HARNESSING IMAGE QUESTION DEPENDENCE FOR BETTER VLM TEST-TIME REINFORCEMENT LEARNING

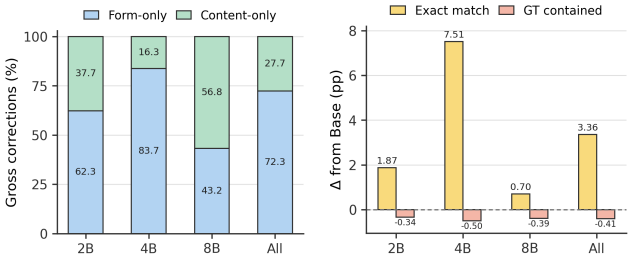
Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time reinforcement learning can adapt vision–language models (VLMs) to unlabeled target data, but relies on reliable self-generated supervision. We analyze consensus-based VLM test-time reinforcement learning across diverse VQA datasets and model sizes and identify two limitations. First, much of its improvement comes from answer normalization rather than content correction. Second, many erroneous responses fail to jointly use the image and the question, allowing consensus rewards to reinforce grounding errors. We propose TTIQ, a test-time reinforcement learning framework that derives supervision from image–question dependence. For each sampled response, TTIQ measures how token likelihoods change under controlled image and question ablations, aggregates the resulting dependence into overall and sustained response-level scores for reward construction, and uses the token-level signals for credit assignment. The reward favors responses that are jointly supported by both inputs and sufficiently confident, while token-level credit emphasizes grounded parts of the response. This shifts supervision from output popularity toward grounded and reliable responses. Across eight VQA datasets and multiple model scales, TTIQ achieves the highest average performance at every evaluated scale, generalizes across VLM families, and transfers to unseen datasets after adaptation on a single source dataset. An anonymized repository is available at `https://anonymous.4open.science/r/TTIQ-BA0B`.

1 INTRODUCTION

Vision–language models (VLMs) are often deployed in settings where the target data distribution differs from the data used for training. Adapting them to such distributions typically requires costly target-domain annotations and supervised fine-tuning (Shu et al., 2022). Test-time reinforcement learning offers a complementary approach by updating a pretrained generative model directly on unlabeled target inputs using learning signals available at inference time (Wei et al., 2025; Zuo et al., 2026). Without access to ground-truth answers, however, the central challenge is how to construct a reliable learning signal that can distinguish responses that should be reinforced from those that should be suppressed.



(a) Composition of exact-match corrections.

(b) Changes in exact match and answer containment.

Figure 1: Where do the gains of TTRV come from? Across eight VQA tasks and model scales, most exact-match corrections are form-only, while improved exact match does not translate into better ground-truth answer containment.

Recent work has explored this challenge largely by deriving rewards from answer consensus across sampled outputs. TTRL (Zuo et al., 2026) treats the majority-vote answer as a pseudo-label for test-time reinforcement learning, while TTRV (Singh et al., 2026) extends this approach to VLMs using response-frequency rewards and an entropy term over the empirical answer distribution. How-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ever, consensus-based supervision has important limitations. First, majority voting can amplify confirmation bias, favor already-solvable examples, and suppress minority correct signals (Wang et al., 2026; Lin et al., 2026). These concerns are particularly relevant to open-ended VQA, where responses vary in both content and form. Because TTRV evaluates its method in a multiple-choice formulation, it remains unclear whether consensus-based adaptation in the free-form setting corrects answer content or primarily normalizes answer form. Second, output-consensus rewards are inherently modality-agnostic: they measure agreement among generated responses without explicitly considering whether those responses are supported by both the image and the question.

To characterize these gaps, we first conduct an in-depth analysis across diverse datasets and VLM scales, summarized here and detailed in Section 3.1. Our analysis reveals two recurring limitations. **First, the gains of consensus-based VLM test-time reinforcement learning largely come from answer normalization rather than content correction.** As shown in Figure 1, across model scales, most apparent corrections are form-only: exact-match accuracy improves even as ground-truth answer containment slightly decreases; **Second, most VLM errors exhibit insufficient joint use of the image and the question, which consensus rewards may then inherit.** As shown in Figure 2, we examine the sampled responses used to construct consensus rewards and find that most incorrect responses lack dependence on at least one input. When such errors recur across rollouts, consensus may treat them as reliable signals and reinforce rather than correct them.

