This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TTRV: Test-Time Reinforcement Learning for Vision Language Models

Akshit Singh[1] **Shyam Marjit**[2] **Wei Lin**[3] **Paul Gavrikov**[1]
Serena Yeung-Levy[4] **Hilde Kuehne**[5,6] **Rogerio Feris**[6] **Sivan Doveh**[4]
James Glass[7] **M. Jehanzeb Mirza**[7]

[1]Independent Researcher [2]IISc Bangalore [3]JKU Linz [4]Stanford
[5]Tübingen AI Center [6]MIT-IBM Watson AI Lab [7]MIT CSAIL

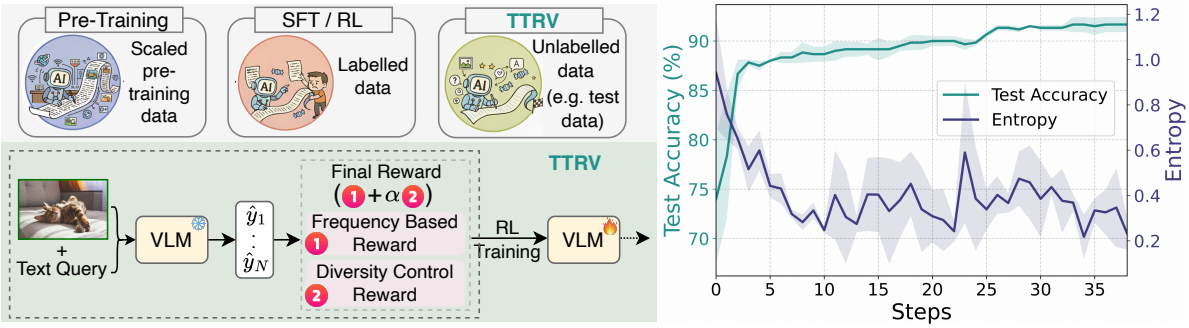


Figure 1. **Test-Time RL for VLMs.** (left) Unlike prior methods that require pre-training splits and post-training via Supervised Finetuning (SFT) or Reinforcement Learning (RL), our approach extracts reward signals directly at test time from unlabeled data. The reward combines (1) frequency-based signals and (2) diversity control, allowing the model to adapt online and improve downstream vision performance without any labeled data. (right) Test accuracy increases while entropy of the output logits decreases, showing that the model becomes more accurate and less uncertain as test-time RL progresses. The solid lines represent the mean, and shaded regions represent the variance of results obtained over 5 independent runs. The dataset is Resics45 [7], task is object recognition, and the model is InternVL-3-2B [73].

Abstract

Existing methods for extracting reward signals in Reinforcement Learning typically rely on labeled data and dedicated training splits, a setup that contrasts with how humans learn directly from their environment. In this work, we propose TTRV to enhance vision–language understanding by adapting the model on-the-fly at inference time, without the need for any labeled data. Concretely, we enhance the Group Relative Policy Optimization (GRPO) framework by designing rewards based on the frequency of the base model's output, while inferring on each test sample multiple times. Further, we also propose to control the diversity of the model's output by simultaneously rewarding the model for obtaining low entropy of the output empirical distribution. Our approach delivers consistent gains across both object recognition and visual question answering (VQA), with improvements of up to 52.4% *and* 29.8%*, respectively, and average boosts of* 24.6% *and* 10.0% *across* 16 *datasets. Remarkably, on image recognition, TTRV applied to* `Intern-VL-8B` *surpasses GPT-4o by an average of* 2.3% *over* 8 *benchmarks, while remaining highly competitive on VQA, demonstrating that test-time reinforcement learning can match or exceed the strongest proprietary models. Finally, we find many interesting properties of test-time RL for VLMs: for example, even in extremely data-constrained scenarios, where adaptation is performed on a single randomly chosen unlabeled test example, TTRV still yields non-trivial improvements of up to* 5.5% *in recognition tasks.*

Correspondence: akshitsingh21112002@gmail.com

1. Introduction

Recent advances in vision–language models (VLMs) [24, 30, 40, 44, 62, 66] have enabled impressive progress on tasks such as object recognition [10, 39] and visual question answering [14, 63]. Yet, unlike humans, who continuously refine their reasoning by interacting with the world and adapting to ambiguous, unlabeled experiences, current VLMs remain largely static once trained. Adaptation typ-

ically requires large amounts of annotated data and costly fine-tuning, limiting their ability to cope with new domains or unseen tasks.

Reinforcement Learning (RL) has shown promise for improving reasoning in Large Language Models (LLMs) [51] and Vision–Language Models (VLMs) [65], and has emerged as an effective post-training method for enhancing task-specific performance. However, most existing approaches still rely on reward signals derived from human-labeled data and remain restricted to curated training splits, which are misaligned with real-world scenarios where train–test distinctions do not naturally exist. This dependence raises a fundamental question:

> *If RL is to embody true learning from experience, should it not arise directly from interaction with unlabeled data in the wild, rather than from curated benchmarks?*

In this work, we move toward this vision by proposing a **T**est-**T**ime **R**einforcement Learning framework for **V**ision Language Models (TTRV) that learns directly from unlabeled test data. Our TTRV extracts reward signals for Group Relative Policy Optimization (GRPO) [51] directly on the test data, as it is encountered. Specifically, our proposed reward formulation consists of two distinct parts, based on frequency and diversity control of the pre-trained model's output for each test sample. The intuition is to encourage the model to frequently produce similar outputs for each test sample and reward the predictions of the model which are more frequent and at the same time, control the diversity of model's output by rewarding lower entropy of the output empirical distribution. An overview of our work, along with one optimization trajectory, is provided in Figure 1. This approach transforms static pretrained VLMs into dynamic systems capable of self-improvement at inference time, bringing RL for multimodal models closer to the human-like paradigm of learning through raw experience.

