Test-Time Consistency in Vision Language Models

Shih-Han Chou*[1,2], Shivam Chandhok*[1,2], James J. Little[1], Leonid Sigal[1,2,3]
[1]University of British Columbia, Canada
[2]Vector Institute for AI, Canada [3]Canada CIFAR AI Chair, Canada
{shchou75, chshivam, little, lsigal}@cs.ubc.ca

Abstract

Vision-Language Models (VLMs) have achieved impressive performance across a wide range of multimodal tasks, yet they often exhibit inconsistent behavior when faced with semantically equivalent inputs—undermining their reliability and robustness. Recent benchmarks, such as MM-R^3, highlight that even state-of-the-art VLMs can produce divergent response across semantically equivalent inputs, despite maintaining high average accuracy. Prior work addresses this issue by modifying model architectures or conducting large-scale fine-tuning on curated datasets. In contrast, we propose a simple and effective test-time consistency framework that enhances semantic consistency without supervised re-training. Our method is entirely post-hoc, model-agnostic, and applicable to any VLM with access to its weights. Given a single test point, we enforce consistent predictions via two complementary objectives: (i) a Cross-Entropy Agreement Loss that aligns predictive distributions across semantically equivalent inputs, and (ii) a Pseudo-Label Consistency Loss that draws outputs toward a self-averaged consensus. Our method is plug-and-play, and leverages information from a single test-input itself to improve consistency. Experiments on the MM-R^3 benchmark show that our framework yields substantial gains in consistency across state-of-the-art models, establishing a new direction for inference-time adaptation in VLMs.[1]

1. Introduction

Vision-Language Models (VLMs) [9, 17, 18, 25] have achieved impressive performance across a wide range of multimodal tasks, including visual question answering [2], captioning [4, 16, 20], and reasoning [10, 28]. While existing evaluations predominantly focus on accuracy, a growing body of work highlights a critical shortcoming: *semantic inconsistency* [6]. That is, VLMs producing divergent outputs when prompted with semantically equivalent inputs—undermining their reliability, interpretability, and deployment in high-stakes settings.

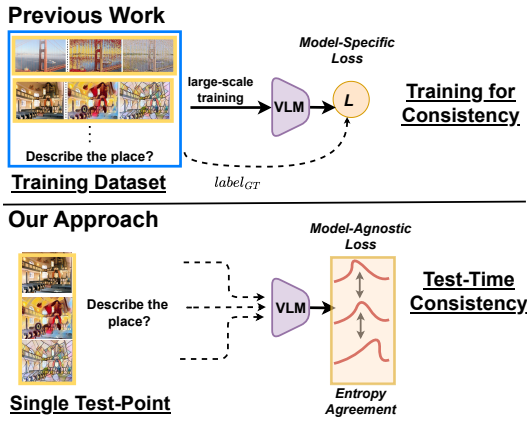

Figure 1. Comparison between **training-time** (top) and our **test-time** (bottom) consistency frameworks. While prior work (*e.g.*, [6]) needs large-scale supervised fine-tuning with curated dataset to enforce consistency, our method operates entirely post-hoc by adapting to a single test point with few gradient steps at test-time

Recent interest in investigating inconsistencies in VLM response has led to the development of dedicated benchmarks for evaluating this phenomenon. Notably, previous efforts such as MM-R^3[6] introduces a curated test suite that assesses VLM consistency under controlled semantics-preserving perturbations, including *question rephrasing*, *image restyling*, and *context masking*. Despite achieving high accuracy, state-of-the-art models exhibit significant variance across these conditions, highlighting that correctness does not imply semantic consistency—a key requirement for reliable multimodal reasoning. While previous approaches [6] enhance consistency via adapter-based large scale fine-tuning, they requires intrusive architectural modifications and access to a sizable curated training dataset.

We tackle the problem of improving consistency through a ***test-time, inference-only*** approach—without access to labels, training data, large-scale fine-tuning, or architectural modifications. Unlike prior methods that rely on supervised re-training or adapter insertion [6], our framework operates entirely post-hoc and leverages information within a single test sample to enhance consistency. This is particularly

*Equal Contribution. Order determined by coin flip.
[1]Code is available at https://github.com/ShihHanChou/testtime_const_VLM

valuable in scenarios where retraining is infeasible.

Approach: We propose a simple, general-purpose ***test-time consistency*** framework that can be applied to any probabilistic vision-language model (VLM) in a ***plug-and-play*** manner. Our method leverages semantically equivalent variants of a single test input and encourages agreement among the resulting predictions via two lightweight objectives: (1) a *Cross-Entropy Agreement Loss*, which penalizes divergence in response token distributions, and (2) a *Pseudo-Label Consistency Loss*, which aligns predictions toward a consensus output. Crucially, our framework departs from training-centric paradigms [6] by adapting model behavior post-hoc—even for a *single* test input—using only the rich signal present in that input's semantic variations. This enables the model to produce invariant predictions across linguistic and visual perturbations, thereby promoting semantic consistency and robust multimodal reasoning.

We evaluate our approach on the established MM-R^3 benchmark [6] and demonstrate substantial improvements in consistency across multiple open-source VLMs. Our method remains lightweight—requiring no architectural changes or large-scale re-training—and introduces only a small amount of inference-time overhead, limited to a few gradient steps on a single test sample. We believe this is a ***worthwhile trade-off*** given the efficiency it offers: it requires no task-specific training, adapts to new data distributions using just a single sample, and remains fully model-agnostic. Notably, even strong models benefit from this targeted adaptation, highlighting the value of test-time regularization. We advocate for consistency, alongside accuracy, to be a central design goal in the development of robust multimodal learning systems.