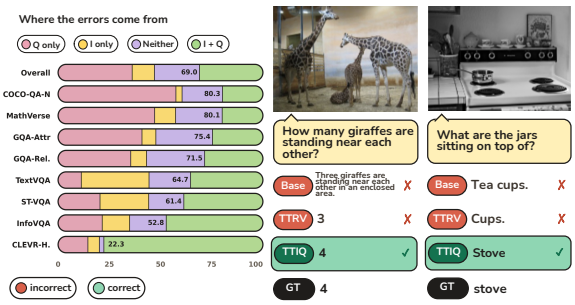


Figure 2: Most incorrect pre-update responses lack joint image–question dependence, with examples showing errors that TTIQ corrects while TTRV preserves.

To address these challenges, we propose TTIQ, a test-time reinforcement learning method for VLMs that rewards responses based on joint image–question dependence. As illustrated in Figure 3, TTIQ teacher-forces each sampled response under the original and image- or question-ablated inputs, using the resulting likelihood changes to estimate dependence on each input. At the **response level**, it aggregates token-level dependence into overall and sustained scores to capture both overall support and whether it extends beyond a few tokens, and combines them with confidence to reward grounded and reliable responses. At the **token level**, it assigns greater positive policy credit to tokens supported by both inputs. This shifts the learning signal from output popularity toward jointly grounded and sufficiently confident responses. Our contributions are:

- We analyze consensus-based VLM test-time reinforcement learning across diverse VQA datasets and models, showing that its gains largely come from answer normalization rather than content correction, while many initial errors reflect limited joint use of the image and the question and can be preserved by consensus-based rewards.
- We introduce TTIQ, the first test-time reinforcement learning method for generative VLMs to explicitly incorporate visual grounding into its training signal. It estimates image and question dependence through image and question ablations, and uses these signals for response-level reward construction and token-level credit assignment.
- Extensive experiments show TTIQ achieves the best average performance at every evaluated model scale. It generalizes across VLM families, and models trained on one dataset consistently improve performance on unseen datasets without further training.

2 Problem Formulation

Let x denote an image, q a question, and $\pi_\theta(y \mid x, q)$ an autoregressive vision–language model. At test time, the pretrained model receives a small unlabeled support set

$$\mathcal{U} = \{(x_j, q_j)\}_{j=1}^{J}, \tag{1}$$

drawn from the target task distribution. Ground-truth answers are not available during adaptation.

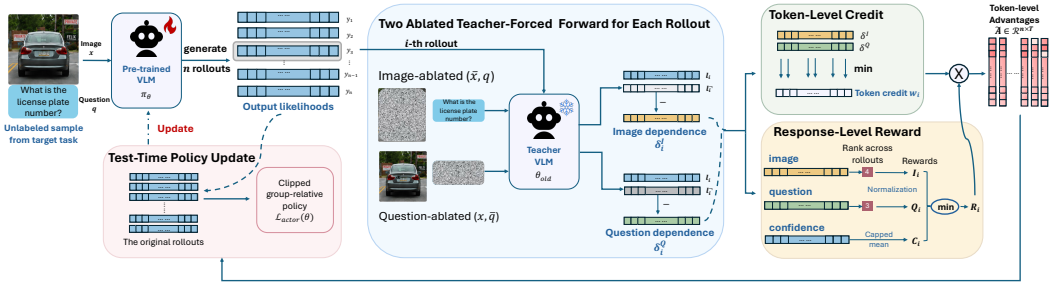


Figure 3: Overview of TTIQ. Given an unlabeled image–question pair, the VLM samples multiple responses. Each response is teacher-forced under image-ablated and question-ablated inputs to estimate token-level image and question dependence. These signals are used to construct a response-level reward and token-level credit assignment, followed by a test-time GRPO update.

At each training iteration, a frozen model $\pi_{\theta_{\text{old}}}$ samples N responses for each support example:

$$y_i = (y_{i,1}, \ldots, y_{i,T_i}), \tag{2}$$

$$y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, q), \qquad i = 1, \ldots, N. \tag{3}$$

The goal is to update π_θ through reinforcement learning using rewards constructed solely from the unlabeled support set and sampled responses, thereby improving performance on the target task.

3 Method

In this section, we first analyze consensus-based VLM test-time training to motivate joint image–question dependence (Section 3.1). We then present our method, including the response-level reward (Section 3.2), joint-token credit assignment (Section 3.3), and GRPO-based update (Section 3.4).

3.1 Limitations of Output-Consensus Signal

We examine whether output consensus provides reliable supervision for VLM test-time reinforcement learning. First, we ask whether its gains reflect actual corrections to answer content or mainly changes in answer form. Second, we study whether consensus remains reliable when sampled responses fail to use both the image and the question. Our findings motivate using image and question dependence, rather than output agreement alone, to construct the learning signal.