We extensively evaluate our TTRV across 16 datasets spanning two tasks: image recognition and visual question answering (VQA). These datasets cover a diverse range of domains, including fine-grained recognition, math reasoning, and general VQA. Our results show that TTRV consistently improves performance, generalizes across model families, and is remarkably data-efficient. For instance, when post-training the InternVL3 [6] model on only 20 randomly sampled test images, GRPO achieves gains of up to 52.4% (42.3% on large-scale ImageNet [10]). Similarly, on VQA benchmarks, TTRV boosts performance by as much as 28.0% on AI2D [22]. Remarkably, our TTRV outperforms one of the strongest proprietary models (GPT-4o) by 2.3% on average (over 8 datasets) for image classification, while remaining highly competitive for VQA. Beyond these gains, our ablation studies uncover several interesting properties of GRPO for vision–language understanding. Notably, GRPO improves cross-dataset generalization: training on one dataset can yield strong gains on a completely unrelated dataset. Further, even in extremely data-scarce scenarios, GRPO remains effective, achieving up to 5.5% improvement from rewards extracted on a single randomly chosen example. These findings suggest that GRPO does not simply adapt to a dataset's distribution but instead activates latent capabilities already learned during large-scale pretraining.

To conclude, we summarize the contributions of our work as follows:

- We introduce the first test-time reinforcement learning framework for vision–language models, which can be bootstrapped to any pre-trained VLM. Powered by carefully designed reward formulations, our method adapts models on the fly without requiring supervised data, thus realizing the true promise of RL.
- Through extensive experiments on 16 diverse benchmarks, we demonstrate that our proposed TTRV provides consistent and substantial improvements across tasks, model families, and domains.
- Our ablation studies further uncover novel properties of GRPO for VLMs, such as effectiveness in extremely low-data regimes and cross-dataset generalization, opening up new directions for future research in test-time adaptation with reward-driven learning.

2. Related Work

Our work is closely related to vision-language models and works that study RL-based fine-tuning and test-time training (TTT) for VLMs.

VLMs. Recent progress in vision–language modeling has led to two major families of approaches. The first is dual-encoder models, where separate vision and text encoders are trained jointly, typically in a contrastive setting. These models excel at recognition-oriented tasks, with representative examples including CLIP [44], ALIGN [21], OpenCLIP [48], SigLIP [66], and MetaCLIP [62], as well as numerous extensions for downstream applications [11, 12, 27, 34, 35, 37, 42]. The second family, often referred to as large multimodal models (LMMs), couples a vision encoder with a large language model (LLM), enabling open-ended multimodal reasoning for tasks such as captioning, visual question answering (VQA), and document understanding. Pioneering approaches in this direction include BLIP-2 [25], InstructBLIP [9], MiniGPT [5, 72], and the LLaVA series [24, 29–31]. More recent models have pushed these capabilities even further: Qwen-2.5 VL [3] advances visual understanding by supporting precise object localization, dynamic resolution processing and strong agentic capabilities such as tool execution. InternVL3 [73] improves percep-

tion and reasoning through native multimodal pretraining and domain-specific data such as 3D scenes, GUIs, and video. Phi-3.5 Vision [1] offers a lightweight but strong alternative with long-context reasoning (128K tokens), robust vision inputs (images, charts, documents), and improved alignment via preference optimization. Several recent studies [13, 17, 20, 28, 38] have further enhanced these models through improved training or adaptation strategies. In this work, we target the most recent open-source LMMs and focus on improving their test-time adaptability for vision-centric tasks such as object recognition, which remains a key weakness highlighted in prior studies [38, 69].

RL-based fine-tuning for VLMs. RL has become a central paradigm for aligning large language models with human preferences and task objectives, with approaches such as RLHF [41] and DPO [45] improving safety, coherence, and instruction-following in both LLMs and VLMs. More recently, rule-based methods like GRPO [51] have demonstrated the feasibility of scaling RL to enhance reasoning capabilities. Building on this foundation, RL-based fine-tuning (RFT) has been extended to multimodal models across a variety of vision-driven tasks. For example, VLM-R1 [52], VisualThinker-R1-Zero [71], and Perception-R1 [65] adapt RFT for open-vocabulary object recognition, spatial reasoning, and visual perception, respectively; CLS-RL [26] applies RFT to few-shot image classification; while R1-VL [67] and related efforts further refine multimodal reasoning. These works demonstrate that RL can significantly enhance vision-centric capabilities of VLMs, but they still rely on curated training splits or labeled feedback. In contrast, our work investigates how reinforcement learning can be performed at *test time*, directly from unlabeled test data, thus bringing RL for VLMs closer to human-like learning from raw experience.

Test-time training (TTT). TTT methods adapt model parameters at inference time without requiring labeled test data, typically by optimizing surrogate objectives such as entropy minimization or self-supervised auxiliary losses [16, 33, 36, 47, 55, 56, 58]. Originally developed for unimodal architectures, these techniques have recently been extended to multimodal systems, with much of the focus on dual-encoder VLMs (*e.g.*, CLIP [44]). Representative examples include TPT [53], which tunes text prompts by minimizing entropy over augmented views, along with its extensions DiffTPT [15] and C-TPT [64], which improve augmentation quality and calibration. RLCF [70] instead adapts the image encoder using feedback from larger models, while several black-box methods operate directly on embeddings without modifying internal parameters.

However, many widely used TTT approaches impose architectural or objective-level constraints that limit their applicability to decoder-based VLMs. For instance, Norm [47] and DUA [33] require batch-normalization layers, and TENT [58] relies on minimizing entropy over a model's output class-probability distribution. Decoder-based VLMs do not produce such class-level distributions; instead, they generate autoregressive token distributions over the full vocabulary, making direct application of TENT-style objectives nontrivial. For comparison, we approximate TENT by minimizing the entropy of the empirical output distribution and report the results in Table 3. Our proposed reward design consistently outperforms this surrogate, highlighting the advantage of our method.