Contributions. We summarize contributions as follows:

- We introduce a simple, model-agnostic framework that addresses the underexplored problem of ***test-time consistency*** in VLMs. It operates entirely *post-hoc*, requiring only access to model weights and information from a *single test input*—with no need for training data, supervised re-training, or access to original loss functions—making it broadly applicable and deployment-friendly.
- Our framework combines two complementary objectives: (1) a *Cross-Entropy Agreement Loss* that minimizes divergence across perturbed input predictions, and (2) a *Pseudo-Label Consistency Loss* that aligns outputs toward a consensus response. Unlike prior approaches that rely on fine-tuning or adapter insertion, our method is fully ***plug-and-play***, requiring no architectural modifications or training-time assumptions.
- We demonstrate substantial improvements in semantic consistency across linguistic and visual perturbations on MM-R^3 consistency benchmark [6] and OKVQA [19].

2. Related Works

Consistency in Vision-Language Models. While existing evaluations of Vision-Language Models (VLMs) predominantly focus on accuracy, a growing body of work highlights a critical shortcoming: *semantic inconsistency* [6]. That is, VLMs often produce divergent outputs when prompted with semantically equivalent inputs—undermining their reliability, interpretability, and applicability in high-stakes settings. The recently established MM-R^3 consistency benchmark [6] systematically investigates this issue, introducing a suite of perturbation-based evaluations across rephrased questions, stylized images, and masked contexts. Their results show that even state-of-the-art VLMs exhibit significant inconsistency across these settings, despite high accuracy—revealing a fundamental gap between correctness and stable reasoning.

While prior efforts to improve consistency [6] typically focus on intrusive architectural modifications, or fine-tuning on curated data, these approaches are computationally intensive and often infeasible in practical settings. To this end, in contrast to previous work [6], we address this challenge from a test-time perspective, proposing a lightweight, post-hoc framework that improves consistency without large-scale retraining or access to labels.

Test-Time Adaptation. Test-time adaptation methods have evolved from entropy-based confidence maximization to more efficient and modular tuning strategies. MEMO [30] enhances robustness by enforcing confident and consistent predictions across augmented test-time views. Test-Time Prompt Tuning [22] adapts CLIP by optimizing prompts at inference to better match shifted input distributions. MedAdapter [21] guides pretrained LLMs toward domain-specific tasks via lightweight adapter updates, while LoRA-TTT [12] reduces adaptation cost by fine-tuning low-rank adapters. Karmanov et al. [11] propose a lightweight VLM adaptation scheme that freezes the core model and updates only a small projection head via entropy minimization. However, these methods primarily target contrastive VLMs like CLIP and are not readily applicable to generative VLMs such as Qwen, LLaVA etc..

In contrast, our work addresses a complementary and underexplored axis: *semantic consistency* in generative VLMs. Our approach operates entirely post-hoc, is model-agnostic, and leverages information from a single test input—requiring no retraining, architectural changes, or access to training data—making it lightweight, scalable, and broadly applicable.

Pseudo-Labeling and Self-Training. Pseudo-labeling has been widely used in semi-supervised learning [3, 15, 23, 27, 29], often paired with augmentations or confidence thresholds. In vision-language models, it has been employed to generate pseudo-captions [26], region–phrase alignments [5], and visual-language prototypes [1]. We

adopt a test-time variant of pseudo-labeling, aggregating model outputs across perturbed inputs into a self-consistent consensus, encouraging stability without requiring external supervision or retraining. Recently, TTRL frames inference on unlabeled data as RL with majority-vote rewards to self-evolve LLMs [32]. While it pursues label-free test-time adaptation, our method instead targets *semantic consistency* in *multimodal generative VLMs* rather than reasoning accuracy in LLMs, and uses lightweight cross-entropy and pseudo-label losses instead of reinforcement learning.

Entropy-Based Adaptation. Entropy minimization has been a foundational strategy for improving robustness under distribution shift. Grandvalet and Bengio [8] introduced it as a regularization objective for unlabeled data, and TENT [24] applied it for test-time adaptation by optimizing batch norm parameters. MEMO [30] improved on this by combining entropy minimization with multi-view consistency during inference. Extensions to large models include entropy-guided generation in LLMs [7, 13] and efficient test-time tuning for vision-language models by updating only lightweight projection heads [11]. Our method builds on this line of work by extending entropy-based objectives to open-ended generative multimodal VLMs—not to improve accuracy, but to enhance semantic consistency, an underexplored yet practically important aspect of reliability and interpretability in multimodal reasoning.

3. Approach

3.1. Problem Setting

We follow the procedure and settings defined in the MM-R^3 benchmark to evaluate consistency under diverse semantic variations. Given a test input $\mathbf{x} = (I, Q)$, the benchmark provides K semantically equivalent variants (I_k, Q_k) constructed via:

- *Question Rephrasing*: Paraphrased variants of Q generated using a language model keeps I fixed.
- *Image Restyling*: Stylized versions of I using neural style transfer (*e.g.*, Mosaic, Candy, Undie, and Grayscale) with Q not altered.
- *Context Reasoning*: Variants of I with different occlusions applied to a specific object region, while keeping Q once again fixed.

Each perturbed pair (I_k, Q_k) is passed through the VLM to obtain response distributions:

$$\mathbf{p}_k = \texttt{VLM}(I_k, Q_k), \quad \text{for } k = 1, \ldots, K \tag{1}$$

3.2. Method

Overview. We propose a lightweight, test-time strategy to improve the semantic consistency of Vision-Language Models (VLMs) by encouraging agreement across semantically equivalent variants of a single test input. Our method operates entirely post-hoc and leverages only the information present in the given test example. It performs a small number of inference-time updates (typically 1–4 steps), requiring no access to training data, ground-truth labels, or model internals.

