), which adapts model parameters at inference using self-supervised signals from each test instance without labeled data. Subsequent methods minimize entropy (Wang et al., 2020), enforce augmentation consistency (Zhang et al., 2022), or adapt vision-language models via prompt learning (Zhou et al., 2022b;a; Khattak et al., 2023). Most of these settings focus on the adapted distribution. Our method produces a qualitatively different effect: because the teacher is recomputed from updated weights at each step, improvements in question-first feed back into a sharper image-first teacher, bootstrapping both orderings toward mutual consistency. Chou et al. (2026a) propose a symmetric consistency objective for VLMs. Through empirical results, we show that treating both orderings as equally reliable recovers substantially less of the observed modality-order performance gap. More broadly, consistency constraints have been used to preserve underlying content across transformations (Zhu et al., 2017). Our setting similarly treats the two modality orderings as semantics-preserving views, but uses a directional consistency objective because image-first prompting is empirically more reliable.

Prompt Sensitivity. LLMs are sensitive to prompt formatting, including option order (Zhao et al., 2021), instruction phrasing (Mizrahi et al., 2024), and the ordering of in-context examples (Lu

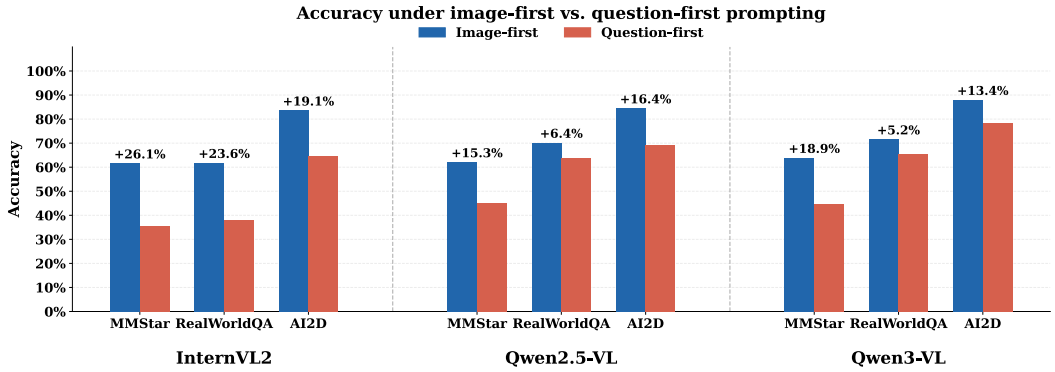


Figure 1: **Accuracy under image-first and question-first prompting.** Image-first prompting consistently outperforms question-first prompting across datasets and model families.

et al., 2022). In VLMs, sensitivity to prompt format has been documented across visual question answering (Liu et al., 2024) and multimodal reasoning (Chen et al., 2024a) tasks. Closest to our work, Deng et al. (2025) and Hejabi et al. (2025) document that image-first prompting outperforms question-first across models and tasks. These works establish the phenomenon. Our work goes further by localizing the failure via activation patching, and proposing a principled test-time repair that can even improve the stronger ordering.

3 Test-Time Training for Order Consistency

In this section, we introduce a label-free test-time training method for reducing modality-order sensitivity in VLMs. We first document the modality-ordering failure, then present our asymmetric adaptation objective and evaluate its effect on both prompt orders.

3.1 Modality-Ordering Failure in Vision-Language Models

For each benchmark instance, we evaluate the same image-question pair under two prompt orders: image-first and question-first. Figure 1 shows that image-first prompting achieves higher accuracy across the evaluated datasets and model families. The two orderings also disagree on a substantial fraction of examples, indicating that modality order changes model behavior even when the underlying image, question, answer choices, and semantic task remain unchanged.

The gap is largest on harder multimodal reasoning benchmarks, but remains visible on natural visual question-answering tasks. Since the two prompts contain the same image and question, the accuracy gap must arise from differences in the computation induced by the two input orders.

3.2 Test-Time Training Consistency Objective

The asymmetry in Fig. 1 motivates a directional adaptation strategy. Rather than forcing both prompt orders toward a common midpoint, we use the stronger image-first prediction as a detached teacher and adapt the question-first prediction toward it. For each test instance, we perform a small number of gradient updates using an asymmetric consistency loss between the two output distributions. This makes the update label-free, instance-specific, and targeted at the weaker ordering identified above.