Most closely related to our work, TTRL [74] introduces the idea of using reinforcement learning at test time for LLMs, where the majority voting across sampled outputs serves as a surrogate reward. While sharing the same high-level motivation, our TTRV differs greatly by extending the paradigm to multimodal models and combining frequency-based rewards with entropy regularization to balance consistency and diversity in predictions. The novel reward formulation helps us achieve better performance than the naïve majority voting (*cf*., Ablations Section 4.3). Further, in contrast to VLM-based approaches that primarily target dual-encoder VLMs or prompt-level adaptation, our work focuses on decoder-based VLMs. To the best of our knowledge, our TTRV is the first framework to leverage GRPO for test-time RL of VLMs. Adapting models through RL at inference presents unique challenges, particularly in designing reward functions that operate in a fully unsupervised setting. We address this by rewarding predictions based on their frequency among the model's own outputs, while simultaneously regularizing diversity by rewarding model's certainty obtained by calculating the entropy of the empirical probability distribution. This formulation enables TTRV to achieve consistent improvements across diverse tasks and benchmarks.

3. TTRV: Test Time RL for VLMs

The goal of our proposed TTRV is to improve downstream vision tasks by extracting reward signals directly from unlabeled test data as it is encountered. To this end, we bootstrap an off-the-shelf VLM (*e.g.*, InternVL [73]) with Group Relative Policy Optimization (GRPO) [51]. A key contribution of our work lies in the design of fully unsupervised reward signals. At a high level, we introduce two complementary rewards: (i) a frequency-based reward that encourages consistent answers from the base VLM, and (ii) an entropy-based reward that regularizes the diversity of responses. An overview of our approach is provided in Figure 2 and a comprehensive `python`-like pseudo-code is provided in the supplementary.

For ease of assimilation, the following subsections first provide a brief recap of Group Relative Policy Optimization (Section 3.1), then describe our proposed reward formulations in detail (Section 3.2), and finally present the resulting

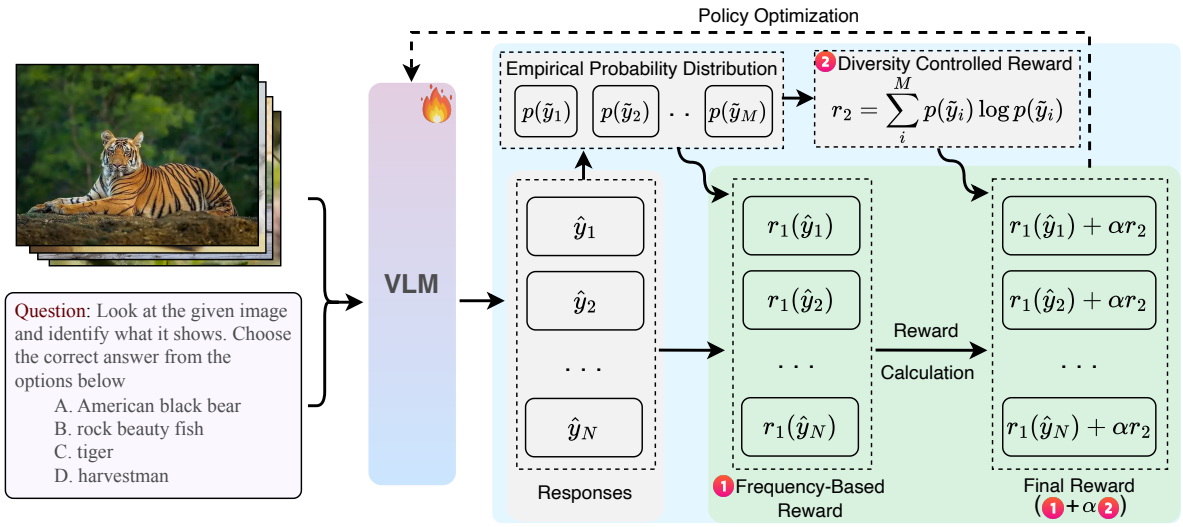


Figure 2. **Overview of TTRV.** For each prompt x, the VLM generates N candidate responses $\{\hat{y}_1, \ldots, \hat{y}_N\}$ from its policy $\pi_\theta(\cdot|x)$. These samples induce an empirical distribution over the unique outputs $\{\tilde{y}_1, \ldots, \tilde{y}_M\}$, from which two reward signals are derived: (i) a *frequency-based reward*, where each response y_j is rewarded in proportion to how often its output occurs among the N responses (*i.e.*, its empirical probability in the distribution), and (ii) a *diversity control reward*, computed from the distribution to regulate diversity and encourage convergence. The final reward is the weighted combination of these terms, which is used to update the policy via GRPO.

optimization objective for our TTRV (Section 3.3).

3.1. Recap: Group Relative Policy Optimization

Let $\mathcal{S}$ denote the space of natural language token sequences. A decoder-based vision-language model $\pi(\cdot|x)$, given an input prompt (image and text) $x \in \mathcal{S}$, produces a probability distribution over possible outputs $y \in \mathcal{S}$, where $y = (y_1, y_2, \ldots, y_T)$ denotes a sequence of tokens. The probability of generating a sequence y is $\pi(y|x) = \prod_{t=1}^{T} \pi(y_t \mid y_{<t}, x)$.