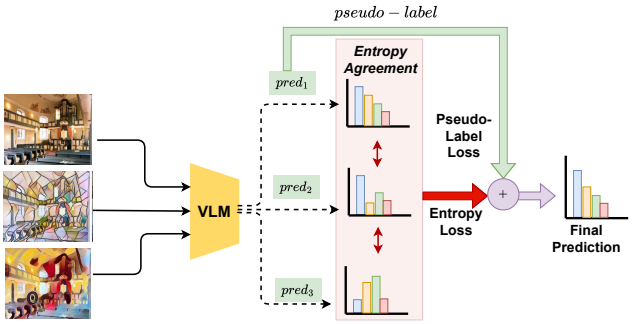

Figure 2. **Overview of our test-time consistency framework**. Given a test input with semantically equivalent input variants (*e.g.*, restyled images), we forward them through a pretrained VLM to obtain predictions. Two complementary objectives are used to improve consistency: (1) Cross-Entropy Agreement Loss, which aligns token-level output distributions across variants, and (2) Pseudo-Label Consistency Loss, which encourages agreement with a consensus pseudo-label. The model is updated with few (1-4) steps using gradients from these objectives, enabling consistent final predictions without access to training data or model internals.

Our approach combines two complementary objectives: (1) a Cross-Entropy Agreement Loss that aligns token-level output distributions across perturbed inputs, and (2) a Pseudo-Label Consistency Loss that enforces convergence toward a stable, consensus output prediction. These objectives guide the model to become more consistent at inference, without altering original architecture or parameters.

3.3. Cross-Entropy Agreement Loss

To promote consistency across semantically equivalent input variants, we introduce a Cross-Entropy Agreement Loss that aligns their token-level output distributions. Given a test input, we generate VLM output for K perturbed variants and obtain token-level logits through a forward pass.

Let $\mathbf{z}_k^j \in \mathbb{R}^V$ denote the logits over the vocabulary V at output token position j of the VLM response for the k-th input variant. Let L_k be the total number of valid output tokens in response for that variant. We compute the average logits across the decoded sequence for each variant k:

$$\bar{\mathbf{z}}_k = \frac{1}{L_k} \sum_{j=1}^{L_k} \mathbf{z}_k^j \tag{2}$$

We then apply softmax to obtain the normalized token distribution:

$$\mathbf{p}_k = \text{softmax}(\bar{\mathbf{z}}_k) \tag{3}$$

The agreement loss is defined as the average of all pairwise symmetric cross-entropies across the K output distri-

butions:

$$\mathcal{L}_{\text{CE}} = \frac{2}{K(K-1)} \sum_{i<j} \text{CE}(\mathbf{p}_i, \mathbf{p}_j) + \text{CE}(\mathbf{p}_j, \mathbf{p}_i) \quad (4)$$

This loss encourages alignment of the global output tokens across K input variants while ensuring the model's generation is *distributionally consistent*, even if wording or phrasing changes.

3.4. Pseudo-Label Consistency Loss

To complement distributional alignment, we introduce a Pseudo-Label Consistency Loss that enforces consistency at the output level by aligning each variant's predicted sequence to a common consensus output prediction.

Let $\{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ be the decoded textual outputs from the K semantically equivalent input variants, generated using greedy decoding. To compute a consensus label, we define a string similarity function $\text{sim}(\cdot, \cdot)$ based on normalized Levenshtein distance (*e.g.*, token set ratio). We cluster the K output responses by assigning two responses $\mathbf{y}_i$ and $\mathbf{y}_j$ to the same cluster if

$$\text{sim}(\mathbf{y}_i, \mathbf{y}_j) \geq \tau, \quad (5)$$

where $\tau \in [0, 1]$ is a fixed similarity threshold (*i.e.*, $\tau = 0.85$). Among all clusters, we identify the largest one, and from within it, select the most frequent response as the pseudo-label:

$$\hat{\mathbf{y}}_{\text{pseudo}} = \text{mode}\left(\mathcal{C}_{\max}\right), \quad (6)$$

where $\mathcal{C}_{\max}$ is the largest similarity-based cluster.

We then tokenize $\hat{\mathbf{y}}_{\text{pseudo}}$ and use it as the supervision target for all K variants. Let $\mathbf{p}_k$ denote the token-level predicted distribution from variant k (*i.e.*, the model's output logits after softmax). The *Pseudo-Label Consistency Loss* is defined as:

$$\mathcal{L}_{\text{PL}} = \frac{1}{K} \sum_{k=1}^{K} \text{CE}(\hat{\mathbf{y}}_{\text{pseudo}}, \mathbf{p}_k), \quad (7)$$

where $\text{CE}(\cdot, \cdot)$ denotes the cross-entropy loss between pseudo-label tokens and predicted distribution. This loss encourages all variants to converge to the dominant semantic response, enhancing answer-level consistency of perturbed inputs.

Complementarity of Losses. The Cross-Entropy Agreement Loss encourages *token-level alignment* by smoothing output distributions across input variants, while the Pseudo-Label Consistency Loss enforces *prediction-level convergence* by aligning decoded outputs with a dominant consensus response. Together, these losses regularize both the internal generation process and final output, yielding improved semantic consistency at test time using only the information in single-test point without modifying the underlying model.

