VEC-TTA: VISUAL-EVIDENCE-CALIBRATED TEST-TIME ADAPTATION FOR COMPACT MULTIMODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time computation provides an opportunity to improve multimodal large language models (MLLMs) beyond their pretrained capabilities. However, reinforcing model predictions does not ensure that the resulting answers are driven by query-relevant visual evidence. Models can favor answers elicited without visual evidence even when the original image is present, and retain original answers after critical visual evidence is altered. To address these limitations, we propose **Visual Evidence Calibrated Test-Time Adaptation (VEC-TTA)**, an instance-specific adaptation framework for compact MLLMs that performs temporary low-rank updates while keeping the backbone frozen. Specifically, VEC-TTA comprises three complementary modules: (1) Query-Conditioned Evidence Elicitation (QEE), which constructs pseudo-supervision from the current image–query instance; (2) Invariant Evidence Consolidation (IEC), which supplements answer fitting with visual consistency, jointly fitting pseudo-targets and aligning predictions across online content-preserving views; and (3) Counterfactual Evidence Calibration (CEC),which constrains persistent answer support by encouraging stronger support for the same pseudo-target when visual content is present than when it is removed. Experiments on two compact MLLMs across multiple benchmarks show consistent improvements over representative test-time methods. Our code is available at `anonymous.4open.science/r/VEC-TTA-E0C3`.

1 INTRODUCTION

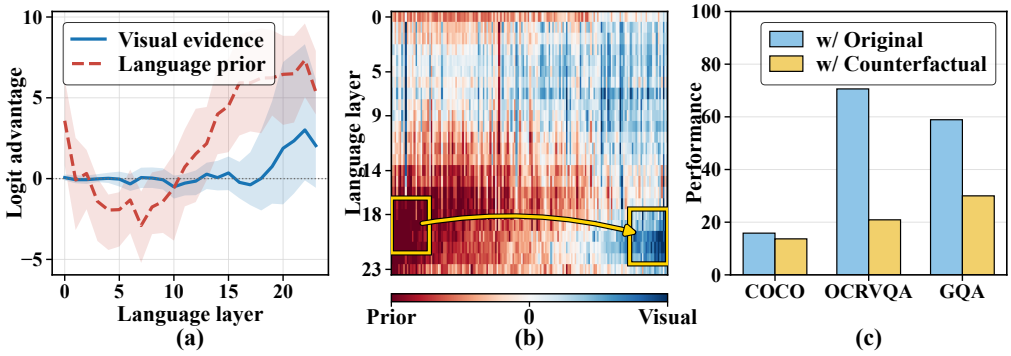


Figure 1: Diagnostics motivating VEC-TTA: (a) Layer-wise visual-evidence and language-prior diagnostics. (b) Per-instance logit margins across language layers. (c) Qwen3-VL-2B performance under original and annotated counterfactual images.

Multimodal large language models (MLLMs) Marafioti et al. (2025); Yang et al. (2025); Lu et al. (2025) support a range of image-conditioned tasks, including visual question answering and image captioning. Beyond their pretrained capabilities, additional computation at test time provides an opportunity to improve their predictions. Test-time scaling (TTS) Wang et al. (2025); Hu et al. (2026) allocates this computation to exploring and aggregating predictions from a frozen model, whereas test-time adaptation (TTA) Lafon et al. (2026); Dastmalchi et al.; Kaya et al. (2026) temporarily

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.

updates model parameters using inference-time signals to enable instance-specific correction. However, accuracy improvements alone do not characterize whether the prediction is grounded in the given image Sun et al. (2025). A generative MLLM may produce the correct answer without relying sufficiently on the relevant visual content Liu et al. (2025b); Shi et al. (2026).

In this work, we focus on compact MLLMs and examine how visual evidence and language priors contribute to their predictions, using the diagnostics in Figure 1. Specifically, Figure 1(a) shows that visual evidence gains influence in later layers, but the language-prior advantage grows more rapidly and remains stronger near the output. This pattern suggests that the presence of visual information does not necessarily translate into a preference for the reference answer near the output. Figure 1(b) further reveals substantial variation across samples. For some instances, the answer elicited under the null-visual condition remains favored in later layers even when the original image is present, whereas others establish a clear preference for the reference answer. These cases suggest that the presence of visual input alone may be insufficient to overcome answer preferences that persist without visual evidence. At the behavioral level, Figure 1(c) shows that the model retains substantial performance against the original references even after counterfactual images alter critical visual evidence. This persistence suggests that predictions can remain anchored to answers that are no longer supported by the modified evidence.

Together, these observations reveal two related limitations in how model predictions reflect visual evidence: (1) even with the original image present, the model can favor an answer elicited without visual evidence over the reference answer; and (2) when critical visual evidence is altered, predictions can remain anchored to the original answer despite the change in its visual support.

Motivated by these observations, we propose a novel framework termed **Visual Evidence Calibrated Test-Time Adaptation (VEC-TTA)**. VEC-TTA adapts a compact MLLM to each image–query pair through temporary low-rank updates while keeping the backbone parameters frozen. As illustrated in Figure 2, VEC-TTA comprises three complementary modules: (1) *Query-Conditioned Evidence Elicitation (QEE)*, which constructs query-conditioned pseudo-supervision from auxiliary questions and visual variations; (2) *Invariant Evidence Consolidation (IEC)*, which supplements answer agreement with visual consistency, jointly fitting pseudo-targets and aligning predictions across content-preserving views; (3) *Counterfactual Evidence Calibration (CEC)*, which constrains persistent answer support when visual content is absent. It encourages stronger support for the same pseudo-target under its associated image than under an all-black counterpart. Specifically, IEC and CEC extend pseudo-answer fitting with constraints on prediction consistency when visual content is preserved and relative answer support when it is removed.

Our contributions are summarized as follows:

- We propose VEC-TTA, an instance-specific adaptation framework that complements query-conditioned pseudo-answer fitting with explicit constraints across visual conditions.
- We develop IEC for joint pseudo-target fitting and cross-view consistency, and CEC for relative-support calibration between associated images and all-black counterparts.
- Experiments on two compact MLLMs show task performance gains and complementary benefits from IEC and CEC, with matched visual interventions indicating increased selective sensitivity to answer-relevant regions.

2 Related Work

2.1 Visual Evidence and Language Priors

Prior work examines how MLLMs use visual evidence and how language priors affect their responses. Diagnostic studies identify failures despite recognizing relevant content Liu et al. (2025a), limitations in fine-grained visual grounding Li et al. (2026), and prompt-induced hallucinations under conflicting textual and visual information Rudman et al. (2026). Related analyses show that extended reasoning can weaken visual grounding Xu et al. (2026), while reinforcement post-training can improve reasoning performance even when essential input information is corrupted Zhang et al. (2026). To mitigate these issues, decoding methods contrast predictions under original and distorted images Leng et al. (2024) or strengthen image–text dependency through mutual-information-based

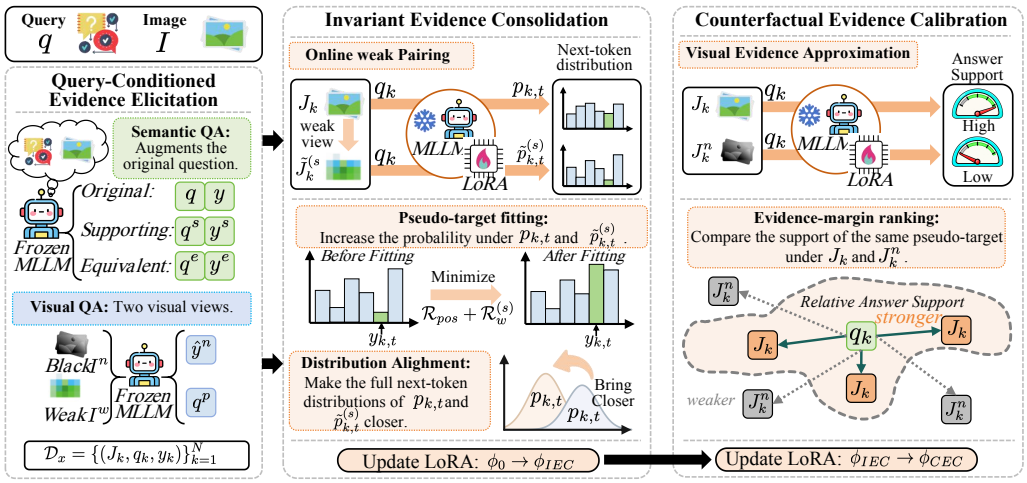


Figure 2: Overview of the VEC-TTA: (1) QEE constructs query-conditioned pseudo-supervision. (2) IEC combines pseudo-target fitting with cross-view consistency. (3) CEC further calibrates relative support for each pseudo-target between its associated image and an all-black counterpart.

calibration Fang et al. (2026). Other inference-time approaches intervene in attention heads or internal representations to reduce hallucinations Ji et al. (2026); Liu et al. (2026). Training-based approaches include visual and textual grounding in PostAlign Wu et al. (2026), modality-aware invariance and sensitivity constraints in MoD-DPO Chaubey et al. (2026).

2.2 Test-Time Scaling and Adaptation

Test-time computation can improve predictions through frozen-model inference or adaptation using inference-time signals. Test-time scaling explores sampled responses through self-consistency Wang et al. (2023b), structured reasoning Xu et al. (2025), and reward-guided decoding Mañas et al. (2025). Multimodal approaches further extend this paradigm through iterative perception Yan et al. (2026), multi-round reasoning and refinement Chen et al. (2026), and self-verification in diffusion-based multimodal models Xin et al. (2026). For compact VLMs, TTAug aggregates token-level predictions across augmented inputs without parameter updates, whereas TTAdapt uses the resulting consensus-based pseudo-labels for parameter adaptation Kaya et al. (2026). Adaptation methods for contrastive VLMs additionally explore statistical anchoring Zanella et al. (2025), Bayesian adaptation Zhou et al. (2025), retrieval Fan et al. (2025), and statistics caching Guan et al. (2026). Evaluations under realistic and unified protocols highlight sensitivity to test-distribution assumptions and trade-offs between accuracy and reliability Zanella et al. (2025); Sheng et al. (2026). For generative language models, QueST performs query-conditioned self-training by fitting generated auxiliary problem–solution pairs with a cross-entropy objective Song et al. (2026).

3 Proposed Method

3.1 Method Overview

Given a test instance $x = (I, q)$, our goal is to answer query q using the visual content of image I. VEC-TTA addresses this through three sequential stages. Specifically, QEE constructs query-conditioned pseudo-supervision, IEC jointly optimizes pseudo-target fitting and prediction consistency across content-preserving visual views, and CEC further calibrates relative answer support by comparing the support for the same answer under each image and its all-black counterpart. Throughout adaptation, the base parameters θ remain frozen and only the LoRA parameters ϕ are updated. For each instance, ϕ starts from ϕ_0 and is reset after the final prediction, making VEC-TTA episodic.

3.2 Query-Conditioned Evidence Elicitation

QEE uses the current MLLM to construct an instance-specific set $\mathcal{D}_x$ from (I, q), providing query-conditioned pseudo-supervision shared by the subsequent IEC and CEC updates.

Semantic QA Construction. A single question provides limited supervision about the visual evidence required by the current instance. QEE therefore augments the original question with supporting and equivalent questions sampled from the current MLLM conditioned on (I, q). The supporting question q^{s} asks for a simpler, directly observable fact that can inform the original answer. The equivalent question q^{e} varies the wording while retaining the required visual evidence and answer type. For each question $q_k \in \{q, q^{\mathrm{s}}, q^{\mathrm{e}}\}$, the corresponding answer is generated independently as

$$y_k = \mathrm{Decode}\left(p_\theta\left(\cdot \mid I, q_k\right)\right), \tag{1}$$

where p_θ indicates the frozen base model and Decode indicates decoding operation. This produces the original, supporting, and equivalent triples (I, q, y), $(I, q^{\mathrm{s}}, y^{\mathrm{s}})$, and $(I, q^{\mathrm{e}}, y^{\mathrm{e}})$.