Post-training with reinforcement learning aims to maximize a scalar reward function $r : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ while constraining deviation from a reference policy π_{ref}. This leads to the KL-regularized optimization problem:

$$\max_{\pi} \mathbb{E}_{x\sim D,\, y\sim\pi(\cdot|x)} \Big[r(x,y) \Big] - \beta D_{\text{KL}}\big(\pi(\cdot|x) \,\|\, \pi_{\text{ref}}(\cdot|x)\big), \tag{1}$$

where D is the dataset of prompts, and $\beta > 0$ controls the KL regularization strength.

Group Relative Policy Optimization (GRPO) [51] provides a stable approach to optimize this objective. Given n sampled responses $\{y_i\}_{i=1}^n$ for a prompt x, the advantage of each response is defined as

$$A_i = \frac{r(x, y_i) - \text{mean}_j(r(x, y_j))}{\text{std}_j(r(x, y_j))}, \tag{2}$$

which measures relative performance within the sampled group. The policy update is performed via clipped importance-weighted objectives to ensure stability, while KL regularization keeps the fine-tuned model close to the reference distribution.

3.2. Test-Time RL with Distributional Rewards

Our TTRV extends the vanilla GRPO framework, which is usually applied by extracting rewards by using labeled data, by introducing an inference-time RL framework. We propose to extract self-supervised reinforcement signals directly from the empirical distribution of model outputs at inference time. Unlike settings that rely on external supervision, our framework generates self-consistent rewards that exploit the variability of rollouts to guide convergence during inference. In particular, we extract two rewards from the unlabeled data.

Frequency-Based Reward. Given a copy of the model at a particular time step during test-time learning, our goal is to infer on the test sample multiple times and reward predictions based on their frequency. The underlying intuition is that responses produced more consistently by the model are more likely to be correct. Formally, for each test sample x (consisting of an image and text prompt), we sample N candidate responses $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$ from the current policy $\pi_\theta(\cdot|x)$. Let $\mathcal{U} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_M\}$ denote the set of unique outputs. We estimate the empirical probability of $\tilde{y}_m$ as

$$p(\tilde{y}_m) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\{\hat{y}_j = \tilde{y}_m\}, \tag{3}$$

where $\mathbf{1}$ is an indicator function. The reward for an individual sample $\hat{y}_j$ is then defined as

$$r_1(\hat{y}_j) = \sum_{m=1}^{M} p(\tilde{y}_m) \cdot \mathbf{1}\{\hat{y}_j = \tilde{y}_m\}, \tag{4}$$

which assigns higher values to responses that occur frequently, while still allocating nonzero reward to less common but potentially meaningful alternatives. This graded structure captures the implicit consensus among repeated rollouts without discarding minority reasoning paths.

Importantly, this differs from the standard best-of-N sampling scheme employed by [74], which selects only the most frequent response and discards all others. Such a hard decision can be problematic when the model is uncertain or when the most frequent prediction is incorrect, since it provides a misleadingly strong but potentially wrong reward signal. In contrast, our reward formulation produces a *soft, probabilistic supervision signal* that reflects the full distribution over responses. This perspective is naturally connected to Bayesian reasoning: rather than collapsing to a single point estimate, our method retains uncertainty over hypotheses and uses it to shape learning. We further validate this design choice through ablations against naïve best-of-N sampling, with results reported in Section 4.3.

Diversity Control Reward. Complementing the frequency-based reward r_1, which provides soft, frequency-proportional credit to repeated model responses, we introduce an entropy-based regularizer to control convergence. For a given test sample we compute the Shannon entropy [49] of the empirical response distribution:

$$H(P) = -\sum_{m=1}^{M} p(\tilde{y}_m) \log p(\tilde{y}_m), \tag{5}$$

and define the auxiliary reward

$$r_2 = -H(P). \tag{6}$$

which penalizes excessive dispersion in the output distribution. This mechanism ensures that while the model explores diverse reasoning modes initially (as encouraged by the frequency-based reward), it gradually consolidates probability mass toward stable, high-probability answers rather than spreading attention too thin across redundant responses.

Combined Reward. The overall reward assigned to a response $\hat{y}_j$ is the combination of probability and entropy terms:

$$R(\hat{y}_j) = r_1(\hat{y}_j) + \alpha r_2, \tag{7}$$

where α is a tunable hyperparameter controlling the trade-off between convergence and diversity. By combining probability-based self-rewarding with entropy regularization, the model adaptively aligns its outputs during inference, striking a balance between exploring diverse reasoning paths and converging to coherent predictions.

3.3. Optimization Objective

The reinforcement learning objective is to maximize the expected reward under the policy:

$$\max_{\theta} \ \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[R(y) \right]. \tag{8}$$

For decoder-based VLMs, optimization is performed through the standard autoregressive language modeling objective, with the reward providing a soft, sample-level weighting of predicted tokens. The parameters are updated via gradient ascent:

$$\theta \leftarrow \theta + \eta \nabla_\theta \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[R(y) \right], \tag{9}$$

where η denotes the learning rate.

We note that GRPO [51] modifies this process by replacing the raw reward with a relative advantage term (as defined in Eq. (2)). This shifts optimization from absolute rewards toward relative comparisons, making it more stable and better aligned with group-level objectives.

4. Results

In this section, we first list the implementation details, which include an introduction to the datasets used for evaluating our TTRV, then provide an overview of the different baselines we compare to, and finally conclude with a discussion of the main results and ablations. The details about implementation and evaluation protocols are delegated to the supplementary material.

4.1. Evaluation Settings

Image Recognition Datasets. We evaluate on eight diverse object recognition benchmarks. These include the two original ImageNet test sets: ImageNet [10] and ImageNet-V2 [46], along with three out-of-distribution variants: ImageNet-Rendition (R) [18], ImageNet-Sketch (S) [59], and ImageNet-Adversarial (A) [19]. In addition, we consider two fine-grained recognition datasets: Food101 [4] and the Describable Textures Dataset (DTD) [8], as well as a remote sensing dataset: Resisc45 [7].