3.5. Final Objective and Inference

Given a test input with K semantically equivalent variants (I_k, Q_k) (*e.g.*, via question rephrasing, image restyling, or context masking), we adapt the model using gradients from two complementary objectives: the Cross-Entropy Agreement Loss $\mathcal{L}_{\text{CE}}$ and the Pseudo-Label Consistency Loss $\mathcal{L}_{\text{PL}}$. The total loss at each update step is computed as a weighted sum:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{PL}}, \quad (8)$$

where α and β are hyperparameters balancing distributional agreement and semantic convergence. We optimize this objective for a small number of gradient-based updates—typically between 1 and 4—using only the current test example, without access to any labeled (training) data.

Adaptive Step Selection Different test inputs may benefit from different numbers of adaptation steps—while some improve with a few updates, others may degrade due to over-adaptation. To address this, we introduce an adaptive mechanism that dynamically selects the optimal number of steps for each test point.

After each update step $t \in \{0, 1, \ldots, T\}$, we decode the model's output responses for the K input variants. To assess internal consistency, we compute the average pairwise token-set similarity—based on normalized Levenshtein distance—among the K decoded answers and the previously generated pseudo-label (used in the Pseudo-Label Consistency Loss). The similarity score for step t is given by:

$$\text{score}_t = \frac{1}{K(K-1)} \sum_{i<j} \text{sim}(a_i^t, a_j^t), \quad (9)$$

where a_i^t and a_j^t denote either one of the K decoded answers or the shared pseudo-label. The step t^* with the highest score is selected as the final output, reflecting the most consistent model behavior during adaptation.

$$t^* = \arg\max_t \text{score}_t. \quad (10)$$

This selection mechanism is fully unsupervised and relies solely on model outputs *without using ground-truth annotations*. It enables per-instance, test-time adaptation that is both robust and efficient, ensuring that predictions remain semantically consistent while avoiding over-updating.

Method Variants. We report results for two variants of our method. In the first, we use a fixed number of adaptation steps ($T = 2$) for all samples, which we refer to as **Test-time (constant T)**. In the second, we use the adaptive step selection mechanism described above to dynamically choose the optimal number of updates per input. We refer to this variant as **Test-time (adapt. T)**. This comparison allows us to assess the trade-offs between simplicity and input-specific adaptivity.

4. Experiments

Dataset. We evaluate our method on the standard MM-R^3 consistency benchmark [6], which provides a test suit for evaluating VLM consistency with three distinct tasks, spanning variations across both visual and linguistic modalities: **Question Rephrasing** (lingual perturbation), **Image Restyling** (semantics-preserving visual perturbation), and **Context Reasoning** (occlusion-based visual perturbation). The *question rephrasing* task assesses whether VLMs produce consistent answers to semantically equivalent questions phrased differently. The *image restyling* task evaluates consistency under visual domain shifts by presenting stylized versions of the image. The *context reasoning* task tests the model's ability to reason under partial occlusion. Our evaluations are conducted on MM-R^3 test set.

Evaluation Settings. In line with prior test-time adaptation work, our default setup assumes access to a single test input (*i.e.*, image, question and its K semantic variants, as provided by MM-R^3), on which adaptation is performed. However, to further ensure fairness and rule out potential benchmark coupling, we additionally evaluate a stricter setting where adaptation uses a *separate set of variants* (generated via independent paraphrasing and restyling), while evaluation is conducted solely on the MM-R^3 benchmark variants. We call this setting **(Disjoint Variant Setup)**.

Models. We evaluate our method on widely used state-of-the-art open-source Vision-Language Models (VLMs). Specifically, LLaVA 1.5M [17] (`llava-v1.5-7b`), LLaVA-Next [18] (`llava-v1.6-mistral-7b`), Qwen2-VL [25] (`Qwen2-VL-7B-Instruct`), InternVL3 [31] (`InternVL3`), and Idefics2 [14] (`idefics2-8b`). Our choice of these models is motivated by the fact that these models are widely considered state-of-art, used as foundations for downstream applications and frequently serve as initialization points for developing more advanced VLMs. Since our method involves modifying model parameters at test time, we restrict our evaluation to open-source models and exclude proprietary systems such as GPT-4V or Gemini. Evaluating on these representative models enables us to assess the generality, practical utility, and broader impact of test-time consistency improvements across VLMs.

Implementation Details. Please see Supplemantary

Evaluation Metrics. Since VLM responses are open-ended and linguistically diverse, we adopt evaluation metrics similar to those introduced in MM-R^3 [6], in order to capture both correctness and consistency. We briefly introduce the core evaluation metrics used to assess correctness and consistency; full metric definitions and implementation details are provided in the supplementary material.

- **Accuracy (Acc):** Measures correctness using fuzzy string matching, accounting for minor lexical variations. A similarity threshold of 85 is used to determine a match.
- **Similarity with Ground Truth ($S_{\mathbf{GT}}$):** Computes semantic similarity between the model's response and the reference answer using BERT sentence embeddings, offering a more flexible alternative to exact match.
- **Consistency Accuracy (Con):** Evaluates semantic agreement across responses to semantically equivalent inputs. Responses are considered consistent if their pairwise similarity exceeds a threshold of 0.7.
- **Consistency Similarity ($S_{\mathbf{C}}$):** Computes the average pairwise similarity across all response variations, providing a smoother measure of output invariance.
- **Overall Score (O_{all}):** The harmonic mean of correctness and consistency metrics.

$$H_{mean}(mean(\mathbf{Acc}, \mathbf{S_{GT}}), mean(\mathbf{Con}, \mathbf{S_C})). \quad (11)$$

We use the harmonic mean to emphasize models that are balanced in both accuracy and consistency, as it penalizes performance when either component is low. This provides a unified measure of overall model quality.