Consensus gains largely come from answer normalization. We first examine what consensus-based VLM test-time training changes in model behavior and performance. Across the main evaluation covering eight VQA datasets and three VLM scales, we identify 4,131 exact-match corrections, defined as cases where the base model's response is incorrect under exact match while the TTRV response is correct. As shown in Figure 1, 72.3% are form-only corrections: the base response already contains a ground-truth answer span, while TTRV mainly changes the answer boundary or canonical form to satisfy exact matching. Consistent with this finding, TTRV improves macro exact-match accuracy by 3.36 points, whereas ground-truth answer containment decreases by 0.41 points. Figure 2 (right) provides representative examples: TTRV converts A car" to Car," compresses an incorrect count of three giraffes into 3," and simplifies Tea cups" to "Cups." Together, these results indicate that a substantial portion of the measured improvement reflects answer normalization rather than correction of the predicted content. More examples are provided in Appendix D.

Consensus is constrained by VLM initial failures. We next examine a more fundamental limitation of output-consensus supervision. Because consensus rewards are derived from the model's own sampled responses, repeated errors can themselves become high-consensus signals. When the same error recurs across rollouts, consensus may treat it as a reliable signal and reinforce rather than correct it. This concern is particularly relevant for VLM errors, as prior work shows that when generating responses, models tends to rely heavily on language priors while underusing visual evidence (Asadi

et al., 2026; Lee et al., 2025). We conduct a diagnostic with InternVL2.5-4B using 200 examples from each of eight datasets and four rollouts per example, yielding 6,400 responses. For each response, we keep its tokens fixed and compare their likelihoods under the original input, a zero-image input, and a question-masked input. We use the likelihood drop after ablating an input as a measure of dependence on that input. As shown in Figure 2, 69.0% of incorrect responses lack dependence on at least one input. This finding identifies insufficient joint use of the image and the question as a prevalent failure mode among incorrect VLM responses.

3.2 Response Reward from Image–Question Dependence

Based on the findings, we propose a training reward based on joint image–question dependence. Given a sampled response $y_i = (y_{i,1}, \ldots, y_{i,T_i})$, where i indexes responses and t indexes tokens, we score the same fixed token sequence under three inputs: the original image–question pair, a zero-image input $\bar{x}$, and a question-masked input $\bar{q}$. To construct $\bar{x}$, we replace the pixel tensor with an all-zero tensor of the same shape, while retaining the visual-token layout. To construct $\bar{q}$, we replace only the question content tokens with the tokenizer's padding token, while preserving the sequence length. We hold the response fixed to isolate how much each input supports that response. The corresponding token log-likelihoods are

$$\begin{aligned} \ell_{i,t} &= \log \pi_{\theta_{\text{old}}}\left(y_{i,t} \mid x, q, y_{i,<t}\right), \\ \ell_{i,t}^{\neg I} &= \log \pi_{\theta_{\text{old}}}\left(y_{i,t} \mid \bar{x}, q, y_{i,<t}\right), \\ \ell_{i,t}^{\neg Q} &= \log \pi_{\theta_{\text{old}}}\left(y_{i,t} \mid x, \bar{q}, y_{i,<t}\right). \end{aligned} \tag{4}$$

We define token-level image and question dependence as

$$\delta_{i,t}^{I} = \ell_{i,t} - \ell_{i,t}^{\neg I}, \qquad \delta_{i,t}^{Q} = \ell_{i,t} - \ell_{i,t}^{\neg Q}. \tag{5}$$

Because the response is fixed, each ablated condition requires only one teacher-forced forward pass, without additional autoregressive decoding. A positive likelihood difference indicates that the corresponding input increases the likelihood assigned to that response token.

Overall and sustained dependence. Overall dependence alone can remain high even when a substantial portion of the response has weak dependence on the input. We therefore introduce a sustained-dependence that emphasizes the lower-dependence tokens, allowing weak support over part of the response to be reflected in the response-level dependence score. For response i and modality $m \in \{I, Q\}$, let $k_i = \lceil T_i/2 \rceil$ and $\mathcal{L}_i^m = \text{Low}_{k_i}(\delta_i^m)$, where $\text{Low}_k(\cdot)$ returns the indices of the k smallest finite values. We define

$$d_i^m = \frac{1}{T_i} \sum_{t=1}^{T_i} \delta_{i,t}^m, \qquad s_i^m = \frac{1}{k_i} \sum_{t \in \mathcal{L}_i^m} \delta_{i,t}^m. \tag{6}$$

The mean d_i^m captures overall dependence, while the lower-half average s_i^m ensures that consistently weak dependence over part of the response lowers the resulting response-level score.