VQA Datasets. We further evaluate our TTRV on eight visual question answering (VQA) datasets covering a broad range of reasoning skills. These include two math reasoning benchmarks: MathVerse [68] and MathVista [32]; three datasets focusing on everyday scenarios and objects: SEED [23], MME [63], and RealWorldQA [2]; two compositional reasoning datasets: Capture [43] and Circular-based Relation Probing Evaluation (CRPE) [61]; and one dataset targeting chart-based questions: AI2D [22].

These 16 datasets were deliberately selected to span a broad spectrum of domains and tasks, including natural images, fine-grained categories, remote sensing, mathematical

	ImageNet	ImageNet-V2	ImageNet-R	ImageNet-S	ImageNet-A	Food101	DTD	Resisc45	Mean
GPT-4o	98.30	95.10	91.70	91.20	90.60	95.60	92.30	92.13	93.37
CLIP	68.33	61.86	76.90	48.27	49.91	87.04	44.73	58.22	61.91
MetaCLIP	70.78	62.64	80.99	57.91	46.75	85.48	55.80	66.19	65.82
EVACLIP	74.72	67.03	81.95	57.73	53.91	87.65	52.71	60.37	67.01
SigLIP	76.05	68.97	90.33	67.91	45.33	89.84	64.79	64.54	70.97
LLaMA-3.2-11b	72.68	39.94	72.26	69.39	83.07	93.99	87.54	82.51	75.17
LLaVA-1.5-7b	97.74	95.85	96.46	94.58	94.39	95.35	76.68	92.29	92.92
Phi-3.5-vision	97.94	95.66	96.05	94.93	85.78	96.10	88.26	85.89	92.58
InternVL3-2B	56.00	67.43	66.01	62.19	67.92	67.19	37.24	72.28	62.03
w/ **TTRV**	98.31	**98.25**	**96.89**	94.74	96.31	95.60	**89.73**	90.06	94.99
Δ	+42.31	+30.82	+30.88	+32.55	+28.39	+28.41	+52.49	+17.78	+32.95
InternVL2.5-4B	93.26	83.07	79.53	65.51	90.67	80.92	47.33	23.44	70.47
w/ **TTRV**	97.11	95.66	88.21	92.01	96.00	94.49	81.98	13.30	82.34
Δ	+3.85	+12.59	+8.68	+26.50	+5.33	+13.57	+34.65	-10.14	+11.88
InternVL3-8B	79.47	62.58	59.32	54.48	57.03	78.32	59.11	83.62	66.74
w/ **TTRV**	**99.31**	97.24	96.88	**95.03**	**96.86**	**97.20**	89.37	**93.82**	95.71
Δ	+19.84	+34.66	+37.56	+40.55	+39.83	+18.88	+30.26	+10.20	+28.97

Table 1. **Image Classification.** Top-1 Accuracy (%) obtained by evaluating multiple different backbones. The results in gray are obtained using the specialized dual-encoder VLMs and the proprietary GPT-4o. For decoder-based VLMs we also evaluate multiple families and model sizes. Our TTRV is applied to different model sizes from the InternVL [73] family of models. The best results obtained for a dataset are highlighted in **bold**, while the second best are underlined.

	Mathverse	Mathvista	SEED	MME	RealWorldQA	Capture	CRPE	AI2D	Mean
GPT-4o	54.40	63.80	69.80	89.75	75.40	85.25	76.60	84.60	71.97
LLaMA-3.2-11b	19.36	35.40	62.56	51.59	41.18	61.48	44.43	59.54	53.34
LLaVA-1.5-7b	26.02	34.29	61.50	49.05	59.87	71.75	64.84	47.49	46.02
Phi-3.5-vision	36.02	51.22	**69.54**	77.22	57.25	72.99	68.33	**75.55**	48.58
InternVL3-2B	44.10	58.26	24.99	17.04	63.47	60.27	71.92	39.68	47.47
w/ **TTRV**	48.51	66.11	48.85	11.06	64.29	78.64	72.00	67.75	57.15
Δ	+4.41	+7.85	+23.86	-5.98	+0.82	+18.37	+0.08	+28.07	+9.69
InternVL2.5-4B	51.69	65.49	57.37	85.27	65.25	80.03	74.33	51.55	66.37
w/ **TTRV**	**53.02**	**66.94**	61.14	**85.79**	**66.00**	**85.99**	**75.22**	61.09	69.40
Δ	+1.33	+1.45	+3.77	+0.52	+0.75	+5.96	+0.89	+9.54	+3.03
InternVL3-8B	34.56	38.84	32.12	49.02	19.01	59.50	55.81	30.95	38.05
w/ **TTRV**	42.15	50.41	59.16	78.77	26.57	80.68	68.26	53.92	55.56
Δ	+7.59	+11.57	+27.04	+29.75	+7.56	+21.18	+12.45	+22.97	+17.50

Table 2. **Visual Question Answering.** Results obtained by evaluating multiple different backbones. For decoder-based VLMs, we evaluate multiple families and model sizes. Our TTRV is applied to different model sizes from the InternVL [73] family of models.

reasoning, everyday commonsense, compositionality, and chart understanding. This diversity ensures that our findings are not confined to a single domain but instead provide a representative view of model capabilities across varied and challenging settings. We further expect that the insights derived here will generalize to other benchmarks, which can be incorporated in future evaluations.

Baselines: For comparison, we evaluate the following dual-encoder VLMs: CLIP [44], MetaCLIP [62], EVA-CLIP [54], and SigLIP [66]. As representative methods for decoder-based VLMs, we choose: LLaMA [57], LLaVA [30], Phi-3.5-vision [1], and also provide results for the proprietary GPT-4o [40]. For the main experiments, we apply our TTRV to different model sizes of the InternVL [6] family. However, we want to point out that TTRV can be applied to any open-source VLM.