Visual QA Construction. QEE further augments the $\mathcal{D}_x$ with two visual views of the original image. The weak view provides semantics-preserving augmentation, while the black-image view provides a reference with the visual content removed. A weak transformation applies random photometric perturbations, such as brightness and contrast changes, while preserving task-relevant content, producing a base weak image I^{w}. Then QEE reuses y to form (I^{w}, q, y). For the black-image view, QEE replaces the original image with a black image I^{n} and decodes

$$\hat{y}^{\mathrm{n}} = \mathrm{Decode}\left(p_\theta\left(\cdot \mid I^{\mathrm{n}}, q\right)\right). \tag{2}$$

We use $\hat{y}^{\mathrm{n}}$ to construct a prior-aware version q^{p} of the original question. QEE then restores the original image and reuses y to form (I, q^{p}, y). Thus, the answer generated from the black image is used only for question construction and never as a pseudo-target.

$$\mathcal{D}_x = \{(J_k, q_k, y_k)\}_{k=1}^{N}, \tag{3}$$

where J_k denotes the visual input associated with each entry. The same dataset $\mathcal{D}_x$ is shared by IEC and CEC without regenerating its questions or targets.

3.3 Invariant Evidence Consolidation

Given $\mathcal{D}_x$, IEC supplements pseudo-target fitting with an explicit cross-view consistency constraint, jointly optimizing answer likelihood and prediction agreement across paired visual inputs.

Online Weak Pairing. The weak image constructed by QEE is a fixed entry in $\mathcal{D}_x$, whereas IEC additionally samples new weak views for each entry in $\mathcal{D}_x$ during adaptation. At each IEC update s, we sample a new weak perturbation $\epsilon_k^{(s)}$ for every $(J_k, q_k, y_k) \in \mathcal{D}_x$ and construct

$$\widetilde{J}_k^{(s)} = \mathcal{T}_{\mathrm{w}}\left(J_k; \epsilon_k^{(s)}\right), \ \widetilde{\mathcal{D}}_x^{(s)} = \left\{(\widetilde{J}_k^{(s)}, q_k, y_k)\right\}_{k=1}^{N}. \tag{4}$$

These online views are resampled at every update and are used only during IEC. Since $\mathcal{T}_{\mathrm{w}}$ preserves task-relevant visual content, each pair shares the same query and pseudo-target.

Invariant Evidence Fitting. Given the paired views $(J_k, \widetilde{J}_k^{(s)})$, IEC fits the same pseudo-target under both views, which are assumed to preserve the task-relevant content. For a visual input J and a query–answer pair (q_k, y_k), the answer-only teacher-forced cross-entropy is

$$\mathcal{L}_{\mathrm{CE}}(J, q_k, y_k) = -\frac{1}{|y_k|}\sum_{t=1}^{|y_k|}\log p_{\theta,\phi}\left(y_{k,t} \mid J, q_k, y_{k,<t}\right). \tag{5}$$

Prompt and padding positions are masked so that supervision is applied only to the pseudo-target answer tokens. IEC applies this objective to both the bank entries and their online weak views:

$$\mathcal{R}_{\mathrm{pos}} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{L}_{\mathrm{CE}}(J_k, q_k, y_k), \ \mathcal{R}_{\mathrm{w}}^{(s)} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{L}_{\mathrm{CE}}(\widetilde{J}_k^{(s)}, q_k, y_k). \tag{6}$$

Beyond fitting the same answer, IEC encourages the two views to produce consistent next-token distributions. For the same query q_k and teacher-forced prefix $y_{k,<t}$, we denote the full-vocabulary next-token distributions under J_k and $\widetilde{J}_k^{(s)}$ by $p_{k,t}$ and $\widetilde{p}_{k,t}^{(s)}$, respectively. We align them using stop-gradient symmetric KL divergence:

$$D_{\mathrm{SKL}}^{(s)} = \frac{1}{2N}\sum_{k=1}^{N}\frac{1}{|y_k|}\sum_{t=1}^{|y_k|}\left[D_{\mathrm{KL}}\left(\mathrm{sg}[p_{k,t}] \,\|\, \widetilde{p}_{k,t}^{(s)}\right) + D_{\mathrm{KL}}\left(\mathrm{sg}[\widetilde{p}_{k,t}^{(s)}] \,\|\, p_{k,t}\right)\right]. \tag{7}$$

Here, sg stops gradients through the reference distribution in each direction. With θ frozen, IEC updates the LoRA parameters from ϕ_0 to ϕIEC with the following objective:

$$\mathcal{L}_{\mathrm{IEC}}^{(s)} = \mathcal{R}_{\mathrm{pos}} + \lambda_{\mathrm{w}} \mathcal{R}_{\mathrm{w}}^{(s)} + \lambda_{\mathrm{KL}} D_{\mathrm{SKL}}^{(s)}. \tag{8}$$

3.4 Counterfactual Evidence Calibration

Given $\mathcal{D}_x$, CEC calibrates relative answer support by comparing the support for each pseudo-target under its associated image and an all-black counterpart.

Visual Evidence Approximation. To constrain answer support in the absence of visual evidence, CEC replaces each associated image with an all-black counterpart. This provides an annotation-free reference for comparing support for the same pseudo-target with and without visual input content. We later analyze the resulting approximation theoretically and compare it empirically with local evidence removal in Appendix D.1. For each $(J_k, q_k, y_k) \in \mathcal{D}_x$, CEC constructs $J_k^{\mathrm{n}} = \mathcal{T}_{\mathrm{n}}(J_k)$, where $\mathcal{T}_{\mathrm{n}}$ replaces J_k with an all-black image of the same size. The two conditions keep q_k and y_k fixed and differ only in the visual input. Thus, J_k^{n} serves as an annotation-free approximation.

Evidence-margin Ranking. CEC keeps q_k and y_k fixed and defines the evidence margin through their length-normalized cross-entropy:

$$\Delta_k = \mathcal{L}_{\mathrm{CE}}(J_k^{\mathrm{n}}, q_k, y_k) - \mathcal{L}_{\mathrm{CE}}(J_k, q_k, y_k),\ \mathcal{L}_{\mathrm{CEC}} = \frac{1}{N} \sum_{k=1}^{N} \mathrm{softplus}(-\Delta_k). \tag{9}$$

Minimizing $\mathcal{L}_{\mathrm{CEC}}$ favors $\Delta_k > 0$, giving the same pseudo-target lower loss under J_k than under J_k^{n}. The black-image condition introduces no additional target and is used only to compare the support for the same pseudo-target. Starting from ϕ^{IEC}, CEC minimizes $\mathcal{L}_{\mathrm{CEC}}$ to obtain ϕ^{CEC}. The resulting adapter is used for the final prediction and reset before processing the next instance.

4 Experiments

4.1 Experimental Setups

Evaluation Setup. We evaluate VEC-TTA on two MLLMs, SmolVLM2-2.2B-Instruct Marafioti et al. (2025) and Qwen3-VL-2B-Instruct Yang et al. (2025), which we refer to as SmolVLM-2B and Qwen-2B, respectively. We conduct experiments on five benchmarks, including COCO Chen et al. (2015), AMBER Wang et al. (2023a), GQA Hudson & Manning (2019), VizWiz Gurari et al. (2018), and HalluBench Guan et al. (2024). We report ROUGE-L on COCO and accuracy on the remaining four benchmarks. We compare VEC-TTA with representative test-time computation methods from two categories: (1) test-time scaling methods that keep model parameters frozen, including Self-Consistency Wang et al. (2023b), which aggregates multiple predictions through majority voting, and SCI Tang et al. (2026); (2) test-time adaptation methods that update model parameters, including TTAdapt Kaya et al. (2026) and QueST Song et al. (2026). Vanilla denotes direct inference without additional test-time computation. Due to space constraints, detailed definitions of the evaluation metrics used throughout the experiments are provided in Appendix B.

Implementation Details. We use LoRA with rank $r = 4$ and scaling factor $\alpha = 8$ for test-time adaptation. For each test instance, we perform 5 update steps in each of IEC and CEC, using AdamW with a learning rate of 1×10^{-4}. For IEC, we set $\lambda_{\mathrm{w}} = 0.5$ and $\lambda_{\mathrm{KL}} = 0.1$. All experiments are conducted on 8 NVIDIA A6000 GPUs.

4.2 Overall Comparisons

Main Results. Table 1 summarizes the overall performance of VEC-TTA. The gains are particularly pronounced on SmolVLM-2B, where VEC-TTA reaches 45.3 % ROUGE-L on COCO, outperforming the strongest methods by 29.4 %. VEC-TTA also achieves strong performance on Qwen-2B, showing consistent improvements across the evaluated benchmarks. Overall, these results demonstrate that VEC-TTA generalizes across different compact MLLM backbones and image-conditioned tasks, supporting the effectiveness of the proposed test-time adaptation framework.

Table 1: Comparison with representative test-time computation methods on two compact MLLMs.

Backbone	Method	Benchmark					Avg.
		COCO	AMBER	GQA	VizWiz	HalluBench	
SmolVLM-2B	Vanilla	8.8	63.9	55.7	8.9	57.0	38.9
	Self-Consistency	8.1	68.9	54.9	11.7	58.5	40.4
	SCI	9.3	74.6	52.5	19.9	53.2	41.9
	TTAdapt	15.9	75.4	53.9	10.2	59.4	43.0
	Quest	11.5	75.2	53.9	26.8	59.9	45.5
	VEC-TTA	**45.3**	**75.7**	**57.7**	**33.0**	**62.4**	**54.9**
Qwen-2B	Vanilla	15.8	62.9	58.9	25.0	64.1	45.3
	Self-Consistency	15.3	65.6	58.2	25.9	64.4	45.9
	SCI	16.7	70.3	**59.5**	25.5	62.7	46.9
	TTAdapt	15.9	61.6	59.1	24.5	64.5	45.1
	Quest	24.1	76.7	58.5	**26.2**	61.4	49.38
	VEC-TTA	**30.8**	**77.4**	**59.5**	**26.2**	**66.5**	**52.1**

4.3 Further Analysis

Ablation Study. Table 2 examines the contributions of IEC and CEC. We keep QEE fixed because it provides the pseudo-supervision shared by both stages, and removing it would change their common supervision source rather than isolate either optimization component. (1) Comparing Experiments 0 and 1 shows that IEC improves performance across all five benchmarks, supporting the effectiveness of combining pseudo-target fitting with cross-view consistency. (2) Comparing Experiments 0 and 2 shows that CEC also improves performance without the IEC stage, supporting the utility of relative answer-support calibration between associated images and all-black counterparts. (2) Experiment 3 improves upon the IEC-only setting in Experiment 1. IEC fits pseudo-targets under semantics-preserving visual variations, while CEC further enlarges their support margin against the all-black counterpart, making them complementary.

Hyperparameter Analysis. Figure 3 examines 3 aspects of the adaptation configuration. (a) Increasing the pseudo-supervision budget improves performance across five benchmarks, with limited gains beyond four to five entries. Our default $N = 5$ covers all five QEE entry types, while expanding to ten entries yields diminishing returns. (b) With $\alpha/r = 2$, performance varies across nonzero LoRA ranks, and larger ranks do not consistently improve results. This suggests that the adaptation benefits do not require a large low-rank update capacity. (c) The two IEC loss weights affect performance: increasing λ_{w} from the lower end of the tested range improves scores, while increasing λ_{KL} produces a non-monotonic response. The stronger results at intermediate KL weights suggest that cross-view consistency benefits from a balance with pseudo-target fitting.

Sequential vs. Joint Optimization. In this section, we aim to answer a natural question: *Why does VEC-TTA adopt sequential IEC–CEC optimization instead of jointly optimizing the two objectives from initialization?* We therefore compare the sequential IEC–CEC schedule with Joint, which optimizes both objectives simultaneously. As shown in Figure 4, sequential optimization achieves better overall performance on both evaluated backbones. This result suggests that first fitting the pseudo-supervision under weak-view consistency and then calibrating its relative answer support is more effective than imposing both objectives from the beginning.

Sequential vs. Joint LoRA Updates. To further characterize the difference between the two optimization schedules, Figure 5 compares their final effective LoRA updates on SmolVLM-2B. Figure 5(a) shows that the Sequential and Joint updates remain positively aligned overall, but exhibit noticeable directional differences across instances and benchmarks. In contrast, Figure 5(b) shows closely matched cumulative spectral-energy profiles in both visual and language modules. Together with the performance results above, these observations suggest that the advantage of sequential optimization is associated with a distinct update direction rather than a substantially different spectral concentration. This further supports fitting the pseudo-supervision before applying relative visual-support calibration, rather than optimizing both objectives jointly from initialization.

Analysis of Stage-wise LoRA Updates. We analyze the stage-specific LoRA updates of IEC and CEC to characterize how the two stages modify the adapter. We compare their update magnitude,

Table 2: Ablation study of our VEC-TTA on SmolVLM-2B.