Rank-normalized response dependence. The magnitude of likelihood differences can vary with response content and model confidence, making the raw dependence statistics difficult to compare directly across responses. Since the policy update only requires relative preferences among rollouts for the same input, we instead use within-group ranks. For a rollout group $\mathcal{G}$ with n valid responses, let $\rho_{\mathcal{G}}(z_i)$ denote the zero-based average rank of z_i, and define $\mathcal{R}_{\mathcal{G}}(z_i) = 2\rho_{\mathcal{G}}(z_i)/(n-1) - 1$.

We rank-normalize the overall and sustained dependence statistics separately so that they contribute on a common scale, and then combine them with equal weight. The response-level image and question dependence scores are

$$\begin{aligned} I_i &= \mathcal{R}_{\mathcal{G}}\big(\mathcal{R}_{\mathcal{G}}(d_i^I) + \mathcal{R}_{\mathcal{G}}(s_i^I)\big), \\ Q_i &= \mathcal{R}_{\mathcal{G}}\Big(\mathcal{R}_{\mathcal{G}}(d_i^Q) + \mathcal{R}_{\mathcal{G}}(s_i^Q)\Big). \end{aligned} \tag{7}$$

The inner rank transformations place overall and sustained dependence on a common scale, preventing either statistic from dominating due to its magnitude. Their sum favors responses with both strong overall dependence and sustained token-level support, while the outer transformation maps the combined score back to a common group-relative scale for subsequent reward construction.

Minimum-bottleneck reward. Image and question dependence measure whether a response is supported by both inputs, but strong dependence alone does not guarantee that the response is reliable. A response may react to both the image and the question while still being assigned low likelihood by the model. We therefore use model confidence as an additional reliability signal.

We measure confidence using the length-normalized likelihood under the original input:

$$C_i = \mathcal{R}_{\mathcal{G}} \left[\exp \left(\frac{1}{T_i} \sum_{t=1}^{T_i} \ell_{i,t} \right) \right]. \tag{8}$$

The exponential corresponds to the geometric mean token likelihood, while the rank transformation places confidence on the same group-relative scale as the dependence scores.

Importantly, we use confidence only to constrain the reward rather than increase it independently. We define a confidence cap $C_{i,\lambda}^{\text{cap}} = (1-\lambda) + \lambda C_i$, where $\lambda \in [0,1]$ controls the strength of this constraint. When $\lambda = 0$, confidence has no effect; increasing λ allows low-confidence responses to impose a tighter upper bound. The final response-level reward is

$$R_i = \mathcal{R}_{\mathcal{G}} \left(\min \left\{ I_i,\, Q_i,\, C_{i,\lambda}^{\text{cap}} \right\} \right). \tag{9}$$

The minimum implements a bottleneck rather than a compensatory combination: a high score on one component cannot offset weak support from another. Thus, a response is favored only when it is strongly supported by both the image and the question and is not constrained by low confidence.

3.3 Joint-Token Credit Assignment

The response-level reward determines whether a sampled response should be reinforced or suppressed, but assigns the same learning signal to all of its tokens. This can be overly coarse: a positively rewarded response may contain grounded answer tokens together with generic templates, formatting phrases, or unsupported explanations. We therefore use token-level image–question dependence to determine where positive credit should be concentrated.

We first standardize the response rewards within each rollout group to obtain the response-level advantage $A_i = (R_i - \mu_R)/(\sigma_R + \epsilon)$. For each token, let $P_i^I(t)$ and $P_i^Q(t)$ denote the percentile ranks of its image and question dependence, respectively. We define its joint support as $g_{i,t} = \min\{P_i^I(t), P_i^Q(t)\}$. The percentile ranks place the two dependence signals on a common scale, while the minimum requires a token to be supported by both inputs rather than strongly by only one.

We convert $g_{i,t}$ into positive credit weights $w_{i,t}$ and normalize them to have unit mean within each response. Thus, the total positive credit assigned to a response is preserved, but tokens with stronger joint support receive a larger share. For negatively rewarded responses, we do not use token-level support to reduce the penalty: joint dependence indicates where positive evidence is concentrated, but does not provide sufficient evidence to exempt individual tokens from a disfavored response. The resulting token-level advantage is

$$\widetilde{A}_{i,t} = \begin{cases} w_{i,t} A_i, & A_i > 0, \\ A_i, & A_i \le 0. \end{cases} \tag{10}$$

Thus, tokens with stronger joint support receive more positive credit in preferred responses, whereas all tokens in disfavored responses retain the same negative response-level advantage.