4.2. Results

Image Classification. In Table 1 we present the top-1 accuracy across eight diverse image recognition benchmarks. We observe that TTRV consistently enhances performance for all evaluated InternVL backbones, with particularly strong gains on challenging distribution shifts such as ImageNet-R and ImageNet-S. For example, when applied to InternVL3-2B, TTRV improves accuracy by up to 89.7% on DTD and 90.0% on Resisc45, while yielding an average gain of around 32.9% across datasets. Similar trends

	Mathvista	SEED	AI2D	Mathverse	CRPE
InternVL2.5-4B (base)	65.49	57.37	51.55	51.69	74.33
w/ Maj Voting	65.08	58.37	47.52	50.75	71.60
Δ vs. base	-0.41	+1.00	-4.03	-0.94	-2.73
TTRV (w/o Freq. reward)	66.81	58.87	52.66	52.00	74.54
Δ vs. base	+1.32	+1.50	+1.11	+0.31	+0.21
TTRV (w/o Diversity reward)	65.08	59.27	53.06	51.74	73.53
Δ vs. base	-0.41	+1.90	+1.51	+0.05	-0.80
TTRV (Freq. + Diversity)	66.94	61.14	61.09	53.02	75.22
Δ vs. base	+1.45	+3.77	+9.54	+1.33	+0.89

Table 3. **Ablating Reward Designs.** We compare the design choices of our TTRV with the reward design proposed by Zuo et al. [74], based on the pseudo-labels obtained from a majority voting scheme. Further, we also ablate the individual effect of our frequency- and diversity-based rewards.

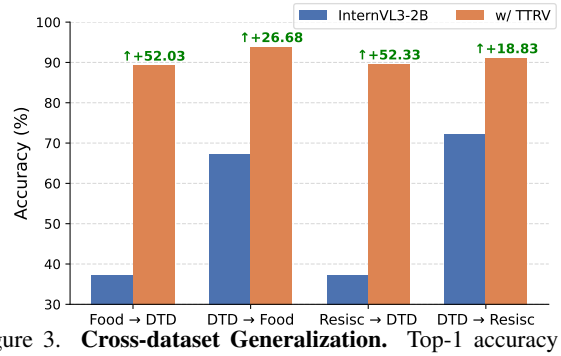


Figure 3. **Cross-dataset Generalization.** Top-1 accuracy (%) achieved by employing TTRV on a base dataset using InternVL3-2B and evaluating on a target dataset from a completely different domain. The results highlight that TTRV enhances core abilities of the model.

appear for larger models: InternVL2.5-4B and InternVL3-8B see systematic boosts on both fine-grained recognition (Food101) and large-scale benchmarks (ImageNet and variants). Notably, TTRV lifts InternVL3-8B to over 99% accuracy on ImageNet, even outperforming proprietary systems (*e.g.*, GPT-4o) and establishing new state-of-the-art results for open-source VLMs. We emphasize that these improvements are achieved by randomly sampling only 20 test instances per dataset, suggesting that TTRV may not be adapting strongly to the test data distribution itself, but rather recovering and amplifying visual recognition capabilities that were present in pre-training and which might be attenuated during instruction tuning. We also observe some particular cases when the model performance decreases (*e.g.*, on the Resics45 dataset with InternVL-2.5-4B model). This can be attributed to the low performance of the base model, which might result in extremely low-quality rollouts from the model or general instability of GRPO [60]. Overall, these findings demonstrate that frequency- and entropy-based test-time rewards allow models to consolidate predictions more effectively, leading to robust improvements in visual recognition.

Visual Question Answering. In Table 2 we report results across eight multimodal reasoning tasks, including mathematical problem solving (MathVerse, MathVista), scientific diagram understanding (AI2D), and general-domain evaluation (RealWorldQA). We find that TTRV provides consistent gains across all datasets and model scales. For instance, on InternVL2.5-4B, accuracy improves by 4.4% on MathVista and 9.5% on AI2D, while the larger InternVL3-8B benefits from 12.4% and 7.5% improvements on CRPE and RealWorldQA, respectively. We also observe that TTRV outperforms other open-source VLMs while also remaining highly competitive with GPT-4o (only lagging behind by $\sim$ 2% on average), and outperforming the strong proprietary model on some benchmarks, like the challenging Mathvista. Furthermore, these gains are also obtained using only 20 sampled instances per dataset, suggesting that the improvements may not stem from distribution-level adaptation but rather from leveraging and re-aligning latent reasoning skills that were learned during pre-training but weakened after instruction tuning. This highlights that TTRV is especially effective at recovering such capabilities under test-time optimization, enabling more robust reasoning across diverse VQA tasks.

4.3. Ablations

In this section, we present extensive ablations of our method and our design choices. We begin by comparing our reward formulation against the naïve best-of-N (majority voting) sampling strategy for RL. Next, we evaluate the robustness of TTRV by training on one dataset and testing on a completely different distribution. We then investigate alternative sampling techniques and reward designs, followed by experiments in extremely data-scarce settings where TTRV is applied to a single randomly chosen test sample. Finally, we report results obtained by applying TTRV to the Qwen-VL model, highlighting generalization of TTRV beyond the Intern-VL models. Due to space constraints, we delegate additional ablations (*e.g.*, on latency) to the supplementary.

Reward Designs. In Table 3 we ablate different reward designs. Specifically, we compare the reward design proposed in our work with the majority voting reward used by [74], and also ablate the effect of the two different rewards used in our work (*cf.*, Section 3). We find that the combination of the two rewards: frequency-based and diversity-control, outperforms all other design choices. We point out that TTRV (w/o Freq. reward) can be considered as the test-time adaptation with entropy minimization proposed by TENT [58]. We see that our method outperforms vanilla entropy minimization as well.