4.1. Main Results

Overview. Table 1 presents the performance of our test-time consistency framework across three tasks in the MM-R^3 benchmark. We report results for each base model with two variants: Test-time (constant T) and Test-time (adaptive T). Across all tasks, our method consistently improves semantic consistency and overall performance, with the adaptive variant yielding the best results.

Question Rephrasing. In the rephrasing task, our adaptive test-time method yields substantial gains in consistency and overall score across all three models while preserving accuracy. For instance, on LLaVA-1.5M, O_{all} improves from 52.73 (base) to 64.08, with consistency rising from 48.55 to 79.11 and **Acc** increasing from 36.18 to 39.58. LLaVA-Next, Qwen2-VL, and InternVL3 also show notable gains, with the adaptive variant achieving the best O_{all} for each model: 68.83 and 85.33, respectively. This validates the ability of our method to enforce semantic invariance across linguistic perturbations without reducing accuracy.

Image Restyling. This task poses a significant domain shift challenge due to stylized visual inputs. Our method leads to especially large improvements in consistency for all models. On LLaVA-Next, consistency improves from 55.34 (base) to 91.85 (Test-time) and further to 91.85 (Adaptive), with O_{all} reaching 46.00. Qwen2-VL sees the highest performance overall, with the adaptive variant achieving O_{all} = 51.20 and nearly perfect consistency (99.14). These results demonstrate the robustness of our framework under visual perturbations.

Context Reasoning. Our approach also improves model behavior in the context reasoning task, which requires stable answers under partial information. Our method delivers

Table 1. **Overall results.** We highlight our approach in orange color and the overall results in gray color. The best-performing method is in bold for each models.

	Models	Acc	S_{GT}	Con	S_C	O_{all}
Question Rephrasing	LLaVa 1.5M	36.18	62.96	48.55	64.10	52.73
	+ Constant T	38.00	65.05	77.67	84.65	63.03
	+ Adapt. T	39.58	65.10	79.11	86.10	**64.08**
	LLaVa-Next	42.89	64.89	49.18	65.69	55.61
	+ Constant T	44.48	68.74	83.39	88.47	68.25
	+ Adapt. T	44.74	68.67	85.18	89.92	**68.83**
	Qwen2-VL	66.72	79.69	65.78	76.16	72.07
	+ Constant T	70.79	82.66	90.44	93.52	83.66
	+ Adapt. T	72.14	83.27	93.6	95.64	**85.33**
	InternVL3	43.51	64.52	46.93	63.20	54.14
	+ Constant T	44.69	66.41	71.31	80.11	63.82
	+ Adapt. T	44.30	65.45	76.86	84.10	**65.41**
	Idefics2	53.21	71.04	64.5	74.8	65.68
	+ Constant T	53.35	71.12	64.47	74.85	65.74
	+ Adapt. T	54.13	71.5	66.0	75.88	**66.64**
Image Restyling	LLaVa 1.5M	9.61	34.85	18.96	56.91	28.03
	+ Constant T	12.09	35.62	20.14	59.01	29.77
	+ Adapt. T	17.94	40.15	33.90	64.46	**36.52**
	LLaVa-Next	17.57	41.47	55.34	71.36	40.27
	+ Constant T	18.99	42.49	88.25	91.25	45.80
	+ Adapt. T	18.71	42.52	91.85	93.16	**46.00**
	Qwen2-VL	21.13	39.25	61.67	75.85	41.96
	+ Constant T	22.60	42.32	98.30	98.97	48.85
	+ Adapt. T	22.58	46.40	99.14	99.45	**51.20**
	InternVL3	9.96	31.10	50.63	67.16	30.44
	+ Constant T	10.75	31.44	57.00	70.66	**31.71**
	+ Adapt. T	10.75	31.44	57.00	70.66	**31.71**
	Idefics2	17.23	42.15	66.31	81.02	42.32
	+ Constant T	17.84	42.42	67.46	81.6	42.91
	+ Adapt. T	18.91	43.05	66.35	81.02	**43.62**
Context Reasoning	LLaVa 1.5M	16.11	42.69	65.64	75.08	41.47
	+ Constant T	22.88	49.49	88.89	93.45	51.81
	+ Adapt. T	31.04	55.14	72.11	81.90	**55.26**
	LLaVa-Next	30.24	27.43	32.11	58.44	35.23
	+ Constant T	32.50	50.84	89.91	90.16	56.97
	+ Adapt. T	32.29	53.85	95.24	96.66	**59.45**
	Qwen2-VL	29.09	40.03	34.58	53.70	38.77
	+ Constant T	29.60	50.11	91.17	91.75	55.52
	+ Adapt. T	30.42	53.00	99.53	99.66	**58.80**
	InternVL3	18.82	44.83	90.08	93.21	47.24
	+ Constant T	18.71	44.76	87.07	91.21	46.81
	+ Adapt. T	18.80	44.83	91.28	94.10	**47.37**
	Idefics2	25.82	49.69	50.56	66.05	45.83
	+ Constant T	25.77	49.65	51.17	66.47	45.96
	+ Adapt. T	25.77	49.65	51.95	67.01	**46.16**