3.4 Test-Time Policy Update

We optimize the model with the standard clipped GRPO objective (Shao et al., 2024), replacing the uniform response-level advantage with $\widetilde{A}_{i,t}$ from Eq. (10). For each generated token, the policy ratio under the original input is

$$r_{i,t}(\theta) = \frac{\pi_\theta \left(y_{i,t} \mid x, q, y_{i,<t} \right)}{\pi_{\theta_{\text{old}}} \left(y_{i,t} \mid x, q, y_{i,<t} \right)}. \tag{11}$$

The training objective is

$$\mathcal{L}(\theta) = -\mathbb{E}_{i,t} \left[\min \left\{ r_{i,t}(\theta) \widetilde{A}_{i,t}, \\ \text{clip}(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\, \widetilde{A}_{i,t} \right\} \right] + \beta \mathcal{L}_{\text{KL}}. \tag{12}$$

Table 1: Comparison with all test-time reinforcement learning baselines. Best and second-best results within each backbone block are shown in bold and underlined.

Method	InfoVQA	TextVQA	GQA-Rel.	GQA-Attr	ST-VQA	MathVerse-VD	COCO-QA-N	CLEVR-H	Mean
LLaMA-3.2-11B	46.36	62.59	2.76	14.72	61.55	8.72	36.38	24.19	32.16
LLaVA-1.5-7B	15.48	53.43	41.54	56.30	43.06	14.58	63.80	36.33	40.57
Qwen2.5-VL-3B	66.06	81.87	37.41	55.61	71.25	28.65	73.75	80.60	61.90
Qwen2.5-VL-7B	70.62	84.40	33.69	47.42	76.69	32.16	78.79	80.70	63.06
Qwen3-VL-2B	62.30	80.52	34.21	48.91	69.17	27.08	72.80	71.69	58.34
Qwen3-VL-4B	70.07	81.83	41.00	54.93	71.56	30.73	74.99	79.98	63.14
InternVL3-2B	51.90	76.27	26.08	46.91	75.55	25.91	67.53	65.40	54.44
w/ TTRL	**53.30**	78.71	33.20	51.72	76.32	26.95	72.72	66.92	57.48
w/ SCOPE	53.21	78.45	33.67	51.60	76.29	26.69	72.54	67.08	57.44
w/ TTRV	52.63	78.37	25.66	51.72	76.16	26.43	73.89	65.62	56.31
w/ TTIQ (Ours)	53.21	**79.48**	**35.11**	**53.09**	**77.69**	**27.21**	**75.14**	**69.19**	**58.77**
InternVL2.5-4B	57.89	73.55	17.25	45.59	80.07	28.91	36.01	78.74	52.25
w/ TTRL	**59.79**	79.32	35.71	56.24	82.00	29.43	68.78	78.84	61.26
w/ SCOPE	59.67	78.82	35.35	56.99	81.98	28.91	70.13	**79.06**	61.36
w/ TTRV	58.35	74.76	31.06	55.50	81.84	29.17	69.58	77.85	59.76
w/ TTIQ (Ours)	59.73	**79.52**	**37.15**	**57.10**	**82.67**	**29.56**	**74.70**	78.81	**62.41**
InternVL3-8B	65.58	84.10	35.84	57.22	84.49	33.46	80.51	84.17	65.67
w/ TTRL	65.42	84.54	39.58	59.56	84.72	34.38	80.48	84.39	66.63
w/ SCOPE	**65.82**	**84.90**	39.79	59.11	84.71	34.11	80.55	84.36	66.67
w/ TTRV	65.51	83.86	39.66	58.13	84.52	34.51	80.37	84.39	66.37
w/ TTIQ (Ours)	65.76	84.48	**40.10**	**60.31**	**84.84**	**34.77**	**80.80**	**84.50**	**66.94**

The policy ratio is always evaluated using the original input; the ablated inputs are used only to construct the response reward and token-level credit assignment.

Training efficiency. Relative to standard GRPO, TTIQ adds only two teacher-forced scoring passes, one for each controlled input ablation. Because the response sequence is fixed, all token likelihoods are computed in parallel without additional autoregressive decoding. Appendix B provides detailed training-time comparisons and analyzes how the number of training examples affects performance and runtime.