Cross-Data Generalization. In the main results (*cf.*, Tables 1 & 2), we evaluate TTRV on test samples drawn from the same dataset (as used for test-time RL). In contrast,

	Imagenet-A	Imagenet-R
InternVL3-4B (base)	90.67	79.53
w/ biased sampling	95.09	88.51
Δ vs. base	+4.42	+8.98
w/ random sampling	96.00	88.21
Δ vs. base	+5.33	+8.68

Table 4. **Biased vs. Random Sampling.** Top-1 accuracy (%) obtained by sampling the test data differently. For biased sampling, we choose a fraction of the data from only a subset of classes (*e.g.*, 4 out of 200 for ImageNet-R). Random sampling results are obtained by sampling the data randomly from all classes.

	Seed	Imagenet-R
InternVL2.5-4B (base)	57.37	79.53
w/ Random Rewards	52.41	78.00
Δ vs. base	-4.96	-1.53
TTRV (Freq. + Diversity)	61.14	88.21
Δ vs. base	+3.77	+8.68

Table 5. **Random Reward vs. TTRV.** We compare the results obtained by using random rewards (following Shao et al. [50]) and our sophisticated reward design. The results highlight that the gains obtained with *spurious rewards* do not transfer to Intern-VL family of models, which were shown for Qwen-based models.

	Mathvista	SEED	Imagenet-A	Imagent-R
InternVL2.5-4B	65.49	57.37	90.67	79.53
w/ TTRV	66.11	58.87	95.28	85.00
Δ	+0.62	+1.50	+4.61	+5.47

Table 6. **Single Example TTRV.** We report results for VQA and image classification after applying TTRV on a single randomly sampled test example.

	Mathverse	Mathvista	Capture	Resisc45
Qwen2.5-VL-3B	45.33	67.35	71.33	90.08
w/ TTRV	48.71	71.48	75.25	92.71
Δ	+3.38	+4.13	+3.92	+2.63

Table 7. **Generalization to Model Families.** We provide results for the two tasks (Image Classification and VQA) by using the Qwen2.5-VL-3B [3].

Figure 3 reports results when TTRV is applied using one dataset and evaluated on a completely different distribution (*e.g.*, Food101 for TTRV and DTD for testing). We observe that TTRV exhibits strong cross-data generalization, indicating that its performance gains stem not from distribution-specific adaptation but from enhancing the model's underlying task ability (*e.g.*, image classification).

Effect of Data Sampling. We find that TTRV does not require sampling data from all classes in the downstream dataset to achieve strong performance gains. In Table 4, we compare *biased sampling*, where data is drawn from only a small subset of classes (*e.g.*, 4 out of 200 in ImageNet-R), against *random sampling*, where data is sampled uniformly across classes. Even under biased sampling, TTRV yields substantial improvements over the base model.

Random Rewards. Shao et al. [50] recently showed that some models trained with GRPO can exhibit performance gains even when optimized with spurious rewards. As a sanity check, we compare TTRV against such random rewards and report results in Table 5. We find that InternVL [6] models do not appear to benefit from random rewards, suggesting that their improvements under TTRV stem from meaningful reward signals rather than spurious correlations.

Single Sample TTRV. To further test whether the gains from TTRV arise from true task enhancement rather than adaptation to the data distribution, we consider an extreme data-scarce setting where adaptation is performed on only a single randomly chosen test example. Results in Table 6 show that even in this case, TTRV yields measurable improvements, lending additional support to our hypothesis.

Generalization to Model Families. While our main results in Tables 1 & 2 focus on post-training InternVL [6] models, we also examine whether TTRV extends to other architectures. In Table 7, we present results with Qwen2.5-VL [3] and observe consistent gains. These findings suggest that TTRV is not restricted to a single model family and can generalize across diverse VLM architectures. Results for more model families are delegated to the supplementary material.

5. Limitations and Conclusion

Limitations. While we empirically show that TTRV enhances task-specific abilities of the base model rather than adapting to the data distribution, we do not yet provide a theoretical explanation for this behavior. Establishing such a foundation remains an important direction for future work.

Conclusion. We introduced TTRV, the first test-time reinforcement learning framework for VLMs, where rewards are extracted on-the-fly from unlabeled test data. Specifically, we proposed two complementary rewards: one based on the frequency of model predictions and another that regulates diversity in rollouts. Extensive evaluation on 16 benchmarks spanning object recognition and visual question answering shows consistent improvements over strong base models, even outperforming GPT-4. Beyond empirical gains, our ablations highlight data-efficient properties of TTRV and how it enhances task-specific abilities without explicit supervision, pointing to test-time optimization through RL as a powerful paradigm for bridging pre-training and downstream deployment.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*, 2024. 3, 6

[2] X AI. Grok-1.5 Vision Preview, 2024. 5

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 8

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. ECCV*, 2014. 5

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning. In *Proc. ICLR*, 2024. 2

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proc. CVPR*, 2024. 2, 6, 8

[7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. In *Proc. IEEE*, 2017. 1, 5

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proc. CVPR*, 2014. 5

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*, 2023. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1, 2, 5

[11] Sivan Doveh, Assaf Arbelle, Sivan Harary, Amit Alfassy, Roei Herzig, Donghyun Kim, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, et al. Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models. In *NeurIPS*, 2023. 2

[12] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching Structured Vision&Language Concepts to Vision&Language Models. In *Proc. CVPR*, 2023. 2

[13] Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards Multimodal In-Context Learning for Vision & Language Models. *arXiv preprint arXiv:2403.12736*, 2024. 3

[14] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. SEED: Self-supervised Distillation for Visual Representation. In *Proc. ICLR*, 2021. 1

[15] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse Data Augmentation with Diffusions for Effective Test-time Prompt Tuning. In *Proc. CVPR*, 2023. 3