No.	Components		Benchmarks				
	IEC	CEC	COCO	AMBER	GQA	VizWiz	HalluBench
0	–	–	8.8	63.9	55.7	8.9	57.0
1	✓	–	44.2	73.2	56.6	32.8	61.3
2	–	✓	26.1	72.9	55.8	26.3	60.6
3	✓	✓	**45.3**	**75.7**	**57.7**	**33.0**	**62.4**

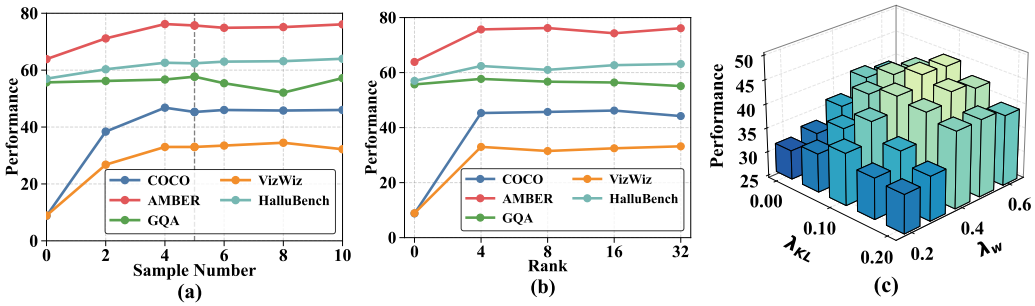


Figure 3: Hyperparameter analysis of (a) the number of pseudo-supervision samples, (b) LoRA rank, and (c) IEC loss weights.

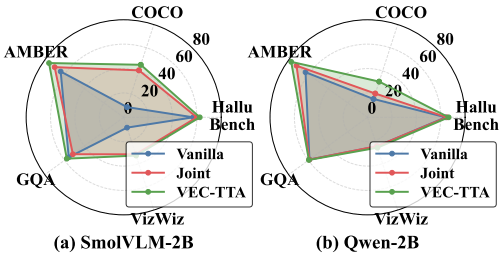


Figure 4: Performance comparison between sequential and joint optimization on SmolVLM-2B and (b) Qwen-2B.

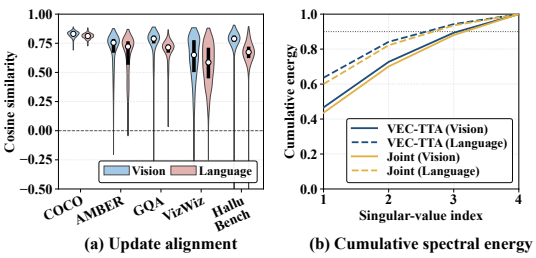


Figure 5: LoRA updates produced by sequential and joint optimization: (a) update-direction cosine similarity and (b) cumulative spectral-energy.

spectral-energy distribution, and stable rank Ipsen & Saibaba (2025). Figure 6(a) shows that IEC and CEC produce updates of comparable magnitude across datasets, indicating that CEC introduces a non-negligible correction after IEC. Despite their similar scale, the two stages exhibit distinct spectral structures. As shown in Figure 6(b), IEC distributes more update energy across the spectrum, whereas CEC concentrates more energy in the leading singular components. Consistently, Figure 6(c) shows that CEC exhibits lower stable rank Ipsen & Saibaba (2025) across both visual and language modules. These results show that IEC and CEC introduce comparable but structurally distinct updates, supporting their complementary roles in the two-stage adaptation process.

Analysis of Visual Evidence. We evaluate on the region-annotated training split of the Vision-OPD-6K Yuan et al. (2026), using reference answers only for scoring. To assess selective sensitivity to answer-relevant visual content, we compare the original image with matched masking of either the annotated evidence region or a same-sized, non-overlapping control region. As shown in Table 3, IEC-only improves original-image accuracy over Vanilla on both backbones and increases the evidence–control NLL gap (ES). These results suggest that pseudo-target fitting with weak-view consistency already contributes to predictions that are more selectively sensitive to answer-relevant visual content. Building on IEC, CEC improves original-image accuracy and increases ES, while accuracy under control masking remains close to that on the original images. Together, these comparisons support the complementary contributions of evidence consolidation and relative-support calibration to improving predictions and their selective sensitivity to visual evidence. We further examine the proxy gap (PG) to assess the discrepancy between black-image replacement and local

evidence masking. Although IEC-only improves accuracy and ES, it does not consistently reduce PG relative to Vanilla. Adding CEC reduces PG on both backbones, tightening the theoretical bound on the calibration-objective discrepancy. The detailed proof is provided in Appendix D.1.

Table 3: Visual intervention results on Vision-OPD-6K. ES denotes the evidence–control NLL gap, and PG denotes the black-image–evidence-masking NLL gap.

Backbone	Method	Accuracy (%)			NLL Diagnostics	
		Original	Evidence	Control	ES	PG
SmolVLM-2B	Vanilla	36.6	29.8	34.1	0.18	0.61
	TTAdapt	36.3	29.8	34.1	0.19	0.61
	IEC-only	37.6	27.1	36.5	0.25	0.80
	VEC-TTA	39.3	23.2	38.6	0.65	0.32
Qwen-2B	Vanilla	45.2	27.3	43.6	0.71	1.71
	TTAdapt	45.0	27.5	44.4	0.71	1.70
	IEC-only	45.4	27.9	44.8	0.87	1.46
	VEC-TTA	46.7	25.9	45.6	1.93	0.90

Table 4: Prediction transitions on SmolVLM-2B relative to the vanilla model and TTAdapt. "T" and "F" denote correct and incorrect predictions, respectively.

Dataset	Vanilla vs. VEC-TTA			TTAdapt vs. VEC-TTA		
	F→T	T→F	T→T	F→T	T→F	T→T
AMBER	17.7	5.9	58.0	6.3	6.0	69.4
GQA	6.4	4.4	51.3	5.1	1.3	52.6
VizWiz	25.7	1.6	7.3	22.9	0.1	10.1
HalluBench	13.8	8.4	48.6	12.5	9.5	49.9

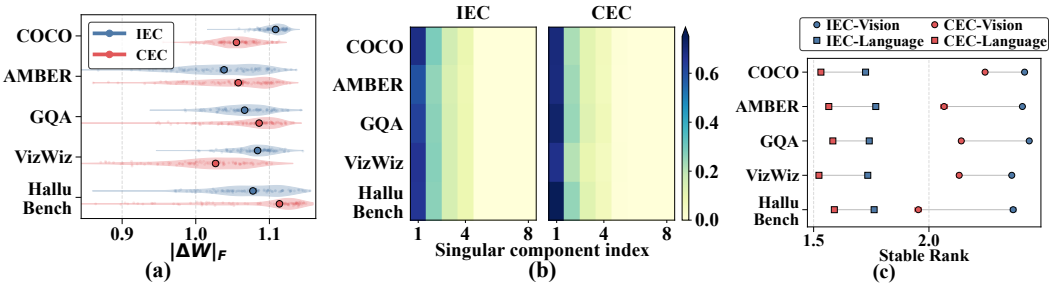


Figure 6: Stage-wise LoRA update geometry of IEC and CEC: (a) Frobenius-norm magnitude, (b) singular-value energy distribution, and (c) stable rank of visual and language modules.

Analysis on Prediction Transition. To better understand how VEC-TTA changes model predictions, we analyze instance-level correctness transitions relative to the vanilla model and TTAdapt. Table 4 reports results on four accuracy-based benchmarks; COCO is excluded because its primary metric is ROUGE-L. Across all benchmarks, F→T transitions consistently exceed T→F transitions, showing that the performance gains are driven primarily by correcting previous errors. This pattern is particularly pronounced on VizWiz, where VEC-TTA yields 25.7% F→T versus 1.6% T→F relative to the vanilla model, and 22.9% versus 0.1% relative to TTAdapt.

Analysis of Training Dynamics. Figure 7 examines the two-stage adaptation process on VizWiz. We extend each stage from the default 5 steps to 10 to track how the optimization metrics evolve over a longer adaptation horizon. During IEC, the pseudo-target cross-entropy decreases for both the entries in D_x and their online weak views in Figure 7(a), while the decreasing symmetric KL in Figure 7(b) indicates increasingly consistent full-vocabulary next-token distributions across paired views. During CEC, the black-image cross-entropy increases substantially while the cross-entropy under the associated image remains comparatively low. This widens the support margin primarily by reducing pseudo-target support in the absence of visual content, consistent with CEC's role in constraining persistent answer support under this condition. Additional training dynamics on the remaining benchmarks are provided in Appendix C.2.

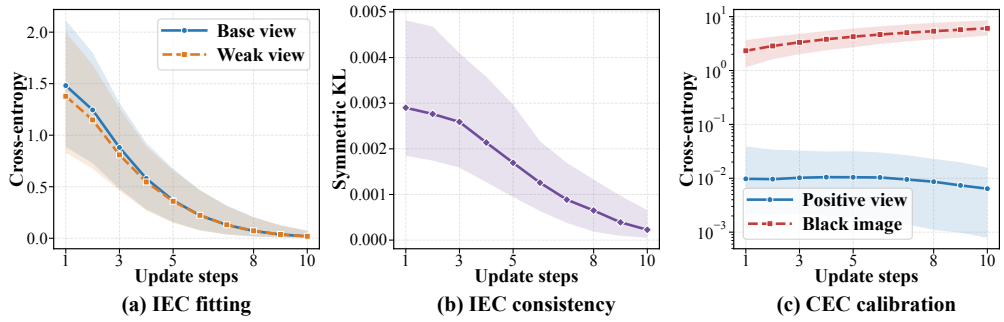


Figure 7: Training dynamics on VizWiz: (a) pseudo-target fitting, (b) weak-view consistency during IEC, and (c) relative visual-support calibration during CEC.

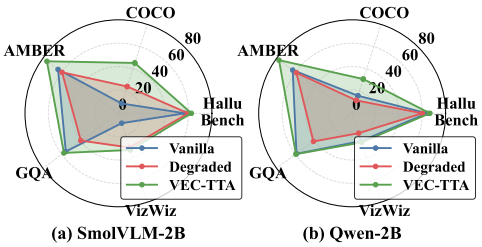


Figure 8: Visualization of performance under degraded pseudo-supervision.

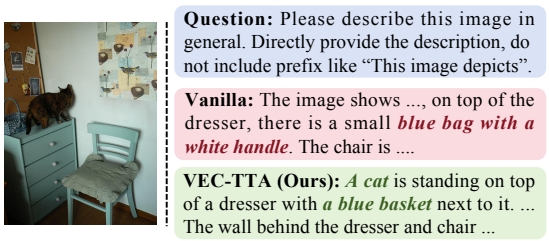


Figure 9: Qualitative comparison on COCO between the vanilla model and VEC-TTA.

Effect of Pseudo-Supervision Quality. To examine the role of pseudo-supervision quality in D_x, we construct a degraded variant in which auxiliary questions are generated from the original image, while their answers are generated using an all-black image. During adaptation, these degraded answers are fitted using the original or weakly augmented images, and CEC continues to compare their support under the original and black-image conditions. Evaluation is performed on the original benchmark images. As shown in Figure 8, this degraded supervision substantially reduces the gains of VEC-TTA across both backbones and most benchmarks. These results highlight the importance of pseudo-supervision quality for both IEC and CEC. IEC fits the supplied pseudo-targets while encouraging cross-view consistency, but agreement across views does not by itself ensure that these targets are correct. CEC subsequently calibrates the relative support for the same targets under the associated and black-image conditions, without directly verifying their correctness. The performance degradation therefore suggests that evidence consolidation and relative-support calibration cannot fully compensate for degraded pseudo-supervision.

Qualitative Analysis. Since most evaluated benchmarks require short direct answers, we use COCO for qualitative visualization, where free-form descriptions provide a clearer view of how visual content is reflected in model outputs. As shown in Figure 9, the vanilla model overlooks the salient cat in the scene and introduces an inaccurate fine-grained attribute by describing the blue handle of the basket as white. In contrast, VEC-TTA identifies the cat and focuses on other directly observable content, including the blue basket. This suggests that our proposed method VEC-TTA ties model predictions more closely to key visual evidence.

5 Conclusion

We introduced VEC-TTA, a test-time adaptation framework for compact MLLMs motivated by the observation that model predictions can remain insufficiently grounded in query-relevant visual evidence. Using temporary low-rank updates, VEC-TTA combines query-conditioned pseudo-supervision with pseudo-target fitting and weak-view consistency, followed by relative-support calibration between associated images and their all-black counterparts. Experiments on two compact MLLMs show improved performance across multiple benchmarks, while ablations support the complementary roles of IEC and CEC. Matched visual interventions further indicate that the adapted predictions are more selectively sensitive to answer-relevant visual regions.