4 Experimental Results

4.1 Experiment Setup

Datasets and Baselines. We evaluate TTIQ on eight VQA tasks: InfoVQA (Mathew et al., 2022), TextVQA (Singh et al., 2019), ST-VQA (Biten et al., 2019), GQA-Relation and GQA-Attribute (Hudson & Manning, 2019), MathVerse Vision Dominant (Zhang et al., 2024), COCO-QA Number (Ren et al., 2015), and CLEVR-Humans (Johnson et al., 2017a;b). We compare against Base, which performs deterministic zero-shot inference, and TTRV (Singh et al., 2026), the closest prior test-time reinforcement learning method for VLMs. We further compare with two consensus-based methods adapted from LLM test-time reinforcement learning: TTRL (Zuo et al., 2026), which uses the majority-vote answer as a pseudo-label, and SCOPE (Wang et al., 2026), which weights the consensus signal using the model's confidence. We also report zero-shot results from Llama-3.2-11B-Vision (Grattafiori et al., 2024), LLaVA-1.5-7B (Liu et al., 2024), Qwen2.5-VL-3B/7B (Bai et al., 2025b), and Qwen3-VL-2B/4B (Bai et al., 2025a) as VLM reference points. For fair comparison, all test-time training methods use the same prompts, answer extraction procedure, and common hyperparameters. Our primary metric is exact-match accuracy (%). We additionally report Token-F1 and ROUGE-L, together with a detailed decomposition of TTIQ's improvements, in Appendix C.

Implementation. We implement TTIQ with InternVL3-2B, InternVL2.5-4B, and InternVL3-8B (Chen et al., 2024; Zhu et al., 2025), and evaluate cross-family transfer on Qwen2.5-VL-3B/7B. Following TTRV, for each task, we randomly sample 20 examples as unlabeled training data and evaluate on the remaining examples in the corresponding split. Following TTRL and TTRV, we sample 32 responses per prompt at temperature 1.0 and retain 16 for training. We use an actor learning rate of 5×10^{-7} with cosine decay and 3% warmup, and reuse the same prompt as TTRL and TTRV. Prompt and response lengths are capped at 7,524 and 1,024 tokens, respectively. For Table 1, λ is the

Table 2: Ablation study results. Bold indicates the best result in each column over the entire table.

Method	Reward	Credit	InfoVQA	TextVQA	GQA-Attr	GQA-Rel.	Avg.
Reference methods							
Base			57.89	73.55	45.59	17.25	48.57
TTRV	answer frequency		58.35	74.76	55.50	31.06	54.92
Reward components							
Image only	I	full response	59.49	78.71	56.13	36.33	57.66
Question only	Q	full response	57.65	78.41	56.93	23.96	54.24
w/o image	$\min(Q, C)$	full response	58.57	73.96	56.47	15.90	51.23
w/o question	$\min(I, C)$	full response	58.14	73.43	44.04	16.24	47.96
Mean fusion	$\mathrm{mean}(I, Q, C)$	full response	57.77	73.67	44.73	17.07	48.31
w/o sustained dependence	$\min(I_{\mathrm{overall}}, Q_{\mathrm{overall}}, C)$	full response	57.62	73.43	45.19	17.04	48.32
Credit assignment; full reward fixed							
Full-response credit	$\min(I, Q, C)$	full response	59.49	76.37	56.19	33.57	56.40
Joint-token credit	$\min(I, Q, C)$	joint token	59.61	76.51	56.99	35.48	57.14
Confidence component							
w/o confidence	$\min(I, Q)$	joint token	59.64	77.95	57.04	36.53	57.79
TTIQ	$\min(I, Q, C_{\lambda>0})$	joint token	**59.73**	**79.52**	**57.10**	**37.15**	**58.38**

only tuned hyperparameter; we report the best observed result over $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. All experiments use NVIDIA A100 and GH200 GPUs.

4.2 Main Results

Table 1 shows that TTIQ achieves the highest average accuracy at every evaluated model scale and consistently outperforms TTRV across all model-dataset settings. From these results, we make three observations. **Firstly**, the largest gains occur on tasks that require fine-grained joint use of the image and the question, including visual relations, attributes, and counting. Case-level comparisons further show that TTIQ corrects semantic errors that remain unchanged by TTRV, suggesting that its improvements extend beyond answer reformatting to better multimodal grounding. **Secondly**, the gains are smaller on the strongest 8B backbone than on the smaller models, where the pretrained model already has stronger multimodal capabilities. Nevertheless, TTIQ improves performance across all evaluated tasks at this scale, indicating that the proposed learning signal remains useful as model capacity increases. The improvements also extend beyond grounding-heavy tasks. **Finally**, TTIQ remains effective on OCR-intensive and visual mathematical reasoning tasks, showing that image–question dependence provides a useful learning signal across diverse forms of visual reasoning.