Setup. Let f_θ denote a vision-language model with parameters θ. For a test instance

$$s = (x, q, \mathcal{A}),$$

where x is an image, q is a question, and $\mathcal{A}$ is the set of answer candidates, we define two prompt orderings:

$$o_{\mathrm{IF}} = \text{image-first}, \qquad o_{\mathrm{QF}} = \text{question-first}.$$

The model produces logits over answer candidates:

$$z_{\mathrm{IF}}(\theta; s) = f_\theta(s, o_{\mathrm{IF}}), \qquad z_{\mathrm{QF}}(\theta; s) = f_\theta(s, o_{\mathrm{QF}}).$$

We convert logits to predictive distributions,

$$p_{\mathrm{IF}}(\theta; s) = \mathrm{softmax}(z_{\mathrm{IF}}(\theta; s)), \qquad p_{\mathrm{QF}}(\theta; s) = \mathrm{softmax}(z_{\mathrm{QF}}(\theta; s)).$$

For each test instance, we initialize from the pretrained parameters $\theta^{(0)} = \theta_0$ and perform K adaptation steps, where each adaptation step is one gradient update on that same instance. At step k, we recompute the image-first distribution using the current parameters and detach it:

$$\bar{p}_{\mathrm{IF}}^{(k)}(s) = \mathrm{stopgrad}\Big(p_{\mathrm{IF}}(\theta^{(k)}; s)\Big).$$

Thus, the teacher is fixed within each gradient step but updated across steps. We use this moving-teacher variant by default; Table 5 shows that recomputing the teacher substantially outperforms keeping the initial image-first prediction fixed. After evaluating the adapted model on both prompt orders, we reset the parameters to θ_0 before processing the next example. This reset ensures that the adaptation remains label-free, instance-specific, and independent of all previous examples.

Objective. At step k, our alignment loss is

$$\mathcal{L}_{\mathrm{align}}(\theta^{(k)}; s) = D_{\mathrm{KL}}\left(\bar{p}_{\mathrm{IF}}^{(k)}(s) \,\middle\|\, p_{\mathrm{QF}}(\theta^{(k)}; s)\right).$$

We perform gradient updates of the form

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla_{\theta^{(k)}} \mathcal{L}_{\mathrm{align}}(\theta^{(k)}; s), \qquad k = 0, \ldots, K-1.$$

After adaptation, we evaluate the updated model on the target prompt order and then reset to θ_0 before beginning adaptation on the subsequent test instance.

This objective is asymmetric: gradients flow through the question-first prediction, while the image-first distribution acts as a detached teacher at each step. This encodes the empirical direction of the failure in Fig. 1; image-first provides the anchor, and question-first is adapted toward it.

In principle, the update can be applied to all model parameters. In practice, our mechanistic analysis in Section 5 shows that the modality-ordering failure is concentrated in a specific depth range. We therefore compare adaptation over early, middle, and late layer windows, and find that mid-layer windows give the best trade-off between question-first recovery and image-first performance. This turns the causal analysis into a mechanism-guided test-time repair procedure, rather than a generic model-wide adaptation step applied uniformly throughout the entire network.

4 Experiments

We evaluate whether asymmetric order-consistency TTT reduces the question-first accuracy gap identified in Section 3.1. Across three VLMs spanning two model families and three benchmarks, our method consistently improves question-first accuracy while slightly improving performance under the stronger image-first ordering, without introducing a trade-off between the two.

4.1 Experimental Setup

Models. We evaluate on three recent open-source vision-language models spanning two model families. **InternVL2-8B** (Chen et al., 2024b) uses a ViT-based visual encoder projected into the language model's token space via an MLP connector. **Qwen2.5-VL-7B** (Yang et al., 2025) and **Qwen3-VL-8B** (Bai et al., 2025) use a native-resolution vision encoder with dynamic image tiling. All models are instruction-tuned and evaluated in zero-shot mode without task-specific fine-tuning.

Datasets. We evaluate on MMStar (Chen et al., 2024a), RealWorldQA, and AI2D (Kembhavi et al., 2016). These benchmarks vary in visual dependency and language-prior strength. MMStar contains 1,500 samples filtered to require visual understanding. RealWorldQA contains 765 real-world perception questions. AI2D contains 3,088 diagram questions and has stronger language priors; for example, Qwen2.5-VL obtains 59.3% text-only accuracy on AI2D compared to 26.6% on MMStar. This provides a useful control: the ordering gap is largest on visually demanding benchmarks, but remains present even when language priors are stronger (Appendix A).