AI USE STATEMENT

In this work, generative AI tools were used in several capacities. First, the evaluated MLLMs were used as part of the proposed VEC-TTA framework to construct query-conditioned pseudo-supervision at test time, including the generation, refinement, and verification of auxiliary question–answer pairs described in Section 3. These model-generated outputs constitute intermediate supervision signals used by the method rather than ground-truth annotations. Second, ChatGPT was used to draft portions of the manuscript based on technical content, experimental results, and interpretations provided by the authors. Finally, it was used to improve the readability, concision, and grammatical quality of author-written text. All AI-assisted manuscript content modifications were reviewed and revised by the authors. The authors verified the technical claims, numerical results, citations, experimental interpretations and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies test-time adaptation for multimodal language models and does not involve the collection of new human-subject data or personally identifiable information. All datasets and pretrained models are used subject to their applicable licenses and terms of use. The proposed method relies on model-generated pseudo-supervision at test time, whose quality may vary across instances; accordingly, our claims are restricted to the evaluated models, benchmarks, and experimental settings. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We support reproducibility by documenting the QEE pseudo-supervision construction procedure and the IEC and CEC objectives in Section 3. The experiments section describes the evaluated backbones, datasets, comparison methods, and default adaptation settings. Appendix A.1 provides additional details on LoRA parameterization, placement, and optimization, while Appendix B defines the evaluation metrics and diagnostic measures. The appendix also describes the prompt templates, evaluation-subset selection procedure, and hyperparameter settings. Together, these descriptions specify the instance-specific adaptation workflow and the protocols used for evaluation and analysis. Our code is available at `https://anonymous.4open.science/r/VEC-TTA-E0C3` to facilitate reproduction of the reported results.

REFERENCES

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. *arXiv preprint arXiv:2603.03192*, 2026.

Leon Liangyu Chen, Haoyu Ma, Zhipeng Fan, Ziqi Huang, Animesh Sinha, Xiaoliang Dai, Jialiang Wang, Zecheng He, Jianwei Yang, Chunyuan Li, et al. Unit: Unified multimodal chain-of-thought test-time scaling. *arXiv preprint arXiv:2602.12279*, 2026.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Hamidreza Dastmalchi, Aijun An, and Ali Cheraghian. Etta: Efficient test-time adaptation for vision-language models through dynamic embedding updates.

Xinqi Fan, Xueli Chen, Luoxiao Yang, Chuin Hong Yap, Rizwan Qureshi, Qi Dou, Moi Hoon Yap, and Mubarak Shah. Test-time retrieval-augmented adaptation for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8810–8819. IEEE, 2025.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *Advances in Neural Information Processing Systems*, 38:105233–105258, 2026.

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

Zenghao Guan, Yucan Zhou, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 38:6616–6639, 2026.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

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.

Jingcheng Hu, Yinmin Zhang, Shijie Shang, Xiaobo Yang, Yue Peng, Zhewei Huang, Hebin Zhou, Xin Wu, Jie Cheng, Fanqi Wan, et al. Pacore: Learning to scale test-time compute with parallel coordinated reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27200–27215, 2026.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Ilse CF Ipsen and Arvind K Saibaba. Stable rank and intrinsic dimension of real and complex matrices. *SIAM Journal on Matrix Analysis and Applications*, 46(3):1988–2007, 2025.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. Causallens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Mehmet Onurcan Kaya, Desmond Elliott, and Dim Papadopoulos. Efficient test-time scaling for small vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Marc Lafon, Gustavo Vargas Hakim, Clément Rambour, Christian Desrosiers, and Nicolas Thome. Cliptta: Robust contrastive vision-language test-time adaptation. *Advances in Neural Information Processing Systems*, 38:71849–71883, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, et al. Groundingme: Exposing the visual grounding gap in mllms through multi-dimensional evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2412–2422, 2026.

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou, and Xuming Hu. Vision-language introspection: Mitigating overconfident hallucinations in mllms via interpretable bi-causal steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38518–38543, 2026.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, et al. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9097. IEEE, 2025a.

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, and Xunliang Cai. The role of visual modality in multimodal mathematical reasoning: Challenges and insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22596–22611, 2025b.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Oscar Mañas, Pierluca D'Oro, Koustuv Sinha, Adriana Romero-Soriano, Michal Drozdzal, and Aishwarya Agrawal. Controlling multimodal llms via reward-guided decoding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1391–1401. IEEE, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Carsten Eickhoff, Ritambhara Singh, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision–language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41894–41912, 2026.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yang Shi, Huanqian Wang, Xie Xie, Huanyao Zhang, Lijie Zhao, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, et al. Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios. *Advances in Neural Information Processing Systems*, 38, 2026.

Chaehee Song, Minseok Seo, Yeeun Seong, Doyi Kim, and Changick Kim. Query-conditioned test-time self-training for large language models. *arXiv preprint arXiv:2605.13369*, 2026.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5158–5171, 2025.

Kaihua Tang, Jiaxin Qi, Jinli Ou, Yuhua Zheng, and Jianqiang Huang. Scaling test-time robustness of vision-language models via self-critical inference framework. *ArXiv*, abs/2603.07659, 2026.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b.

Zixiao Wang, Yuxin Wang, Xiaorui Wang, Mengting Xing, Jie Gao, Jianjun Xu, Guangcan Liu, Chenhui Jin, Zhuo Wang, Shengzhuo Zhang, et al. Test-time scaling with reflective generative model. *arXiv preprint arXiv:2507.01951*, 2025.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for mllms. In *International Conference on Learning Representations*, volume 2026, pp. 78875–78894, 2026.

Yi Xin, Siqi Luo, Tianxiang Xu, Qi Qin, Haoxing Chen, Kaiwen Zhu, Zhiwei Zhang, Yangfan He, Rongchao Zhang, Jinbin Bai, et al. dmllm-tts: Self-verified and efficient test-time scaling for diffusion multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35726–35735, 2026.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098. IEEE, 2025.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905, 2026.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Gengwei Zhang, Jie Peng, Zhen Tan, Mufan Qiu, Hossein Nourkhiz Mahjoub, Vaishnav Tadiparthi, Kwonjoon Lee, Yanyong Zhang, and Tianlong Chen. Understanding the role of hallucination in reinforcement post-training of multimodal reasoning models. *arXiv preprint arXiv:2604.03179*, 2026.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29999–30009. IEEE, 2025.

A Implementation Details

In this section, we provide additional implementation details of VEC-TTA that are omitted from the main paper due to space limitations. We first describe how the temporary test-time updates are parameterized and optimized, and then provide the definitions used in our LoRA update analyses.

A.1 LoRA Parameterization and Optimization

LoRA Parameterization. VEC-TTA performs instance-specific adaptation through LoRA Hu et al. (2022) while keeping the pretrained multimodal backbone frozen. For a linear module m with pretrained weight $W_m^0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we parameterize its temporary correction as

$$\Delta W_m(\phi) = \frac{\alpha}{r} B_m A_m, \tag{10}$$

where $A_m \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B_m \in \mathbb{R}^{d_{\text{out}} \times r}$ are the trainable low-rank factors, collectively denoted by ϕ. The corresponding forward computation is

$$h' = \left(W_m^0 + \Delta W_m(\phi)\right) h. \tag{11}$$

Adapter dropout is omitted from the above expression for clarity. For each test instance, the temporary LoRA adapter is reinitialized to yield zero effective corrections, optimized using the adaptation objectives, and discarded after the final prediction.

LoRA Placement and Configuration. We use rank $r = 4$, scaling factor $\alpha = 8$, and dropout rate 0.05. We apply LoRA to the query, key, value, and output projections in the attention modules of both the vision and language transformers. No LoRA modules are applied to the MLP layers, vision–language projector or connector, token embeddings, or language-model head. We do not adapt bias parameters.

Optimization Settings. Both Sequential and Joint use AdamW with a learning rate of 1×10^{-4}, zero weight decay, and a maximum gradient norm of 1.0. All adaptation updates are performed in BF16 precision.

Sequential IEC–CEC Optimization. IEC and CEC sequentially optimize the same LoRA adapter. Starting from the initial adapter, IEC first performs five update steps using the objective in Equation 8. CEC then continues from the LoRA parameters obtained after IEC and performs another five update steps using Equation 9. The optimizer is reinitialized when transitioning from IEC to CEC, so the adapted LoRA parameters are retained while the optimizer states are reset.

Joint IEC–CEC Optimization. For comparison, Joint uses the same IEC and CEC objectives but optimizes them simultaneously from the initial LoRA adapter. At each update step, the IEC and CEC losses are computed separately and combined as

$$\mathcal{L}^{(s)}_{\text{Joint}} = \mathcal{L}^{(s)}_{\text{IEC}} + \mathcal{L}^{(s)}_{\text{CEC}}. \tag{12}$$

Joint performs ten consecutive update steps using this combined objective. Sequential and Joint therefore use the same number of optimizer updates, while differing in how the IEC and CEC objectives are scheduled during adaptation.

A.2 Datasets

We evaluate VEC-TTA on five benchmarks covering image captioning, general visual question answering, real-world visual understanding, and hallucination-sensitive evaluation.

COCO. COCO Chen et al. (2015) is a widely used image-captioning benchmark containing natural images paired with human-written descriptions. In our evaluation, COCO is used for free-form image captioning. This setting provides an open-ended evaluation of whether the generated description captures salient objects, attributes, and relations in the image.

AMBER. AMBER Wang et al. (2023a) is designed to evaluate hallucination in MLLMs by examining whether model responses remain consistent with the provided visual content. Its discriminative evaluation includes questions about object existence, object attributes, and relations between visual entities. These questions evaluate whether predictions agree with the observed image.

GQA. GQA Hudson & Manning (2019) is a compositional visual question answering benchmark built on real-world images. Its questions cover objects, attributes, spatial relations, comparisons, and compositions of multiple visual concepts. The benchmark therefore evaluates the model's ability to combine multiple pieces of visual information when answering structured questions.

VizWiz. VizWiz Gurari et al. (2018) is a visual question answering benchmark collected from real-world user queries and images. Its images exhibit challenging conditions such as blur, poor framing, unusual viewpoints, low illumination. The naturally occurring questions and diverse image quality make it a challenging setting for visual understanding under realistic conditions.

HalluBench. HalluBench Guan et al. (2024) is an image-context reasoning benchmark designed to evaluate visual illusion and knowledge hallucination in MLLMs. It contains visual questions spanning diverse domains and visual formats and organizes them into controlled groups for examining model responses under potentially misleading visual or knowledge cues.

A.3 Prompts

Figures 10 and 11 present the prompt templates used in QEE. For semantic QA construction, the supporting prompt asks the model to generate a simpler question about a clearly observable fact that can help answer the original query, while the equivalent prompt rewrites the original question while preserving the same visual fact. For prior-aware construction, the model receives the original question together with the all-black image and generates a concise prior response using the corresponding answer prompt. The prior response is then inserted into a fixed template with the original

Supporting prompt

Look at the image.
Original question is: {original_question}
Write one easier question about a clearly visible fact that helps answer the original question.Output one complete question only. End it with a question mark.

Equivalent prompt

Look at the image.
Original question: {original_question}
Rewrite the original question in simpler words while asking about exactly the same visual fact.Output only one complete question ending with a question mark.

Figure 10: Prompt templates for supporting and equivalent question generation.

Prior-aware Answer

{question}
Give only the most likely concise answer. Do not explain.

Prior-aware Question

Someone proposed {prior} without seeing the image.Verify the answer from the image: {question}

Figure 11: Prompt templates for prior-response generation and prior-aware question construction.

question to construct the prior-aware question. Thus, the model generates only the prior response in this step, while the prior-aware question is constructed from the template.

A.4 Experimental Protocol

Evaluation Subsets. To keep the computational cost of instance-specific adaptation manageable across methods and backbones, we use a fixed evaluation budget to support comparisons across methods, backbones, and ablation settings. For COCO, AMBER, GQA, and VizWiz, we select 1,000 image–query pairs from each benchmark. Given an ordered list of M eligible instances, we set the stride to $\max(1, \lfloor M/1000 \rfloor)$ and retain the first 1,000 entries obtained by strided selection. For HalluBench, we exclude text-only questions and evaluate all 951 remaining image-present pairs. This deterministic selection procedure yields fixed, reproducible evaluation sets comprising 4,951 image–query pairs across five benchmarks. For Vision-OPD-6K dataset, we apply the same fixed-stride procedure to select 1,000 image–query pairs from its 6,241 original instances for the visual-intervention experiments.