4.3 Ablation Study

We conduct comprehensive ablation studies on InternVL2.5-4B across four representative tasks, using the same training split, common hyperparameters, and eight updates for all variants. We evaluate the effects of reward construction, token-level credit assignment, and confidence. Table 2 shows that both image and question dependence are important for reward construction. Removing either signal reduces average performance, while relying on a single input is not consistently reliable across tasks. The aggregation strategy also matters: minimum fusion outperforms mean fusion, and incorporating sustained dependence substantially improves over using only overall dependence. These results support a reward that requires strong and persistent support from both inputs. Joint-token credit further improves performance by assigning more positive credit to tokens supported by both the image and the question, while calibrated confidence helps suppress unreliable responses. Combining these components yields the strongest overall performance.

4.4 Generalization

Across datasets. We evaluate cross-dataset generalization by adapting InternVL3-2B and InternVL2.5-4B on a single source dataset and evaluating them directly on unseen target datasets.

As shown in Table 3, TTIQ consistently improves over Base and TTRV across all target datasets, including transfers between substantially different tasks such as GQA-Relation, COCO-QA-N, and MathVerse-VD. These results indicate that TTIQ learns transferable updates

Table 3: Cross-dataset generalization of TTIQ. Bold values indicate the best result for each setting.

	IVL3-2B				IVL2.5-4B			
Source: GQA-Attr								
Target →	GQA-Rel	COCO-N	MathV-VD	Avg.	GQA-Rel	COCO-N	MathV-VD	Avg.
Base	26.08	67.53	25.91	39.84	17.25	36.01	28.91	27.39
TTRV	29.46	72.72	26.82	43.00	30.00	71.12	29.17	43.43
TTIQ	**30.34**	**73.31**	**26.95**	**43.53**	**33.44**	**74.19**	**29.43**	**45.68**
Source: GQA-Rel								
Target →	GQA-Attr	COCO-N	MathV-VD	Avg.	GQA-Attr	COCO-N	MathV-VD	Avg.
Base	46.91	67.53	25.91	46.78	45.59	36.01	28.91	36.84
TTRV	50.46	71.66	26.17	49.43	54.41	66.54	29.56	50.17
TTIQ	**51.55**	**74.37**	**26.43**	**50.78**	**56.70**	**75.28**	**30.21**	**54.06**
Source: TextVQA								
Target →	GQA-Rel	GQA-Attr	COCO-N	Avg.	GQA-Rel	GQA-Attr	COCO-N	Avg.
Base	26.08	46.91	67.53	46.84	17.25	45.59	36.01	32.95
TTRV	29.72	48.51	72.91	50.38	16.91	48.91	33.53	33.12
TTIQ	**30.52**	**49.14**	**73.67**	**51.11**	**30.16**	**55.67**	**72.39**	**52.74**

that generalize across distinct task distributions and remain effective on unseen datasets.

Across VLM families. We further test cross-family generalization on Qwen2.5-VL-3B and Qwen2.5-VL-7B. As shown in Table 4, TTIQ achieves the highest average performance at both model scales and remains competitive across all evaluated tasks. These results demonstrate that the effectiveness of TTIQ transfers beyond the InternVL family and is not specific to a particular VLM backbone.

Table 4: Generalization to Qwen2.5-VL.

Method	InfoVQA	GQA-A	GQA-R	COCO-N	Avg.
Qwen2.5-VL-3B					
Base	**66.06**	55.61	37.41	73.75	58.21
TTRV	65.97	55.44	37.31	73.49	58.05
TTIQ	66.03	**55.90**	**37.70**	**74.19**	**58.45**
Qwen2.5-VL-7B					
Base	70.62	47.42	33.69	78.79	57.63
TTRV	70.53	50.11	36.97	78.90	59.13
TTIQ	**70.87**	**51.78**	**37.67**	**79.01**	**59.83**

4.5 Correction Composition

We further analyze whether the exact-match improvements come from changes in answer form or corrections to the predicted content. As shown in Figure 4, TTRV's corrections relative to Base are largely form-only, while TTIQ produces a larger proportion of content-changing corrections. More importantly, among cases where TTRV remains incorrect but TTIQ becomes correct, content corrections account for slightly more than half of the improvements. These results show that TTIQ preserves the benefits of improved answer formulation while extending test-time adaptation to more substantive corrections of the predicted content, consistent with its stronger Token-F1 and ROUGE-L results.