Model	Dataset	Image-first baseline	Question-first baseline	Question-first TTT	Δ_{QF}	Image-first TTT	Δ_{IF}
InternVL2-8B	**MMStar**	61.6	35.5	61.5	+26.1	62.3	**+0.7**
	RealWorldQA	61.8	38.2	60.0	+21.8	62.1	**+0.3**
	AI2D	83.6	64.5	83.0	+18.5	83.6	+0.0
Qwen2.5-VL-7B	**MMStar**	62.1	45.3	58.7	+13.4	62.7	**+0.5**
	RealWorldQA	67.8	60.0	64.1	+4.1	68.4	**+0.5**
	AI2D	84.5	69.3	82.3	+12.9	84.5	+0.0
Qwen3-VL-8B	**MMStar**	63.9	44.9	64.4	+19.5	65.1	**+1.1**
	RealWorldQA	71.6	65.5	70.7	+5.2	72.8	**+1.2**
	AI2D†	88.0	78.5	87.0	+8.5	88.5	**+0.5**

Table 1: **Main results.** Asymmetric order-consistency TTT improves question-first (QF) accuracy across all evaluated model–dataset pairs while maintaining or slightly improving image-first (IF) accuracy. Results are reported at a fixed model-specific step count: K=5 for InternVL2-8B, K=4 for Qwen2.5-VL-7B, and K=3 for Qwen3-VL-8B.

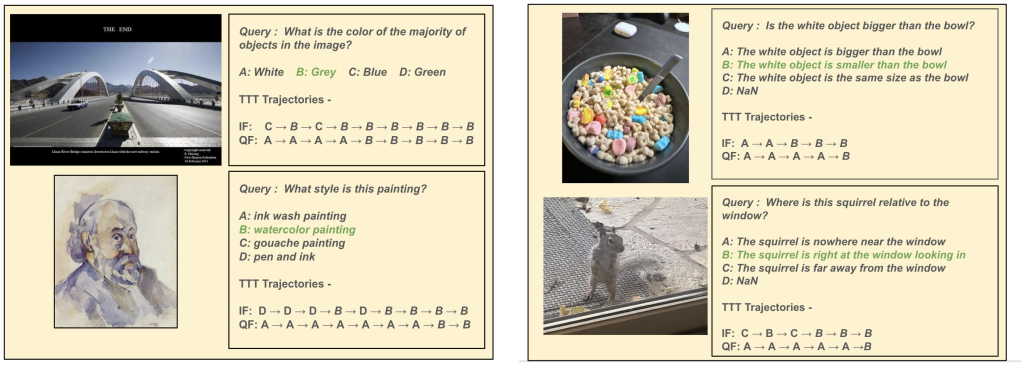


Figure 2: **Qualitative TTT trajectories.** Each panel shows image-first (IF) and question-first (QF) predictions across TTT steps; the correct answer is highlighted in green. In these examples, IF reaches the correct answer earlier, while QF catches up over the adaptation trajectory.

Prompt orders. For each instance, we evaluate two prompts that contain the same image, question, and answer candidates. The image-first prompt places the image before the question; the question-first prompt places the question and options before the image. Both prompts use the model's standard chat template, and only the relative position of the image token block is changed.

Test-time adaptation. For each test instance, we follow the protocol described in Section 3.2. Adaptation targets only the language model backbone parameters within the selected layer window; vision encoder parameters are held fixed throughout. We use SGD with momentum 0.9. We set the learning rate to $\eta = 5 \times 10^{-4}$ for InternVL2-8B and $\eta = 5 \times 10^{-5}$ for Qwen2.5-VL and Qwen3-VL. For each model, we select the learning rate, adaptation window, and step count once using a fixed, randomly sampled 100-question MMStar calibration subset. The resulting configuration is then frozen across all evaluation datasets. We use $K = 5$ for InternVL2-8B, $K = 4$ for Qwen2.5-VL-7B, and $K = 3$ for Qwen3-VL-8B. Thus, our deployment protocol uses labeled data for one-time model-level hyperparameter selection, while the per-instance adaptation objective uses no answer labels. Appendix K provides additional calibration results.