Table 2. **Results for Disjoint Variant Setup**

	Models	Acc	S_{GT}	Con	S_C	O_{all}
Question Rephrasing	LLaVa-Next	42.89	64.89	49.18	65.69	55.61
	+ Constant T	43.37	66.03	77.81	84.54	**63.81**
	+ Adapt. T	42.67	65.44	73.54	81.71	63.73
	Qwen2-VL	66.72	79.69	65.78	76.16	72.07
	+ Constant T	73.37	83.32	84.67	89.48	**82.48**
	+ Adapt. T	71.92	82.37	81.12	87.04	80.46
	InternVL3	43.51	64.52	46.93	63.20	54.14
	+ Constant T	46.65	68.21	68.20	77.6	**63.66**
	+ Adapt. T	46.08	67.56	64.87	75.32	62.76
Image Restyling	LLaVa-Next	17.57	41.47	55.34	71.36	40.27
	+ Constant T	17.47	41.51	86.43	89.56	**44.18**
	+ Adapt. T	17.62	41.48	83.11	87.34	43.88
	Qwen2-VL	21.13	39.25	61.67	75.85	41.96
	+ Constant T	21.32	45.89	98.12	98.80	**50.11**
	+ Adapt. T	21.39	45.96	96.02	97.26	49.95
	InternVL3	9.96	31.10	50.63	67.16	30.44
	+ Constant T	11.01	30.98	54.88	69.47	**31.39**
	+ Adapt. T	10.91	31.24	51.44	67.78	31.14

Table 3. **Comparison of our approach with supervised fine-tuned model on LLaVa 1.5M model.**

Models	Acc	S_{GT}	Con	S_C	O_{all}
Question Rephrasing					
LLaVa 1.5M	36.18	62.96	48.55	64.1	52.73
+ Finetuning [6]	42.55	69.03	63.79	75.83	62.02
+ Adapt. T	39.58	65.10	79.11	86.10	**64.08**
Image Restyling					
LLaVa 1.5M	9.61	34.85	18.96	56.91	28.03
+ Finetuning [6]	25.45	50.67	50.94	66.06	**46.11**
+ Adapt. T	17.94	40.15	33.90	64.46	36.52
Context Reasoning					
LLaVa 1.5M	16.11	42.69	65.64	75.08	41.47
+ Finetuning [6]	63.93	76.62	75.00	83.91	**74.58**
+ Adapt. T	31.04	55.14	72.11	81.9	55.26

Table 4. **Ablation Studies on contribution of different loss functions we use in our approach**

	$\mathcal{L}_{CE}$	$\mathcal{L}_{PL}$	Acc	S_{GT}	Con	S_C	O_{all}
Question Rephrasing			61.44	69.71	52.29	66.86	62.43
	✓		59.48	71.70	52.94	66.36	62.48
		✓	66.67	76.21	85.62	88.90	78.56
	✓	✓	66.01	77.18	90.20	93.10	**80.39**
Image Restyling			14.16	38.36	54.33	70.77	36.99
	✓		16.12	39.86	61.86	74.10	39.65
		✓	17.93	40.73	83.97	89.86	43.86
	✓	✓	19.25	40.35	84.94	90.40	**44.48**
Context Reasoning			32.68	27.23	31.37	55.81	35.51
	✓		28.10	52.19	55.56	71.80	49.25
		✓	32.77	56.13	97.14	97.1	60.99
	✓	✓	33.33	55.76	98.69	99.26	**61.44**

Table 5. **Hyper-parameter search on LLaVa-Next.**

	α	β	Acc	S_{GT}	Con	S_C	O_{all}
Question Rephrasing	0.1	1	66.01	76.38	86.27	89.81	78.73
	0.5	1	66.01	77.18	90.20	93.10	**80.39**
	1	1	64.71	75.63	83.00	87.82	77.04
	1	0.5	64.71	75.68	83.01	87.88	77.07
	1	0.1	65.36	75.61	79.08	84.83	75.79
Image Restyling	0.1	1	17.91	40.23	86.22	91.01	43.78
	0.5	1	19.25	40.35	84.94	90.4	**44.48**
	1	1	17.93	40.28	85.24	90.13	43.70
	1	0.5	17.93	40.28	85.26	90.47	43.73
	1	0.1	17.84	40.60	82.37	87.19	43.46
Context Reasoning	0.1	1	32.68	55.32	97.39	98.11	60.68
	0.5	1	33.33	55.76	98.69	99.26	**61.44**
	1	1	33.33	55.69	97.39	98.30	61.19
	1	0.1	32.68	55.17	96.08	97.25	60.4
	1	0.5	32.68	55.21	94.77	96.30	60.20

both higher consistency and improved accuracy. Specifically, LLaVA-1.5M shows a dramatic gain in O_{all} from 41.47 to 55.26 (adaptive), while LLaVA-Next reaches the highest score of 59.45. Interestingly, even though Qwen2-VL starts from a stronger baseline, our method boosts its O_{all} to 58.80 and consistency to 99.53. These results suggest that test-time consistency not only stabilizes outputs but also improves factual grounding under ambiguity.

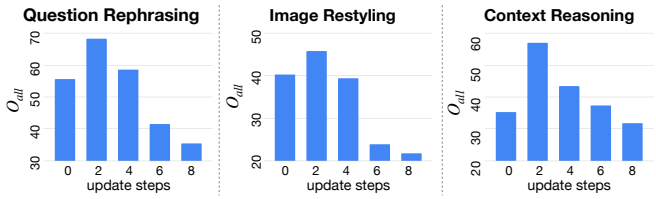

Figure 3. Effect of different number of update steps for each task.

4.1.1. Disjoint Variant Setup.

Beyond the standard setting where adaptation and evaluation use the benchmark-provided variants, we also evaluate a stricter setup in which the model adapts on independently generated variants (paraphrased questions and restyled images) and is evaluated solely on the original MM-R^3 variants. Results in Table 2 show that our method continues to yield substantial improvements under this setup (e.g., Qwen2-VL Rephrasing O_{all}: 72.07 $\rightarrow$ **82.48**; Restyling: 41.96 $\rightarrow$ **50.11**), confirming that the observed consistency gains are robust and not tied to overlap between adaptation and evaluation variants.