Pseudo-supervision Budget. The default QEE set $\mathcal{D}_x$ contains five entries, one from each construction type. For the sample-number analysis in Figure 3(a), budgets below five are formed by randomly selecting entry types without replacement. For budgets from six to nine, we include one entry of each type and randomly select distinct types to provide the additional entries. At a budget of ten, each type contributes two entries. The resulting set is shared by IEC and CEC.

LoRA Rank and Scaling. We use $r = 4$ and $\alpha = 8$ to limit the number of trainable parameters while retaining performance comparable to larger ranks. Figure 3(b) compares $r \in \{4, 8, 16, 32\}$ with $\alpha/r = 2$ throughout; $r = 0$ denotes the unadapted baseline. The default configuration is shared across both backbones and all benchmarks.

IEC Hyperparameter Selection. The default IEC loss weights were fixed during method development using only a development subset sampled from COCO with the same fixed-stride strategy described above. We used COCO for development because its free-form outputs provide richer diagnostic signals for analyzing adaptation behavior than the predominantly short-answer benchmarks. This development subset was constructed separately from the evaluation protocol and was not identical to the COCO subset used for the reported benchmark results. Based on these development experiments, we fixed $\lambda_w = 0.5$ and $\lambda_{\mathrm{KL}} = 0.1$ and then used the same values for both backbones and all benchmarks without further dataset-specific tuning. The sensitivity study in Figure 3(c) was conducted after fixing these defaults and is intended to characterize robustness to the IEC loss weights rather than to select them.

B Evaluation and Analysis Metrics

This section provides detailed descriptions of the metrics and diagnostic measures used throughout the paper. Specifically, we organize them according to the question addressed by each analysis: task performance and prediction transitions, visual evidence diagnostics, training dynamics, and LoRA update geometry. Each subsection first introduces the shared scoring or computational setup and then describes the individual measures and their aggregation.

B.1 Task Performance and Prediction Transitions

Task scores measure overall performance on each benchmark, while prediction transitions characterize how VEC-TTA changes instance-level correctness relative to a baseline.

Task Performance Metrics. We report ROUGE-L for COCO and accuracy for AMBER, GQA, VizWiz, and HalluBench, following the scoring conventions used in the main experiments. ROUGE-L measures longest-common-subsequence overlap between a generated caption and its reference, while accuracy measures the proportion of predictions scored as correct under the benchmark-specific evaluation procedure.

Prediction Transitions. We compute prediction transitions for the four accuracy-based benchmarks and exclude COCO because its primary evaluation metric is ROUGE-L rather than binary correctness. For Table 4, let $c_i^b, c_i^v \in \{0, 1\}$ denote the correctness of the baseline and VEC-TTA predictions on instance i, respectively, where the baseline is either Vanilla or TTAdapt depending on the comparison, and let n denote the number of evaluated instances. Here, F and T denote incorrect and correct predictions, respectively. The reported transition percentages are

$$\begin{aligned} \mathrm{F}\to\mathrm{T} &= \frac{100}{n}\sum_{i=1}^{n}(1-c_i^b)c_i^v, \\ \mathrm{T}\to\mathrm{F} &= \frac{100}{n}\sum_{i=1}^{n}c_i^b(1-c_i^v), \\ \mathrm{T}\to\mathrm{T} &= \frac{100}{n}\sum_{i=1}^{n}c_i^b c_i^v. \end{aligned} \tag{13}$$

All transition percentages are normalized by the full evaluated sample set rather than by the subset of instances that are initially correct or incorrect.

B.2 Visual Evidence Diagnostics

We assess visual evidence using two complementary analyses based on layer-wise answer preferences and controlled visual interventions. The layer-wise analysis characterizes how the model's preference between answer candidates changes across layers, while the intervention analysis evaluates selective sensitivity to annotated evidence regions and the discrepancy between black-image and local evidence-removal scores. These diagnostics capture different aspects of model behavior under controlled changes to the visual input. For an evaluated instance i, let (I_i, q_i) denote its image–query pair and $I_i^{\mathrm{n}} = \mathcal{T}_{\mathrm{n}}(I_i)$ denote the corresponding black-image input. We use J to denote a generic visual input, consistent with Equation 5, and n to denote the number of evaluated instances, distinct from N, which denotes the number of entries in $\mathcal{D}_x$.

Layer-wise Evidence Strength. Because the model is instructed to output only the answer, we examine how its answer preference evolves across language layers rather than relying on an explicit reasoning trace. For each instance i, we compare two answer candidates, namely the reference answer y_i^{ref} and the response $\hat{y}_i^{\mathrm{n}}$ generated under the null-visual input I_i^{n}. We use their first tokens $y_{i,1}^{\mathrm{ref}}$ and $\hat{y}_{i,1}^{\mathrm{n}}$ because the model directly outputs the answer, making the first answer token a consistent signal of its initial answer preference across layers. At language layer l, we pass the hidden state at the final prompt position through the model's final normalization and output head to obtain the diagnostic logit $z_l(v \mid J, q_i)$ for vocabulary token v. We first measure the relative preference between the reference token and the null-visual response token under the original and null-visual inputs as

$$\begin{aligned} M_{i,l}^{I} &= z_l(y_{i,1}^{\mathrm{ref}} \mid I_i, q_i) - z_l(\hat{y}_{i,1}^{\mathrm{n}} \mid I_i, q_i), \\ M_{i,l}^{\mathrm{n}} &= z_l(y_{i,1}^{\mathrm{ref}} \mid I_i^{\mathrm{n}}, q_i) - z_l(\hat{y}_{i,1}^{\mathrm{n}} \mid I_i^{\mathrm{n}}, q_i). \end{aligned} \tag{14}$$

Here, $M^I_{i,l}$ measures the answer preference when the original image is available, while $M^{\mathrm{n}}_{i,l}$ measures the same preference when visual evidence is restricted. We then separate the image-induced gain from the preference that remains under the null-visual condition as

$$\begin{aligned} V_{i,l} &= M^I_{i,l} - M^{\mathrm{n}}_{i,l}, \\ P_{i,l} &= -M^{\mathrm{n}}_{i,l}. \end{aligned} \tag{15}$$

$V_{i,l}$ measures how much restoring the original image increases the reference token's relative advantage over the null-visual response token, which we use as the layer-wise visual-evidence strength. $P_{i,l}$ measures how strongly the null-visual response token is preferred over the reference token when visual evidence is restricted, which we use as the layer-wise language-prior strength. These quantities satisfy $M^I_{i,l} = V_{i,l} - P_{i,l}$, so the original-image answer preference reflects the balance between the visual-evidence gain and the preference retained under the null-visual condition. Figure 1(a) reports the median and interquartile range of $V_{i,l}$ and $P_{i,l}$ across instances at each language layer.

Per-sample Logit Margin. To examine how the balance between visual-evidence strength and language-prior strength varies across instances, we compute their net margin at each language layer. Using the definitions above, this margin is

$$D_{i,l} = V_{i,l} - P_{i,l} = M^I_{i,l} = z_l(y^{\mathrm{ref}}_{i,1} \mid I_i, q_i) - z_l(\hat{y}^{\mathrm{n}}_{i,1} \mid I_i, q_i). \tag{16}$$

A positive $D_{i,l}$ indicates that visual-evidence strength exceeds language-prior strength, corresponding to a higher logit for the reference token under the original image. A negative $D_{i,l}$ indicates that language-prior strength remains larger, corresponding to a higher logit for the null-visual response token even when the original image is available. Figure 1(b) visualizes $D_{i,l}$ across instances and language layers, with each column representing an instance and each row representing a language layer. The columns are sorted by their average margin over the final quarter of language layers, exposing instance-level differences in how this balance evolves toward the model output.

Counterfactual Evidence Removal. To provide a behavioral view of visual dependence, we additionally compare model performance under the original image and an evidence-removed counterpart. For each evaluated instance, we remove the annotated image region containing the key visual evidence required by the query while preserving the remaining image content. The modified image therefore retains the surrounding visual context but removes the evidence intended to support the original answer. We then evaluate the model on the same query using either the original or the evidence-removed image, while keeping the evaluation target and all other settings unchanged. Figure 1(c) reports performance under these two conditions across the three datasets. A large performance decrease after evidence removal indicates that the prediction is sensitive to the query-relevant visual evidence, whereas retained performance indicates that the model can continue producing the original answer despite removal of that evidence.

Visual-Intervention Scores. We evaluate whether model predictions are selectively sensitive to answer-relevant visual regions by comparing evidence masking with matched control masking. For each instance i in Table 3, let I_i denote the original image, I^{E}_i its evidence-masked counterpart, and $I^{\mathrm{C}}_{i,1}, \ldots, I^{\mathrm{C}}_{i,K}$ the K matched control-masked images. Evidence masking removes the annotated answer-relevant region, while each control mask removes a same-sized region that does not overlap the evidence region. The superscripts E and C denote evaluation interventions and are distinct from the semantic-question superscripts used in QEE. Across visual conditions, we keep the query, scored answer, and model parameters fixed, using the same instance-specific adapter for adapted models. Accuracy is evaluated against the original reference answer.

Evidence Selectivity. ES measures whether removing the annotated evidence decreases model support for the ground-truth answer more than removing a matched control region. Let g_i denote the ground-truth option letter and $T_i = (t_{i,1}, \ldots, t_{i,|T_i|})$ its token sequence. We measure its support using the token-averaged NLL

$$L_{i,\mathrm{GT}}(J) = -\frac{1}{|T_i|} \sum_{u=1}^{|T_i|} \log p_{\theta,\phi_i}(t_{i,u} \mid J, q_i, t_{i,<u}). \tag{17}$$

Here, ϕ_i denotes the evaluated adapter for instance i and remains fixed across visual conditions, while the unadapted baseline uses p_θ. The NLL is computed for the ground-truth option letter rather than the full option text and is not renormalized over the candidate options.

For each instance, we measure the increase in ground-truth NLL caused by evidence masking and matched control masking as

$$\begin{aligned}\delta_i^{\mathrm{E}} &= L_{i,\mathrm{GT}}(I_i^{\mathrm{E}}) - L_{i,\mathrm{GT}}(I_i),\\ \delta_i^{\mathrm{C}} &= \frac{1}{K}\sum_{j=1}^{K}\left[L_{i,\mathrm{GT}}(I_{i,j}^{\mathrm{C}}) - L_{i,\mathrm{GT}}(I_i)\right].\end{aligned} \tag{18}$$

We then define the evidence-selectivity score as

$$\begin{aligned}\mathrm{ES} &= \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i^{\mathrm{E}} - \delta_i^{\mathrm{C}}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[L_{i,\mathrm{GT}}(I_i^{\mathrm{E}}) - \frac{1}{K}\sum_{j=1}^{K} L_{i,\mathrm{GT}}(I_{i,j}^{\mathrm{C}})\right].\end{aligned} \tag{19}$$

Subtracting the two intervention effects removes the shared original-image NLL term. A positive ES therefore indicates that removing the annotated evidence decreases support for the ground-truth answer more than removing a same-sized control region. ES measures the selectivity of this visual sensitivity and is interpreted together with the corresponding accuracy results.

Proxy Gap. PG measures the discrepancy between black-image replacement and local evidence removal using the same fixed QEE pseudo-answer. For each instance i, let y_i denote the pseudo-answer generated by QEE and write $L_i(J) = \mathcal{L}_{\mathrm{CE}}(J, q_i, y_i)$. We define

$$\begin{aligned}\epsilon_i &= \left|L_i(I_i^{\mathrm{n}}) - L_i(I_i^{\mathrm{E}})\right|,\\ \mathrm{PG} &= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i.\end{aligned} \tag{20}$$

For each instance, the same QEE pseudo-answer y_i is used across all compared methods, stages, and visual conditions without regeneration after adaptation. The query, answer tokens, teacher-forced prefixes, and token normalization are therefore held fixed when comparing the two visual inputs. PG averages the absolute black–evidence NLL difference within each instance before averaging across instances. A smaller PG indicates a smaller discrepancy between the answer scores induced by black-image replacement and local evidence removal. Appendix D.1 shows how PG bounds the corresponding discrepancy between the two calibration objectives.

B.3 Training Dynamics

Figures 7 and 12 visualize the optimization signals recorded during the actual IEC and CEC updates. During IEC, we track the answer-token cross-entropy on the entries in $\mathcal{D}_x$ and their online weak views, together with the symmetric KL divergence between the paired next-token distributions. These quantities correspond to R_{pos}, $R_w^{(s)}$, and $D_{\mathrm{SKL}}^{(s)}$ in the IEC objective. During CEC, we track the answer-token cross-entropy under the positive and black-image conditions, whose difference forms the evidence margin Δ_k used by the CEC objective.