4.2 Main Results

TTT closes the question-first gap. Table 1 shows that asymmetric order-consistency TTT improves question-first accuracy in all nine model–dataset settings. The gains range from $+5.2$ to $+26.1$ percentage points. In several settings, the adapted question-first accuracy nearly reaches the image-first baseline: InternVL2 improves from 35.5% to 61.5% on MMStar, compared to an image-first

Benchmark	IF base	QF base	Δ_{IF}	Δ_{QF}
MMMU	48.7	41.1	**+0.8**	+6.9
SEEDBench	72.3	42.7	**+0.7**	+30.0
OCRBench	65.2	58.2	**+0.1**	+5.7

Table 2: **Additional benchmark results for InternVL2-8B.** Baseline scores and changes after TTT, reported on a 0–100 scale. The adaptation configuration is fixed across benchmarks.

baseline of 61.6%, and from 38.2% to 60.0% on RealWorldQA, exceeding the image-first baseline of 61.8%. These results show that a large fraction of the modality-ordering gap can be repaired at test time without labels.

Image-first accuracy is preserved or improves. Although the adaptation objective targets the weaker question-first ordering, image-first accuracy does not decrease in any setting. It improves in seven out of nine model–dataset pairs and remains unchanged in the other two. The gains are small but consistent, ranging from $+0.0$ to $+1.2$ percentage points, with the largest gains on Qwen3-VL RealWorldQA and MMStar. This indicates that aligning question-first predictions to the image-first teacher does not trade off against the stronger ordering and can slightly improve both prompt orders.

Generalization to additional benchmarks. We extend the evaluation to MMMU, SEEDBench, and OCRBench, covering knowledge-intensive reasoning, fine-grained perception, and OCR/document understanding, respectively. We use the same InternVL2-8B configuration throughout: $K = 5$, $\eta = 5 \times 10^{-4}$, and adaptation at layers 18–22, without additional benchmark-specific tuning. As shown in Table 2, TTT improves question-first performance across all three benchmarks, with gains ranging from 5.7 to 30.0 percentage points, while image-first performance changes by at most 0.8 points. For OCRBench, adaptation uses KL divergence between the first-token predictive distributions, extending the evaluation to OCR-oriented open-ended outputs.

Computational cost. On a single NVIDIA H100 80GB GPU, InternVL2-8B with $K = 5$ adaptation steps takes 5.63 seconds per instance, compared with 2.99 seconds for evaluating both prompt orders without adaptation ($1.88\times$ overhead). Peak memory is 36 GB when gradient computation is restricted to the selected layer window, compared with 76 GB before this restriction, without changing accuracy. TTT therefore provides improved order robustness at additional inference cost. Appendix L reports the dependence on the number of adaptation steps.

Teacher reliability. Image-first prompting is more accurate on average, but does not provide a correct teacher for every instance. On MMMU, we stratify 800 examples by the initial correctness of both prompt orders. TTT repairs 86 of 130 initially IF-correct/QF-wrong examples, but also corrupts 32 of 69 IF-wrong/QF-correct examples. Accounting for transitions in all four groups, QF accuracy increases from 329/800 to 384/800, a net gain of 6.9 percentage points. Thus, aggregate improvement coexists with individual regressions; IF preservation in aggregate does not imply that the adaptation is harmless on every example. Appendix I reports the complete breakdown.

4.3 Ablations

Loss function. Table 3 ablates the consistency objective used for test-time adaptation. Our default objective is the asymmetric KL loss defined in Section 3.2, which follows the teacher-student distillation view of matching a student distribution to a fixed teacher distribution (Hinton et al., 2015). We compare against three alternatives. Asymmetric CE keeps the same teacher-student direction but replaces KL with cross-entropy:

$$\mathcal{L}_{\text{asym-CE}} = -\sum_{v} \text{stopgrad}(p_{\text{IF}}(v)) \log p_{\text{QF}}(v).$$

Since the teacher distribution is fixed, this has the same gradients as asymmetric KL up to a constant.