4.2. Comparison with Full Fine-Tuning

To contextualize the effectiveness of our approach, we compare it against the fully fine-tuned model from MM-R^3 [6], which retrains LLaVA-1.5M through large-scale supervised training using task-specific data from the curated MM-R^3 training set. In contrast, our method adapts the model using only a single test point and two test-time gradient steps, without access to labeled data, or large-scale training.

Table 3 presents the results on three MM-R^3 tasks. Despite being significantly lighter in terms of computational cost and supervision, our method achieves competitive—and in some cases superior—performance compared to full fine-tuning. Specifically, on the *Question Rephrasing* task, it achieves an O_{all} score of 64.08, outperforming the fine-tuned model (62.02) by a notable margin.

On *Context Reasoning*, although full fine-tuning achieves the highest score (74.58), our method still improves substantially over the base model (55.26 vs. 41.47), again without any retraining. For *Image Restyling*, our method narrows the gap considerably (36.52 vs. 46.11), demonstrating strong robustness to visual perturbations even without additional training data. Notably, our method underperforms on these tasks in overall score because full fine-tuning jointly learns the novel task (unsupported by the base VLM) through curated training dataset and improves consistency. It can be seen that the performance of our approach on consistency (i.e Con) is nearly equivalent to that of full-finetuning, while on accuracy the improvement drops. This is not surprising as [6] learns from voluminous training data, which our model is not designed to do being a test-time approach.

4.3. Ablation Study

All experiments in the ablation studies are performed on the LLaVA-Next model, unless specified otherwise.

4.3.1. Contribution of each Objective

To understand the contribution of each component in our test-time consistency framework, we perform an ablation study by selectively enabling the Cross-Entropy Agreement Loss ($\mathcal{L}_{\text{CE}}$) and the Pseudo-Label Consistency Loss ($\mathcal{L}_{\text{PL}}$). Table 4 reports results across all three MM-R^3 tasks.

Complementary Benefits. We observe that both losses independently contribute to improving consistency and overall performance. Applying only $\mathcal{L}_{\text{CE}}$ improves consistency over the base model in all tasks, while $\mathcal{L}_{\text{PL}}$ alone often yields stronger gains in Acc.

Best Results with Combined Loss. The full method—using both $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{PL}}$—achieves the highest overall performance across all tasks. For instance, in the question rephrasing task, the combination yields $O_{all} = 80.39$ and consistency of 90.20, outperforming both individual losses. Similar trends are observed in image restyling and context reasoning, where the joint objective achieves the best O_{all} scores of 44.48 and 61.44, respectively. These results show the complementary roles of two losses: $\mathcal{L}_{\text{CE}}$ promotes token-level alignment of outputs across input perturbations, while $\mathcal{L}_{\text{PL}}$ anchors model predictions to a consensus output.

4.3.2. Ablation on Loss Weighting Coefficients

We ablate the loss weighting coefficients α and β in our total loss $\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{PL}}$, using LLaVA-Next across the MM-R^3 tasks. Results in Table 5 show that our method is robust to a range of settings, but performance is highest when both objectives are appropriately balanced.

The best results are obtained with $\alpha = 0.5$ and $\beta = 1$, yielding top O_{all} scores across all tasks: 80.39 (Rephrasing), 44.48 (Restyling), and 61.44 (Reasoning). Performance drops slightly when either loss dominates—for example, using $\beta = 0.1$ reduces consistency and overall score.

4.3.3. Ablation on Number of Updated Steps.

Figure 3 shows the impact of varying the number of gradient update steps (T) in our **Test-time (constant T)** variant, where a fixed number of updates is applied to all test inputs. We observe that performance improves initially but degrades beyond a certain point, revealing a trade-off between effective adaptation and overfitting. Across all three tasks, setting $T = 2$ yields the best score (O_{all}).The performance drop beyond $T = 2$ is most pronounced in the *Question Rephrasing* and *Context Reasoning* tasks, likely due to over-adaptation and overfitting on linguistic variations or ambiguous inputs.

This ablation is specific to the **Test-time (constant T)** setup. Our alternative variant, **Test-time (adapt. T)**, auto-

Table 6. **Results on original OKVQA dataset task.**

Acc	LLaVA 1.5M	LLaVA-Next	Qwen2-VL
Original	55.09	54.69	54.13
+ Constant T	53.98	56.10	58.61

matically selects the optimal number of updates per instance using the adaptive step selection mechanism described in Section 3.5. As such, it does not require manual tuning or per-task sensitivity analysis. Together, these two variants allow us to assess the trade-offs between simplicity and input-specific adaptability.

4.3.4. Does Adaptation Preservation Original Base Model Capabilities?

To ensure that our test-time consistency framework does not degrade the model's original capabilities, we evaluate performance on the unperturbed OKVQA dataset [19] before and after adaptation. For this experiment, we generate three semantically equivalent rephrasings of each original question using GPT-4V. These rephrasings are used during adaptation, while the final evaluation is performed on the original (unmodified) question from OKVQA dataset.

Results are shown in Table 6. Both LLaVA-Next and Qwen2-VL improve in accuracy on original unperturbed input after test-time adaptation—rising from 54.69 to 56.10 and from 54.13 to 58.61, respectively. This indicates that our method not only preserves but can even enhance model performance on standard benchmarks. LLaVA 1.5M shows a minor drop (55.09 $\rightarrow$ 53.98), suggesting slightly higher sensitivity in smaller models. Overall, these results show that our approach does not degrade on the original task distribution, and instead enables consistency improvements.