Answer Fitting and Relative Support. We measure answer fitting using the answer-token-averaged cross-entropy in Equation 5, computed only over pseudo-answer tokens. For each instance, we average this quantity over the entries in $\mathcal{D}_x$ separately for each visual condition. During IEC, the averages for the base and online weak views correspond to $\mathcal{R}_{\mathrm{pos}}$ and $\mathcal{R}_{\mathrm{w}}^{(s)}$ in Equation 6. During CEC, we report the cross-entropy under the positive and black-image conditions. For each entry, their black-minus-positive difference gives the evidence margin Δ_k in Equation 9. Lower cross-entropy indicates stronger support for the corresponding pseudo-answer.

Weak-view Consistency. We measure weak-view consistency using $D_{\mathrm{SKL}}^{(s)}$ in Equation 7. This quantity compares the full-vocabulary next-token distributions of the paired base and online weak views through the two directional KL divergences. The divergences are averaged over pseudo-answer token positions within each entry and then over the entries in $\mathcal{D}_x$. Lower values indicate closer agreement between the paired distributions.

Aggregation Across Instances. At each update step, we summarize each quantity using the median and interquartile range of its instance-level values.

B.4 LoRA Update Geometry

We characterize the effective LoRA updates in terms of their magnitude, direction, and spectral structure. All measures are computed from the resulting effective weight corrections rather than from the individual low-rank factors. We distinguish between stage-wise increments for analyzing IEC and CEC and final corrections for comparing Sequential and Joint optimization.

Notation. For test instance i and LoRA-augmented module m, let $W_{i,m}^{(1)}$ and $W_{i,m}^{(2)}$ denote the effective LoRA corrections after IEC and CEC, respectively. Since the effective correction is zero before adaptation, we define the stage-wise increments as

$$\begin{aligned} U_{i,m}^{\mathrm{IEC}} &= W_{i,m}^{(1)}, \\ U_{i,m}^{\mathrm{CEC}} &= W_{i,m}^{(2)} - W_{i,m}^{(1)}. \end{aligned} \tag{21}$$

Thus, $U_{i,m}^{\mathrm{IEC}}$ is the effective correction accumulated during IEC, while $U_{i,m}^{\mathrm{CEC}}$ is the net change in the effective correction introduced during CEC.

For comparing Sequential and Joint optimization, let $F_{i,m}^{h}$ denote the final effective LoRA correction produced by optimization schedule h. It is given by

$$F_{i,m}^{h} = \frac{\alpha}{r} B_{i,m}^{h} A_{i,m}^{h}, \qquad h \in \{\mathrm{Sequential}, \mathrm{Joint}\}. \tag{22}$$

For Sequential optimization, $F_{i,m}^{\mathrm{Sequential}} = W_{i,m}^{(2)}$.

Let $\mathcal{M}$ denote the set of analyzed LoRA-augmented modules. For $g \in \{\mathrm{Vision}, \mathrm{Language}\}$, let $\mathcal{M}_g \subseteq \mathcal{M}$ denote the modules belonging to group g.

Frobenius Norm and Inner Product. For a matrix update U, the Frobenius norm measures its overall update magnitude. Together with the Frobenius inner product, it provides the basis for comparing update scale and directional alignment.

$$\begin{aligned} \|U\|_F &= \left(\sum_{a,b} U_{ab}^2 \right)^{1/2}, \\ \langle U, V \rangle_F &= \sum_{a,b} U_{ab} V_{ab} = \mathrm{tr}(U^\top V). \end{aligned} \tag{23}$$

Update Magnitude and Norm Ratio. For stage $s \in \{\mathrm{IEC}, \mathrm{CEC}\}$, we measure the overall update magnitude for each instance by aggregating the Frobenius energy across the analyzed modules.

$$N_i^{(s)} = \left(\sum_{m \in \mathcal{M}} \|U_{i,m}^{(s)}\|_F^2 \right)^{1/2}. \tag{24}$$

We summarize $N_i^{(s)}$ across instances using the median and a percentile 95% bootstrap interval. To compare the relative scales of the two stage-wise increments within each module, we further define

$$\rho_{i,m} = \frac{\|U_{i,m}^{\mathrm{CEC}}\|_F}{\|U_{i,m}^{\mathrm{IEC}}\|_F}. \tag{25}$$

The ratio is evaluated only for modules with nonzero IEC updates. A ratio of one indicates equal update magnitudes, while values above or below one indicate a larger or smaller CEC increment.

Update Alignment. We measure directional alignment between two corresponding nonzero updates using their Frobenius cosine similarity.

$$\cos_F(U, V) = \frac{\langle U, V \rangle_F}{\|U\|_F \|V\|_F}. \tag{26}$$

A value of one indicates the same direction up to positive scaling, zero indicates orthogonality, and negative values indicate opposing directions. For the comparison between Sequential and Joint optimization, we compute the similarity between their final effective corrections.

$$\begin{aligned} a_{i,m} &= \cos_F(F_{i,m}^{\mathrm{Sequential}}, F_{i,m}^{\mathrm{Joint}}), \\ a_{i,g} &= \frac{1}{|\mathcal{M}_{i,g}^{+}|} \sum_{m \in \mathcal{M}_{i,g}^{+}} a_{i,m}. \end{aligned} \tag{27}$$

Here, $\mathcal{M}_{i,g}^{+}$ contains the matched modules in group $g \in \{\mathrm{Vision}, \mathrm{Language}\}$ whose final updates are nonzero under both optimization schedules. We summarize the instance-level group similarities using their median and interquartile range within each benchmark. For the stage-wise comparison between IEC and CEC, we instead compute

$$b_{i,m} = \cos_F(U_{i,m}^{\mathrm{IEC}}, U_{i,m}^{\mathrm{CEC}}). \tag{28}$$

For the projection-wise layer analysis in Figure 13, we focus on the query, key, and value projections within the visual and language attention modules. These projections provide a direct decomposition of the attention computation, with the query and key projections determining attention weights and the value projection transforming the information aggregated by attention. We therefore examine their update scale and directional alignment separately across layers. The output projection is not included in this projection-wise visualization. For each instance, layer, and projection, we compute the CEC-to-IEC Frobenius-norm ratio $\rho_{i,m}$ and the IEC–CEC cosine similarity $b_{i,m}$. Layer indices refer to the respective visual and language networks and do not imply correspondence between layers of the two networks. We summarize each quantity across instances using the median and interquartile range.

Shared Spectral Computation. For each effective module update U, we compute singular values ordered as $\sigma_1(U) \geq \sigma_2(U) \geq \cdots \geq 0$. Their squared values represent spectral energy and satisfy

$$\sum_j \sigma_j^2(U) = \|U\|_F^2. \tag{29}$$

We use this decomposition to characterize absolute spectral energy, cumulative energy concentration, and stable rank.

Absolute Spectral Energy. For the stage-wise IEC and CEC analysis, we aggregate squared singular values of the corresponding increments $U_{i,m}^{(s)}$ at each singular rank. For instance i and singular rank j, the aggregated energy is $\sum_{m\in\mathcal{M}} \sigma_j^2(U_{i,m}^{(s)})$. We then report its median across instances.

$$E_j^{(s)} = \mathrm{median}_i \left[\sum_{m\in\mathcal{M}} \sigma_j^2(U_{i,m}^{(s)}) \right]. \tag{30}$$

When multiple trailing singular ranks are grouped for visualization, their rank-wise median energies are summed after aggregation across instances.

Cumulative Spectral Energy. For the comparison between Sequential and Joint optimization, let $\sigma_{i,m,j}^h$ denote the j-th singular value of the final correction $F_{i,m}^h$. For each instance i, module group g, and optimization schedule h, we define

$$C_{i,g}^h(k) = \frac{\sum_{m\in\mathcal{M}_g} \sum_{j=1}^{k} (\sigma_{i,m,j}^h)^2}{\sum_{m\in\mathcal{M}_g} \sum_{j\geq 1} (\sigma_{i,m,j}^h)^2}. \tag{31}$$

A faster increase with k indicates that a larger fraction of the update energy is concentrated in the leading singular components. Because this quantity is normalized by the total spectral energy, it characterizes spectral concentration independently of the overall update magnitude.

Stable Rank. For a nonzero module update U, the stable rank Ipsen & Saibaba (2025) is

$$\mathrm{srank}(U) = \frac{\|U\|_F^2}{\|U\|_2^2} = \frac{\sum_j \sigma_j^2(U)}{\sigma_1^2(U)}. \tag{32}$$

We first average the module-level stable ranks within each instance separately over visual and language modules. We then report the median of these instance-level averages together with a percentile 95% bootstrap interval. Stable rank characterizes the effective spectral dimensionality of an update independently of its overall scale.

C Additional Experimental Results

C.1 Additional Analysis on COCO

To complement the ROUGE-L results reported in the main paper, Table 5 provides BLEU-1, BLEU-4, and CIDEr-D for a broader evaluation of open-ended caption quality. BLEU measures modified n-gram precision against reference captions with a brevity penalty; BLEU-1 considers unigram precision, while BLEU-4 aggregates modified precisions from one- to four-grams using their

Table 5: Additional COCO results on two compact MLLMs.

Method	SmolVLM-2B			Qwen-2B		
	BLEU-1	BLEU-4	CIDEr	BLEU-1	BLEU-4	CIDEr
Vanilla	7.6	1.3	0	11.2	2.0	0.0048
Self-Consistency	7.4	1.2	0	13.8	2.3	0.025
SCI	8.0	1.4	0	11.3	2.0	0.0067
TTAdapt	8.0	1.4	0	11.2	2.0	0.010
QueST	9.4	1.7	0.31	4.8	1.0	5.4
VEC-TTA	**58.87**	**18.84**	**67.04**	**29.02**	**6.30**	**5.62**

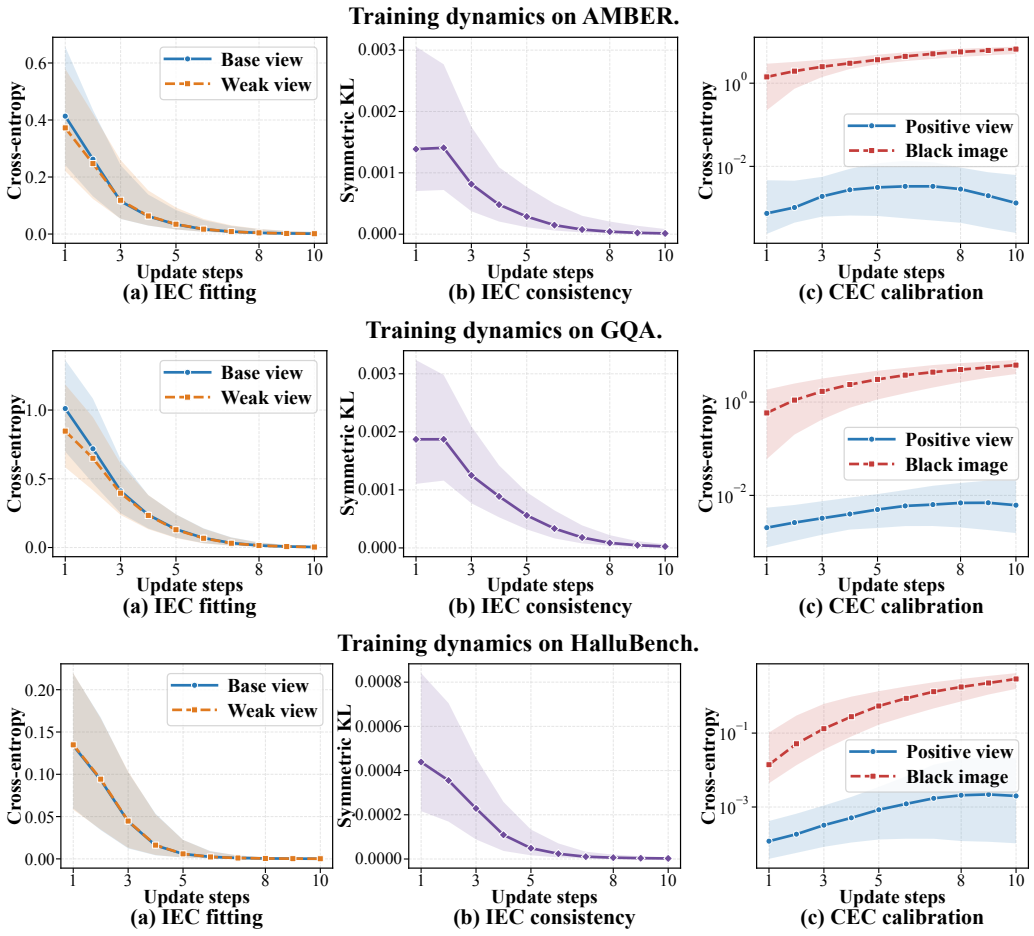


Figure 12: Visualization of training dynamics on AMBER, GQA and HalluBench, where each including (a) pseudo-supervision fitting, (b) weak-view consistency during IEC, and (c) relative answer-support calibration during CEC. Lines and shaded regions denote medians and interquartile ranges across instances.

geometric mean. We additionally report CIDEr-D Vedantam et al. (2015), computed using the `pycocoevalcap` implementation with $n = 4$ and $\sigma = 6$. For readability, CIDEr-D values smaller than 10^{-5} are reported as 0.