Hard pseudo-labeling also keeps the same direction, but replaces the soft image-first distribution with its argmax prediction (Lee, 2013):

$$\mathcal{L}_{\text{hard}} = -\log p_{\text{QF}}(\hat{y}_{\text{IF}}), \qquad \hat{y}_{\text{IF}} = \arg\max_{v} p_{\text{IF}}(v).$$

Objective	IF base	Image-first TTT	Δ_{IF}	QF base	Question-first TTT	Δ_{QF}
Asym. KL (ours)	61.6	**62.3**	**+0.7**	35.5	**61.5**	**+26.0**
Asym. CE	61.6	62.1	+0.5	35.5	**61.5**	**+26.0**
Hard pseudo-label	61.6	61.6	+0.0	35.5	60.9	+25.4
Sym. CE	61.6	61.3	-0.3	35.5	45.3	+9.8

Table 3: **Loss-function ablation (InternVL2-8B/MMStar).** Asymmetric objectives improve question-first accuracy more than symmetric CE. Symmetric CE also slightly reduces image-first accuracy, consistent with the risk of treating the weaker ordering as an equally reliable training signal.

Window	IF base	Image-first TTT	Δ_{IF}	QF base	Question-first TTT	Δ_{QF}
Layers 5–9	61.6	61.6	+0.0	35.5	60.7	+25.2
Layers 8–12	61.6	61.6	+0.0	35.5	60.9	+25.4
Layers 13–17	61.6	62.1	+0.5	35.5	61.0	+25.5
Layers 16–20	61.6	61.9	+0.3	35.5	**61.7**	**+26.2**
Layers 18–22	61.6	**62.3**	**+0.7**	35.5	61.5	+26.0
Layers 20–24	61.6	61.9	+0.3	35.5	60.7	+25.2

Table 4: **Layer-window ablation on InternVL2-8B/MMStar.** All windows recover more than 25 points of question-first accuracy. Layers 16–20 gives the largest question-first gain, while layers 18–22 gives the largest image-first gain and is used as the default adaptation window.

This discards uncertainty in the image-first distribution.

Symmetric CE follows prior test-time consistency objectives that align predictions across semantically equivalent variants (Chou et al., 2026b):

$$\mathcal{L}_{\text{sym-CE}} = \frac{1}{2}\left[-\sum_{v} p_{\mathrm{IF}}(v)\log p_{\mathrm{QF}}(v) - \sum_{v} p_{\mathrm{QF}}(v)\log p_{\mathrm{IF}}(v)\right].$$

Unlike the asymmetric objectives, this treats image-first and question-first as equally reliable signals.

The results support the directional framing in Section 3.2: question-first should be adapted toward image-first, rather than forcing both prompt orders toward a shared midpoint.

Layer window. Table 4 compares adaptation across different layers. Mid-layers produce the strongest question-first recovery, while early and late windows are less effective.

Moving versus frozen teacher. Our default method recomputes the image-first teacher after each TTT update, while stopping gradients through the teacher branch at each step. We compare this to a frozen-teacher variant, where the image-first pseudo-label is computed once from the pretrained model and kept fixed throughout adaptation. Table 5 shows that recomputing the teacher is critical. With a frozen teacher, question-first accuracy improves from 35.5% to only 43.9%. With the moving teacher, question-first accuracy reaches 61.5% at the same step count. The moving teacher also improves image-first accuracy by $+0.7$ points, while the frozen teacher slightly decreases it. This supports the bootstrapping interpretation: as the shared weights improve question-first predictions, the recomputed image-first teacher also becomes sharper, providing a stronger target for later updates.

5 Mechanistic Analysis

Section 4 shows that order-consistency TTT reduces the question-first gap and slightly improves the image-first baseline too. We now ask where this degradation of question-first appears inside the model, and which internal components change after adaptation. The analysis supports two claims. First, activation patching localizes the ordering failure to a narrow mid-layer band. Second, TTT primarily reduces the image-first/question-first mismatch in late-layer MLP sublayers. We focus

Teacher	IF base	IF TTT	Δ_{IF}	QF base	QF TTT	Δ_{QF}
Moving (ours)	61.6	**62.3**	**+0.7**	35.5	**61.5**	**+26.0**
Frozen	61.6	61.5	-0.1	35.5	43.9	+8.5

Table 5: **Moving versus frozen teacher on InternVL2-8B/MMStar.** Both variants use $K{=}5$ TTT steps and the same mid-layer adaptation window. Recomputing the image-first teacher after each update substantially improves question-first recovery and also preserves image-first accuracy.

on InternVL2-8B in the main text because it exhibits the largest ordering gap in Table 1; additional Qwen3-VL activation-patching results show the same qualitative localization pattern and are reported in Section C, providing evidence that the mechanism generalizes across distinct model families.