4.3.5. Effect of Decoding Temperature.

Please see the supplementary material for more details.

4.3.6. Qualitative Results.

We show qualitative results for the three tasks in Figure 4. Across all three tasks, our method is able to improve consistency with just two gradient steps on a single test-input at inference (as also supported in Table 1). Please see the supplementary material for more results.

5. Conclusion.

We present a simple yet effective *test-time consistency* framework for vision–language models that requires no access to curated training data, model internals, or supervised fine-tuning. By leveraging semantically equivalent variants of each input and enforcing agreement through two lightweight losses, our method seamlessly adapts VLMs at inference-time using inherent information in single test-input. Experiments on the MM-R^3 benchmark show that our approach significantly improves consistency while preserving or enhancing accuracy. We advocate for consistency as a core evaluation criterion for building reliable, real-world VLM systems in future work.

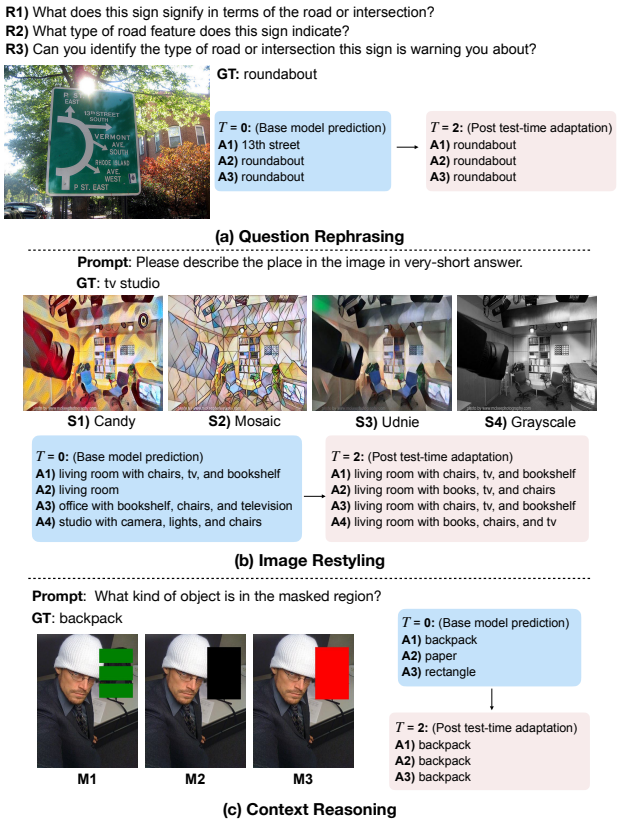

Figure 4. **Qualitative results on the three tasks.** Across three tasks, our method can improve the consistency even when the base model predictions (T=0) are inconsistent. More examples are shown in the supplementary material.

Limitations Our analysis is limited by the scope of the MM-R^3 dataset and its predefined perturbations, which may not fully capture the diversity of real-world consistency challenges. While our method improves consistency without access to training data or model internals, it requires multiple forward and backward passes per test input, which increases inference-time latency. However, it remains significantly more efficient and scalable overall compared to full fine-tuning, as it avoids large-scale training and need for supervision. Additionally, since adaptation is performed locally on a single test point, it may not correct broader model deficiencies or systematic biases, and different inputs could in principle lead to different local optima. Finally, because our approach updates model parameters during inference, it may not be suitable for deployment in strictly frozen or closed-source model environments.

Acknowledgments. This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs, NSERC Canada Research Chair (CRC), and NSERC Discovery Grants. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, the Digital Research Alliance of Canada[2], companies sponsoring the Vector Institute, and Advanced Research Computing at the University of British Columbia. Additional hardware support was provided by John R. Evans Leaders Fund CFI grant and Compute Canada under the Resource Allocation Competition award.

References

[1] Eman Ali, Sathira Silva, and Muhammad Haris Khan. Dpa: Dual prototypes alignment for unsupervised adaptation of vision-language models. In *WACV*, 2025. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019. 2

[4] 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*, 2015. 1

[5] Shih-Han Chou, Zicong Fan, James J Little, and Leonid Sigal. Semi-supervised grounding alignment for multi-modal feature learning. In *CRV*, 2022. 2

[6] Shih-Han Chou, Shivam Chandhok, James J Little, and Leonid Sigal. Mm-r^3: On (in-) consistency of multi-modal large language models (mllms). *ArXiv*, 2024. 1, 2, 5, 6, 7

[7] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. 3

[8] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 2004. 3

[9] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *ArXiv*, 2024. 1

[10] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[11] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, 2024. 2, 3

[12] Yuto Kojima, Jiarui Xu, Xueyan Zou, and Xiaolong Wang. Lora-ttt: Low-rank test-time training for vision-language models. *ArXiv*, 2025. 2

[13] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv*, 2023. 3

[14] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *NeurIPS*, 2024. 5

[15] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 2

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1, 5

[18] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 5

[19] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2, 8

[20] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1

[21] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. MedAdapter: Efficient test-time adaptation of large language models towards medical reasoning. In *EMNLP*, 2024. 2

[22] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 2022. 2

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020. 2

[24] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, 2024. 1, 5

[26] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *CVPR*, 2022. 2

[27] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, 1995. 2

[28] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 1

[2] https://vectorinstitute.ai/#partners

[29] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 2021. 2

[30] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *NeurIPS*, 2022. 2, 3

[31] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, 2025. 5

[32] Yuxin Zuo and et al. Ttrl: Test-time reinforcement learning. *ArXiv*, 2025. 3