Results and Interpretation. Table 5 reports complementary COCO results using BLEU-1, BLEU-4, and CIDEr-D. VEC-TTA achieves the best performance across all three metrics on both backbones, consistent with the improvements observed under ROUGE-L. The gains in BLEU-1 and BLEU-4 indicate improved agreement with the reference captions at both the word and phrase lev-

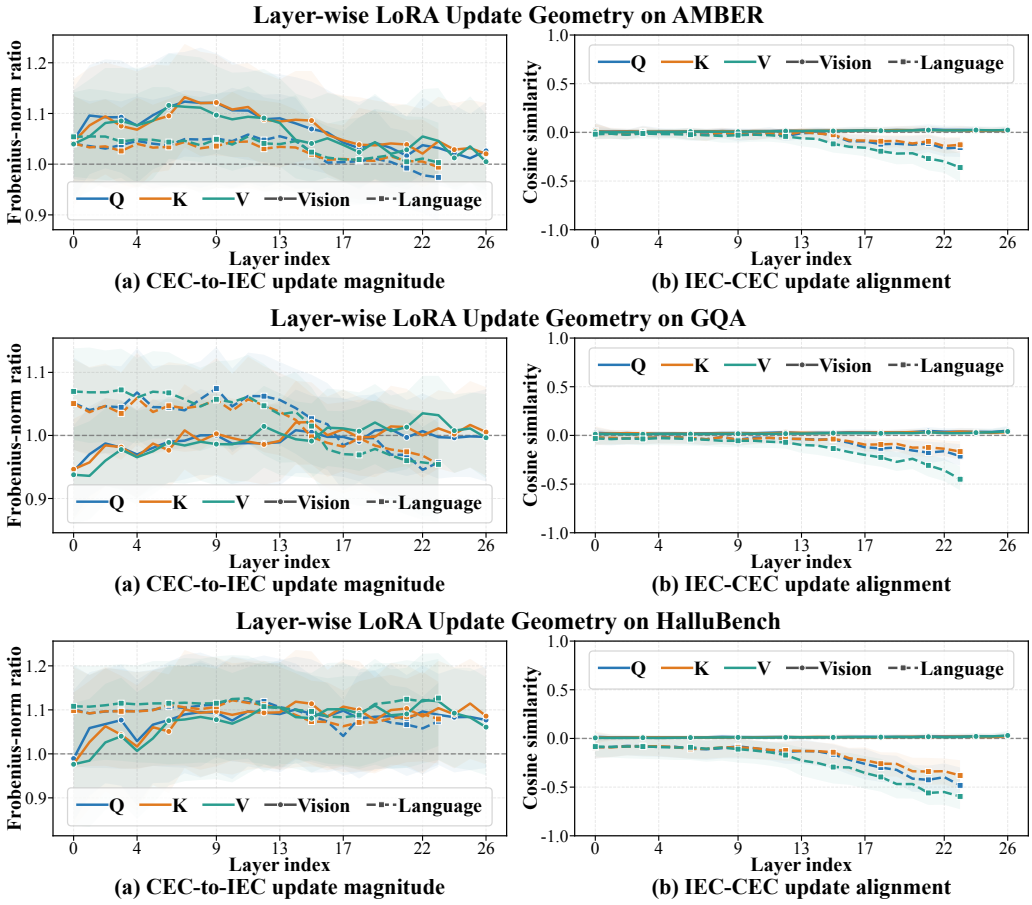


Figure 13: Layer-wise LoRA updates on AMBER, GQA and HalluBench: (a) the norm ratio of the CEC increment to the IEC update and (b) their cosine similarity. Colors distinguish query, key, and value projections; solid and dashed lines denote visual and language modules, respectively.

els. CIDEr-D exhibits a similar overall trend; however, we interpret the substantially larger improvement in this metric with caution, as CIDEr-D can be sensitive to differences in output length and response style. We additionally observe that some comparison outputs contain speculative content or unnecessarily verbose descriptions beyond the directly observable visual information, which can substantially affect the CIDEr-D score. Therefore, the magnitude of the CIDEr-D gain should not be interpreted in isolation. Overall, despite this potential confounding effect on CIDEr-D, the consistent improvements in ROUGE-L, BLEU-1, and BLEU-4, together with the same directional trend in CIDEr-D, provide complementary evidence for the effectiveness of VEC-TTA on image description.

C.2 Additional Training Dynamics

Figure 12 extends the training-dynamics analysis in the main text to AMBER, GQA, and HalluBench. Across all the datasets, IEC progressively reduces the cross-entropy on the base and online weak views while also decreasing the symmetric KL divergence between their next-token distributions. During CEC, the black-image cross-entropy increases substantially, while the positive-view cross-entropy remains comparatively low despite a moderate increase in some cases. These results exhibit the same stage-wise pattern observed on VizWiz, with IEC consolidating the pseudo-supervision under weak-view consistency and CEC increasing the separation in pseudo-answer support between the positive and black-image conditions.

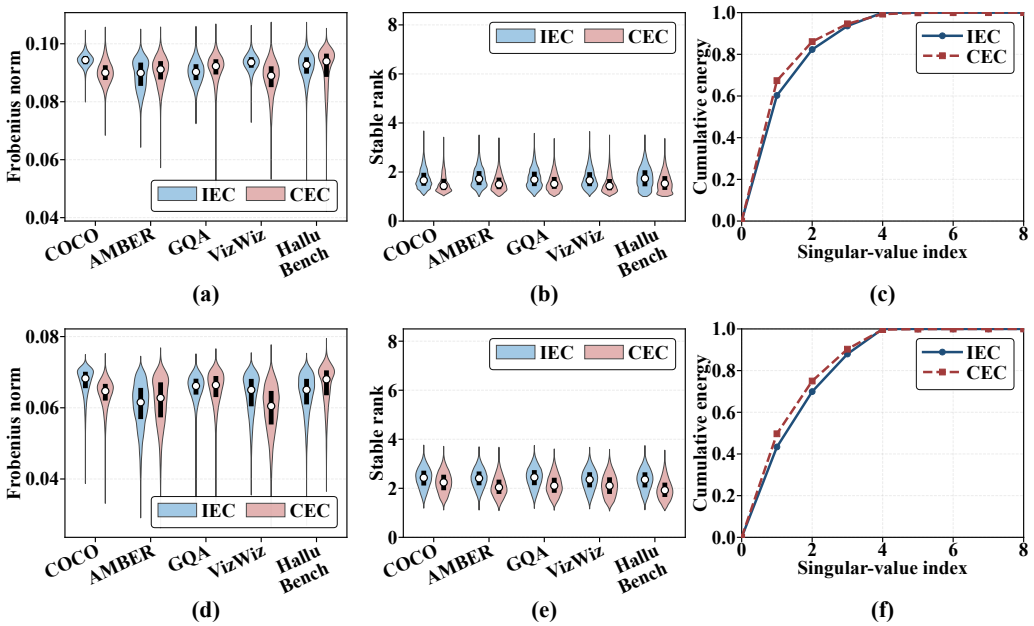


Figure 14: Modality-specific analysis of the rank-4 stage-wise LoRA updates. The top row reports visual modules and the bottom row reports language modules. Panels (a,d) show Frobenius-norm magnitudes, panels (b,e) show stable ranks, and panels (c,f) show cumulative spectral-energy profiles for IEC and CEC.

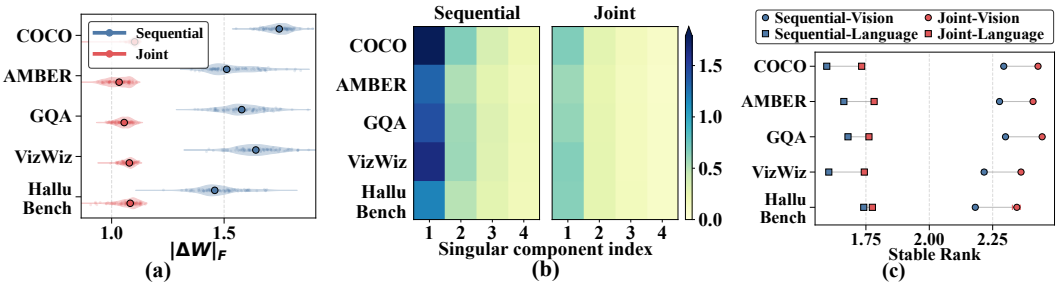


Figure 15: LoRA update geometry of sequential and joint update: (a) Frobenius-norm magnitude, (b) singular-value energy distribution, and (c) stable rank of visual and language modules.

C.3 Layer-wise Update Geometry

Figure 13 compares the IEC and CEC stage-wise updates across layers on AMBER, GQA, and HalluBench for the query, key, and value projections. Panel (a) shows CEC-to-IEC Frobenius-norm ratios generally close to one, indicating that the two stages produce updates of comparable magnitude. Panel (b) shows a different pattern in update direction. Cosine similarities in the visual modules remain largely around zero, while those in the language modules become increasingly negative in deeper layers, with the strongest negative alignment appearing in the value projections. Thus, CEC introduces updates that are similar in scale to the preceding IEC updates but differ substantially in direction, and some deeper language-module updates partially oppose the IEC direction. These results indicate that CEC does not simply amplify the update established by IEC, but instead introduces an additional correction with a distinct directional structure.

C.4 Modality-specific Stage-wise LoRA Analysis

We further decompose the default-rank stage-wise LoRA analysis into visual and language modules. As shown in Figure 14, the relative IEC–CEC pattern is preserved within both modalities. For visual

Table 6: Per-instance computational costs: trainable parameters (#Params.), adaptation and end-to-end latency, and average generated tokens for the full process (Total) and final answer (Answer).

Method	#Params.↓	Wall-clock Latency (s)		Avg. Output Tokens	
		Adaptation↓	End2End↓	Total	Answer
TTAdapt	2246.78M	190.53	207.50	736.75	2.03
VEC-TTA	**2.32M**	**93.69**	**103.69**	65.82	2.02

modules, IEC and CEC produce updates of comparable magnitude, while CEC exhibits lower stable rank and faster accumulation of spectral energy in the leading singular components. The same pattern is observed for language modules, indicating that the stage-wise difference in spectral structure persists after separating the two modalities. The visual and language modules nevertheless differ in their absolute update geometry. Visual-module updates have smaller Frobenius norms and broader spectral support, whereas language-module updates have larger magnitudes and more concentrated spectra. Despite these modality-specific differences, CEC shows greater spectral concentration than IEC within both visual and language modules.

C.5 Computational Efficiency

Table 6 compares the per-instance computational cost of VEC-TTA and TTAdapt. VEC-TTA reduces adaptation latency from 190.53 to 93.69 seconds and end-to-end latency from 207.50 to 103.69 seconds per instance. It also reduces the number of trainable parameters from 2246.78M to 2.32M by replacing full-parameter adaptation with LoRA-based updates. In addition, the average number of generated output tokens decreases from 736.75 to 65.82, while the final answer lengths remain comparable. The lower generation overhead results from constructing pseudo-supervision with a compact set of auxiliary questions and visual variations rather than extensive response generation. Together, these reductions yield substantially lower adaptation and end-to-end cost while preserving the task-performance gains reported in the main experiments.