5.1 Activation Patching Localizes the Ordering Failure

We use activation patching, following causal mediation and causal tracing methods for localizing transformer computations (Vig et al., 2020; Meng et al., 2022), to test where image-first computation can repair question-first computation. For each example, we run the model twice: once with image-first prompting and once with question-first prompting. The image-first run is the donor pass, and the question-first run is the recipient pass. At a chosen transformer layer L, we replace the final-token hidden state in the recipient pass with the corresponding hidden state from the donor pass:

$$h_{\text{QF}}^{(L)} \leftarrow h_{\text{IF}}^{(L)}.$$

The remaining layers are then run normally, and we read out the final answer.

We report two quantities. First, we measure patched question-first accuracy over all examples. Second, we measure recovery on the fixable subset,

$$\mathcal{F} = \{s : \hat{y}_{\text{IF}}(s) = y(s),\ \hat{y}_{\text{QF}}(s) \neq y(s)\},$$

where image-first is correct and question-first is wrong. Recovery at layer L is

$$\text{Recovery}(L) = \frac{1}{|\mathcal{F}|} \sum_{s \in \mathcal{F}} \mathbf{1}\left[\hat{y}_{\text{patch}}^{(L)}(s) = y(s)\right].$$

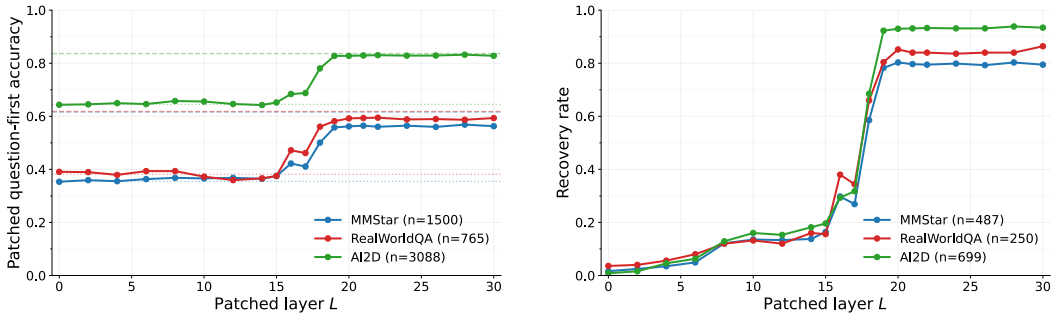
This subset isolates examples where the ordering gap produces an error that the image-first hidden state could in principle repair. In our experiments, the fixable subset contains 487 examples for MMStar, 250 for RealWorldQA, and 700 for AI2D.

If the two prompt orders diverged uniformly across depth, patching image-first activations into question-first computation would be expected to help across many layers. Figure 3 shows a different pattern. Panel (a) reports patched question-first accuracy over all examples, while panel (b) reports recovery on the fixable subset. Patching early layers has little effect: recovery remains low before layer 17 on all datasets. Around layers 18–20, recovery rises sharply, and after this band the curves plateau. Thus, the ordering failure is not spread uniformly across depth; it is causally concentrated in a narrow mid-layer stage of computation.

Post-adaptation component alignment. Complementary analysis of late-layer components shows that TTT increases overlap between the MLP sublayers selected under the two prompt orders: normalized intersection-over-union rises from -0.82 to $+0.84$ for the default adaptation window, while attention-head overlap changes only modestly. This supports increased cross-order alignment after adaptation, but does not establish MLP sublayers as the causal source of the ordering failure. Appendix H provides the metric definition and full results; Appendix D compares the internal changes induced by different adaptation windows.

Because raw IoU depends on the number of selected components, we normalize it against a size-matched random baseline:

$$\text{NIoU}^c = \frac{\text{IoU}^c - \mathbb{E}[\text{IoU}_{\text{rand}}^c]}{1 - \mathbb{E}[\text{IoU}_{\text{rand}}^c]}.$$



(a) **Patched accuracy on all examples.** (b) **Recovery on IF-correct / QF-wrong examples.**

Figure 3: **Activation patching localizes the ordering failure.** We replace the question-first final-token hidden state at layer L with the corresponding image-first hidden state and continue the forward pass normally. Patching becomes effective only in a narrow mid-layer band, where patched question-first accuracy increases and IF-correct/QF-wrong examples are recovered at high rate.

Under this normalization, $\mathrm{NIoU} = 1$ indicates perfect overlap, $\mathrm{NIoU} = 0$ indicates random-level overlap, and $\mathrm{NIoU} < 0$ indicates lower overlap than the random baseline.