[16] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-Time Training with Masked Autoencoders. In *NeurIPS*, 2022. 3

[17] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Are Vision Language Models Texture or Shape Biased and Can We Steer Them? *arXiv preprint arXiv:2403.09193*, 2024. 3

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proc. ICCV*, 2021. 5

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *Proc. CVPR*, 2021. 5

[20] Irene Huang, Wei Lin, M Jehanzeb Mirza, Jacob A Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuhene, Trevor Darrel, et al. ConMe: Rethinking Evaluation of Compositional Reasoning for Modern VLMs. *arXiv preprint arXiv:2406.08164*, 2024. 3

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proc. ICML*, 2021. 2

[22] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth A Dozen Images. In *Proc. ECCV*, 2016. 2, 5

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. ICML*, 2023. 2

[26] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. Think or Not Think: A Study of Explicit Thinking in Rule-Based Visual Reinforcement Fine-Tuning. *arXiv preprint arXiv:2503.16188*, 2025. 3

[27] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge. In *Proc. ICCV*, 2023. 2

[28] Wei Lin, Muhammad Jehanzeb Mirza, Sivan Doveh, Rogerio Feris, Raja Giryes, Sepp Hochreiter, and Leonid Karlinsky. Comparison Visual Instruction Tuning. *arXiv preprint arXiv:2406.09240*, 2024. 3

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge, 2023. 2

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023. 1, 6

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proc. CVPR*, 2024. 2

[32] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *Proc. ICLR*, 2024. 5

[33] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The Norm Must Go On: Dynamic Unsupervised Domain Adaptation by Normalization. In *Proc. CVPR*, 2022. 3

[34] M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Rogerio Feris, and Horst Bischof. TAP: Targeted Prompting for Task Adaptive Generation of Textual Training Instances for Visual Classification. *arXiv preprint arXiv:2309.06809*, 2023. 2

[35] Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. LaFTer: Label-Free Tuning of Zero-shot Classifier using Language and Unlabeled Image Collections. In *NeurIPS*, 2023. 2

[36] M. Jehanzeb Mirza, Pol Jane Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. ActMAD: Activation Matching to Align Distributions for Test-Time Training. In *Proc. CVPR*, 2023. 3

[37] M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Sivan Doveh, , Jakub Micorek, Mateusz Kozinski, Hilde Kuhene, and Horst Possegger. Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs. In *Proc. ECCV*, 2024. 2

[38] M Jehanzeb Mirza, Mengjie Zhao, Zhuoyuan Mao, Sivan Doveh, Wei Lin, Paul Gavrikov, Michael Dorkenwald, Shiqi Yang, Saurav Jha, Hiromi Wakaki, et al. GLOV: Guided Large Language Models as Implicit Optimizers for Vision Language Models. *arXiv preprint arXiv:2410.06154*, 2025. 3

[39] Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification Over a Large Number of Classes. In *Proc. ICVGIP*, 2008. 1

[40] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 6

[41] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 3

[42] Priyank Pathak, Shyam Marjit, Shruti Vyas, and Yogesh S Rawat. LR0.FM: Low-Res Benchmark and Improving Robustness for Zero-Shot Classification in Foundation Models. In *Proc. ICLR*, 2025. 2

[43] Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. CAPTURe: Evaluating Spatial Reasoning in Vision Language Models via Occluded Object Counting. *arXiv preprint arXiv:2504.15485*, 2025. 5

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. ICML*, 2021. 1, 2, 3, 6

[45] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*, 2023. 3

[46] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proc. ICML*, 2019. 5

[47] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving Robustness Against Common Corruptions by Covariate Shift Adaptation. In *NeurIPS*, 2020. 3

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2

[49] Claude E Shannon. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):379–423, 1948. 5

[50] Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious Rewards: Rethinking Training Signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025. 8

[51] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3, 4, 5

[52] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *arXiv preprint arXiv:2504.07615*, 2025. 3

[53] 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. In *NeurIPS*, 2022. 3

[54] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389*, 2023. 6

[55] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proc. ICML*, 2020. 3

[56] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (Learn at Test Time): RNNs with Expressive Hidden States. *arXiv preprint arXiv:2407.04620*, 2024. 3

[57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*, 2023. 6

[58] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-time Adaptation by Entropy Minimization. In *Proc. ICLR*, 2021. 3, 7

[59] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. In *NeurIPS*, 2019. 5

[60] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. *arXiv preprint arXiv:2504.08837*, 2025. 7

[61] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The All-Seeing Project V2: Towards General Relation Comprehension of the Open World. In *Proc. ECCV*, 2024. 5

[62] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP Data. In *Proc. ICLR*, 2023. 1, 2, 6

[63] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*, 2023. 1, 5

[64] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-TPT: Calibrated Test-Time Prompt Tuning for Vision-Language Models via Text Feature Dispersion. *arXiv preprint arXiv:2403.14119*, 2024. 3

[65] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering Perception Policy with Reinforcement Learning. *arXiv preprint arXiv:2504.07954*, 2025. 2, 3

[66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-training. In *Proc. ICCV*, 2023. 1, 2, 6

[67] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. *arXiv preprint arXiv:2503.12937*, 2025. 3

[68] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *Proc. ECCV*, 2024. 5

[69] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are Visually-Grounded Language Models Bad at Image Classification? *arXiv preprint arXiv:2405.18415*, 2024. 3

[70] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Language Models. *arXiv preprint arXiv:2305.18010*, 2023. 3

[71] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132*, 2025. 3

[72] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *Proc. ICLR*, 2024. 2

[73] 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 preprint arXiv:2504.10479*, 2025. 1, 2, 3, 6

[74] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. TTRL: Test-Time Reinforcement Learning. *arXiv preprint arXiv:2504.16084*, 2025. 3, 5, 7