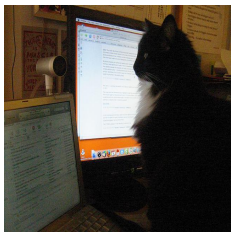
C.6 Supplementary Analysis of LoRA Update Geometry

Figure 15 provides a complementary comparison of the final LoRA updates produced by Sequential and Joint optimization under the default rank-4 setting. Panel (a) shows that Sequential produces consistently larger Frobenius-norm updates than Joint across the evaluated benchmarks, indicating a systematic difference in the overall scale of the final adapter corrections. Panel (b) further decomposes these updates by singular component. For both optimization schedules, most of the absolute spectral energy is concentrated in the leading components, while Sequential exhibits substantially larger absolute energy, particularly in the first singular component. Because this panel reports unnormalized spectral energy, the difference also reflects the larger update magnitudes observed in Panel (a). Panel (c) compares stable rank separately for visual and language modules. The two optimization schedules exhibit broadly similar effective spectral dimensionality, with only moderate differences across benchmarks and modalities. Together, these results show that Sequential and Joint differ more clearly in the scale and absolute spectral energy of their final updates than in their effective spectral dimensionality.

C.7 Additional Qualitative Analysis

We provide additional qualitative comparisons on COCO, with examples from SmolVLM-2B shown on the top and those from Qwen-2B shown on the bottom. For each example, we show the input image, the evaluation prompt, and the descriptions generated by the vanilla model and VEC-TTA. Highlighted phrases mark representative differences between the two outputs.

Success Cases. Figure 16 presents additional examples in which VEC-TTA produces descriptions that more closely reflect the visible image content. In the first example, VEC-TTA captures both the black-and-white appearance of the cat and its spatial relation to the nearby computer monitor. In the second example, VEC-TTA emphasizes the central interaction between the hand and the phone together with the visible appearance of the phone, while avoiding peripheral details introduced by



Question: Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".

Vanilla: A black cat sitting on a desk ... The ***monitor displays a web page***. The cat's fur is predominantly black...

VEC-TTA (Ours): ***A black and white cat*** is sitting on ***a desk next to a computer monitor***....

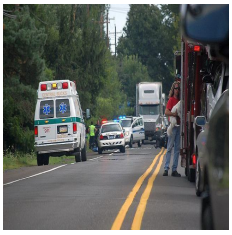


Question: Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".

Vanilla: ... the hand is holding the phone vertically ... ***a plastic cup, and a pair of sunglasses*** ..... and the plastic cup .....

VEC-TTA (Ours): A hand holds ***a white and black cell phone... white cups and a black item*** ...

Figure 16: Successful cases on COCO for SmolVLM-2B (top) and Qwen-2B (bottom).



Question: Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".

Vanilla: ...truck with a red and blue emergency light ..., including police cars and a fire truck ... ***a serious situation*** ...

VEC-TTA (Ours): An ambulance driving down road with police cars, and a truck...



Question: Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".

Vanilla: ...teddy bears are dressed in ***military-style outfits***, complete with helmets and what seems to be camouflage fabric....

VEC-TTA (Ours): ...Teddy bear wearing a green hat is holding a sandwich in its hand. there is a boat with teddy bears ...

Figure 17: Failure cases on COCO for SmolVLM-2B (top) and Qwen-2B (bottom).

the vanilla model that are not clearly supported by the image, such as sunglasses. These examples illustrate how VEC-TTA can preserve salient objects, attributes, and relations while reducing unsupported details in generated descriptions. This qualitative pattern is consistent with the stronger selective sensitivity to answer-relevant visual evidence observed in the quantitative analysis.

Failure Cases. Figure 17 presents representative cases in which VEC-TTA does not fully improve the generated description. In these examples, VEC-TTA still identifies several locally visible objects and attributes but does not always integrate them into a coherent interpretation of the overall scene. For example, recognizing individual emergency vehicles does not necessarily capture the broader emergency context, while describing visible clothing and objects separately may fail to express their shared thematic relation. These cases indicate that improved sensitivity to local visual cues does not necessarily translate into equally strong integration of multiple cues into higher-level scene descriptions.

D Additional Analysis

D.1 Analysis of the Black-Image Proxy

An all-black image removes visual content globally and therefore does not exactly reproduce the query-specific removal of relevant evidence. We analyze this approximation by comparing the CEC calibration objective under black-image replacement with the corresponding objective under local evidence masking. The resulting bound relates their objective discrepancy to the measurable proxy gap defined in Appendix B. The matched visual interventions in Table 3 provide complementary empirical evidence for the practical utility of this proxy.

Setup. Consider n image–query–answer triples (I_i, q_i, y_i) evaluated at fixed model checkpoints. Under episodic adaptation, the checkpoint may differ across instances but is held fixed across the visual conditions compared for each instance. Let I_i^{n} denote the all-black counterpart of I_i, and let I_i^{E} denote the corresponding evidence-masked image. Write $L_i(J) = \mathcal{L}_{\mathrm{CE}}(J, q_i, y_i)$ for the answer-token-averaged NLL in Equation 5. The scored answer and its token-level normalization are identical across the compared visual conditions. We define the black-image and evidence-masking margins as

$$\begin{aligned} \Delta_i^{\mathrm{n}} &= L_i(I_i^{\mathrm{n}}) - L_i(I_i), \\ \Delta_i^{\mathrm{E}} &= L_i(I_i^{\mathrm{E}}) - L_i(I_i). \end{aligned} \tag{33}$$

Using the CEC margin loss $\ell(u) = \log(1 + \exp(-u))$, the corresponding empirical calibration objectives are

$$J_v = \frac{1}{n} \sum_{i=1}^{n} \ell(\Delta_i^v), \qquad v \in \{\mathrm{n}, \mathrm{E}\}. \tag{34}$$

Here, J_{n} uses the same black-image margin as CEC, while J_{E} replaces the black-image condition with local evidence masking while retaining the same positive image and scored answer.

Proposition. For finite answer NLLs, let

$$\epsilon_i = \left| L_i(I_i^{\mathrm{n}}) - L_i(I_i^{\mathrm{E}}) \right|, \qquad \mathrm{PG} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i, \tag{35}$$

as in Equation 20. Then

$$\begin{aligned} \left| \ell(\Delta_i^{\mathrm{n}}) - \ell(\Delta_i^{\mathrm{E}}) \right| &\le \epsilon_i, \\ |J_{\mathrm{n}} - J_{\mathrm{E}}| &\le \mathrm{PG}. \end{aligned} \tag{36}$$

Thus, PG provides an empirical upper bound on the discrepancy between the black-image and local evidence-masking calibration objectives.

Proof. The CEC margin loss satisfies

$$\ell'(u) = -\frac{1}{1 + \exp(u)}, \qquad |\ell'(u)| \le 1, \tag{37}$$

so ℓ is 1-Lipschitz. Therefore,

$$\begin{aligned} \left| \ell(\Delta_i^{\mathrm{n}}) - \ell(\Delta_i^{\mathrm{E}}) \right| &\le \left| \Delta_i^{\mathrm{n}} - \Delta_i^{\mathrm{E}} \right| \\ &= \left| L_i(I_i^{\mathrm{n}}) - L_i(I_i^{\mathrm{E}}) \right| \\ &= \epsilon_i. \end{aligned} \tag{38}$$

The second equality follows because the shared original-image term $L_i(I_i)$ is removed when the two margins are subtracted. Applying the triangle inequality after averaging across instances gives

$$\begin{aligned} |J_{\mathrm{n}} - J_{\mathrm{E}}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \left[\ell(\Delta_i^{\mathrm{n}}) - \ell(\Delta_i^{\mathrm{E}}) \right] \right| \\ &\le \frac{1}{n} \sum_{i=1}^{n} \left| \ell(\Delta_i^{\mathrm{n}}) - \ell(\Delta_i^{\mathrm{E}}) \right| \\ &\le \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \\ &= \mathrm{PG}. \end{aligned} \tag{39}$$

Algorithm 1 VEC-TTA Algorithm.

Require: Test instance $x = (I, q)$, frozen backbone θ, initial LoRA ϕ_0
Ensure: Final prediction $\hat{y}$

Initialize the instance-specific adapter: $\phi \leftarrow \phi_0$
Stage 1: Query-Conditioned Evidence Elicitation (QEE)
Generate the original pseudo-answer: $y \leftarrow \mathrm{Decode}_\theta(I, q)$
Generate supporting and equivalent questions: $(q^{\mathrm{s}}, q^{\mathrm{e}}) \leftarrow \mathrm{GenerateQuestions}_\theta(I, q)$
Obtain their pseudo-answers: $y^{\mathrm{s}} \leftarrow \mathrm{Decode}_\theta(I, q^{\mathrm{s}})$, $y^{\mathrm{e}} \leftarrow \mathrm{Decode}_\theta(I, q^{\mathrm{e}})$
Construct a semantics-preserving weak view: $I^{\mathrm{w}} \leftarrow \mathrm{WeakAug}(I)$
Probe the all-black view: $\hat{y}^{\mathrm{n}} \leftarrow \mathrm{Decode}_\theta(I^{\mathrm{n}}, q)$
Build the prior-aware query using the fixed template: $q^{\mathrm{p}} \leftarrow \mathcal{T}_{\mathrm{p}}(q, \hat{y}^{\mathrm{n}})$
Form the instance-specific pseudo-supervision:

$$\mathcal{D}_x \leftarrow \{(I, q, y), (I, q^{\mathrm{s}}, y^{\mathrm{s}}), (I, q^{\mathrm{e}}, y^{\mathrm{e}}), (I^{\mathrm{w}}, q, y), (I, q^{\mathrm{p}}, y)\}$$

Stage 2: Invariant Evidence Consolidation (IEC)
for $s = 1, \ldots, S_{\mathrm{IEC}}$ **do**
    Sample online weak views: $\widetilde{\mathcal{D}}_x^{(s)} \leftarrow \mathrm{WeakAug}(\mathcal{D}_x)$
    Update ϕ using the IEC objective: $\phi \leftarrow \mathrm{Update}\Big(\phi; \mathcal{L}_{\mathrm{IEC}}(\mathcal{D}_x, \widetilde{\mathcal{D}}_x^{(s)})\Big)$
end for
Stage 3: Counterfactual Evidence Calibration (CEC)
Retain the IEC adapter and reinitialize the optimizer state
for $s = 1, \ldots, S_{\mathrm{CEC}}$ **do**
    Construct all-black counterparts: $\mathcal{D}_x^{\mathrm{n}} \leftarrow \mathrm{BlackView}(\mathcal{D}_x)$
    Continue updating ϕ with the CEC objective: $\phi \leftarrow \mathrm{Update}(\phi; \mathcal{L}_{\mathrm{CEC}}(\mathcal{D}_x, \mathcal{D}_x^{\mathrm{n}}))$
end for
Generate the final answer from the original input: $\hat{y} \leftarrow \mathrm{GreedyDecode}_{\theta,\phi}(I, q)$
Discard the temporary adapter before the next instance
return $\hat{y}$

Interpretation and Empirical Support. The proposition bounds the objective-level discrepancy between black-image replacement and local evidence masking at a fixed evaluated checkpoint. A smaller PG therefore guarantees a smaller upper bound on the difference between the corresponding empirical calibration objectives. Table 3 computes PG using the same fixed QEE pseudo-answer for the paired visual conditions, directly connecting the measured discrepancy to the answer support used by CEC. Although CEC does not directly optimize PG, PG decreases from 0.80 after IEC to 0.32 after CEC on SmolVLM-2B and from 1.46 to 0.90 on Qwen-2B, yielding a smaller upper bound on the corresponding objective discrepancy for both backbones. The matched masking experiment provides a complementary behavioral comparison by evaluating whether model predictions are selectively sensitive to annotated answer-relevant regions. VEC-TTA increases ES relative to IEC-only on both backbones, while preserving substantially more accuracy under matched control masking than under evidence masking and improving accuracy on the original images. Together, the objective-discrepancy bound and the matched visual interventions support the black-image replacement as a practical annotation-free proxy for relative visual-support calibration.

Overall Algorithm. Algorithm 1 summarizes the complete instance-wise workflow of VEC-TTA, including pseudo-supervision construction with QEE, sequential LoRA adaptation through IEC and CEC, and final prediction on the original image–query pair.

Limitations. VEC-TTA relies on pseudo-supervision generated by the base model, so errors in the elicited answers may be reinforced during adaptation. Its black-image proxy removes visual content globally, leaving query-specific evidence removal approximate despite the objective-level agreement examined in our analysis. More localized interventions could further refine this calibration. Our qualitative analysis also identifies occasional cases in which improved local visual grounding does not fully translate into more accurate overall descriptions. Finally, our evaluation is limited to direct-answer settings, and the effectiveness of VEC-TTA on tasks requiring longer multimodal reasoning remains to be evaluated.