Figure 7 shows that TTT substantially changes late-layer MLP alignment between the two prompt orders. Before TTT, late-layer MLP components used by image-first and question-first have a strongly below-random overlap, while attention-head overlap is weakly above random. After TTT, the late-layer MLP overlap becomes strongly positive across adaptation windows, whereas attention-head overlap changes only modestly. Thus, TTT also makes the late-layer MLP computation more similar across the two prompt orders, thus improving accuracy.

Together, Figs. 3 and 7 provide complementary evidence. Activation patching localizes where the ordering failure can be causally repaired, while late-layer NIoU shows that successful TTT makes the final MLP-mediated computation more order-consistent. This is also consistent with Table 4, where the strongest adaptation windows overlap the layer range where patching becomes effective. Additional drift analyses in Appendix D show that different adaptation windows can reach high output agreement through different internal update footprints.

6 Discussion, Limitations, and Future Work

The main implication of our findings is that prompt formatting can induce systematic computational differences, even when the task content is unchanged. This makes modality order a robustness axis for VLM evaluation and a useful source of self-supervision at test time. More broadly, our results suggest that test-time adaptation should not always enforce symmetric consistency: when one input variant is empirically more reliable, the objective should preserve that directionality. The mechanistic evidence further suggests that such failures may be repaired more effectively when adaptation targets the internal region where the inconsistency emerges, rather than updating the model uniformly.

Limitations. Our method increases inference cost because each test instance requires two prompt orders and per-instance gradient updates. Our main evaluation focuses on multiple-choice benchmarks, so broader evaluation on captioning, long-form reasoning, and free-form VQA is still needed. Our mechanistic analysis is strongest for InternVL2-8B; additional Qwen3-VL patching results show the same qualitative pattern, but a full cross-family circuit analysis remains future work.

Future work. Our ablations and results suggest that image-first gains arise because the moving-teacher update can sharpen shared computation across both orderings. A natural next step is to make this causal account more precise. More broadly, semantically irrelevant ordering changes may expose useful self-supervised signals for improving robustness and accuracy at test time.

AI USE STATEMENT

In this work, we conceived and carried out the original research questions, core ideas, proposed methodology, theoretical contributions, experimental design, and overall execution; generative AI tools were used only in an assistive capacity.

We used generative AI tools to assist with editing and polishing the text, checking mathematical arguments and derivations, and assisting with the implementation and verification of computational methods. We additionally used generative AI tools for improving clarity and presentation, brainstorming and refining research ideas, searching for and identifying potentially relevant literature, assisting with coding and research workflow, and reviewing the manuscript for consistency, clarity, and completeness.

We have reviewed all AI-assisted work. We have independently checked all mathematical arguments, derivations, and claims. We have also verified suggested references against the original sources. The code was inspected and tested, and all reported experimental results were verified. All substantive research decisions, including the formulation of the method, theoretical claims, experimental protocols, interpretation of results, and conclusions, were made by us.

We take full responsibility for the final content of this work, including text, mathematical claims, experimental results, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work advances fundamental methodology for generative modeling and does not involve human subjects, sensitive personal data, or deployment in high-stakes applications. We do not identify any specific societal or ethical concerns arising from the contributions of this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. For the theoretical results, the Appendix provides the full mathematical settings, assumptions, statements, and proofs. For the empirical results, the Appendix provides detailed descriptions of the datasets and benchmarks, experimental protocols, model and training configurations, evaluation procedures, and relevant hyperparameters. Upon acceptance, we will release the source code required to reproduce our experiments in a public GitHub repository and provide the corresponding repository link in the paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024b.

Shih-Han Chou, Shivam Chandhok, James J Little, and Leonid Sigal. Test-time consistency in vision language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7789–7798, 2026a.

Shih-Han Chou, Shivam Chandhok, James J. Little, and Leonid Sigal. Test-time consistency in vision language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876, 2025.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SHMi1b7sjXk.

Parsa Hejabi, Elnaz Rahmati, Alireza S Ziabari, and Morteza Dehghani. Flip-flop consistency: Unsupervised training for robustness to prompt perturbations in llms. *arXiv preprint arXiv:2510.14242*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

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*, pages 216–233. Springer, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*, 2025a.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. In *Advances in Neural Information Processing Systems*, 2025b.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. In *Advances in Neural Information Processing Systems*, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.