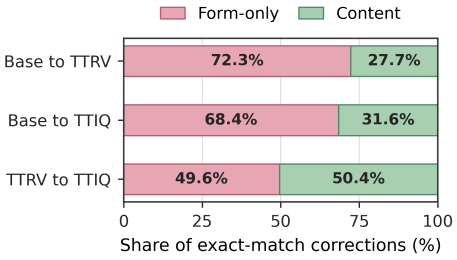


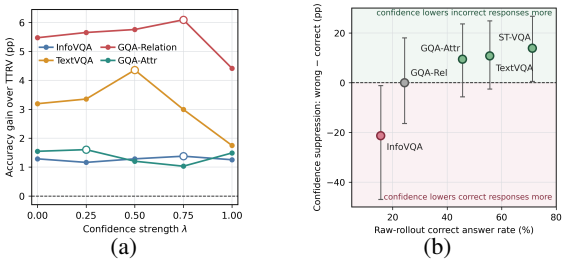
Figure 4: Composition of exact-match corrections into form-only and content-changing cases for Base→TTRV, Base→TTIQ, and TTRV→TTIQ.

4.6 Sensitivity Analysis

Confidence strength. To study sensitivity to the only method-specific hyperparameter, we vary the confidence strength λ on InternVL2.5-4B while keeping all other settings fixed. Figure 5(a) shows that TTIQ outperforms TTRV throughout the sweep. Moderate values work best on TextVQA and GQA-Relation, whereas $\lambda = 1$ reduces their gains; InfoVQA and GQA-Attribute vary by less than 0.6 points. Thus, a practical guideline is that TTIQ is generally robust to λ, but sensitive tasks benefit from a moderate rather than the strongest confidence constraint.

Confidence and initial rollout quality. We further examine when stronger confidence is beneficial using InternVL3-1B. Figure 5(b) relates raw-rollout accuracy to the difference in confidence-based suppression between incorrect and correct responses, where

positive values indicate stronger suppression of incorrect responses. On InfoVQA, where correct rollouts are scarce, confidence suppresses correct responses more strongly; on ST-VQA, where correct rollouts are more common, it preferentially suppresses incorrect responses. Thus, the same confidence signal can either filter erroneous responses or weaken useful ones, depending on the quality of the initial policy. This also explains why applying the strongest confidence constraint is not uniformly beneficial across tasks. This suggests refining the guideline according to initial rollout quality: stronger confidence is more appropriate when the initial policy already produces reliable responses, while weaker or moderate confidence is safer otherwise.



(a) (b)

Figure 5: Sensitivity to confidence strength and initial rollout quality.

5 Related Work

LLM and VLM test-time reinforcement learning. Test-time training Sun et al. (2020; 2025) adapts pretrained models using unlabeled target inputs and has recently been extended to large language-model (LLM) reasoning through reinforcement learning. Existing work in this direction has primarily focused on LLMs. TTRL uses majority-vote responses from repeated sampling as pseudo-labels Zuo et al. (2026); subsequent work improves this consensus signal using step-wise confidence Wang et al. (2026), or analyzes how majority voting can suppress initially rare correct responses Lin et al. (2026). In comparison, test-time reinforcement learning for generative VLMs remains less explored. TTRV applies this paradigm to VLMs using response-frequency rewards and the empirical answer distribution Singh et al. (2026). Existing approaches therefore largely derive supervision from consistency within the model's own outputs He et al. (2025), leaving it unclear whether the resulting improvements reflect stronger multimodal understanding or merely greater output consistency.

Grounding in VLMs. Despite their strong performance, VLMs often produce responses without faithfully using the provided visual evidence. Recent analyses reveal different forms of this failure Tong et al. (2024); Li et al. (2024). Mirage shows that frontier VLMs can retain substantial benchmark performance and generate detailed visual reasoning even without image access Asadi et al. (2026). VLind-Bench further isolates language-prior-induced blindness through counterfactual tests Lee et al. (2025). Motivated by these grounding failures, a complementary line of work leverages visual grounding signals to improve model predictions at inference time. VCD contrasts predictions under original and distorted images to reduce reliance on language priors Leng et al. (2024), while entropy-based methods use model uncertainty to identify or retrieve relevant visual evidence Gröpl et al. (2026). Grounding signals have also been incorporated into VLM training through fine-grained feedback and reward modeling Yu et al. (2024); Yan et al. (2024). Given the prevalence of grounding failures and the effectiveness of grounding signals in inference and training, their use for test-time fine-tuning remains underexplored.

6 Conclusion

We studied test-time reinforcement learning for VLMs and identified two limitations of output-consensus supervision. Motivated by these findings, we proposed TTIQ, which estimates token-level image and question dependence through controlled input ablations, aggregates these signals into overall and sustained response-level scores, combines them with calibrated confidence for reward construction, and uses the same token-level signals for fine-grained credit assignment. Experiments across diverse VQA tasks and VLM scales show that TTIQ achieves the best average performance at every evaluated scale. It also generalizes across VLM families and transfers to unseen datasets after adaptation on a single source dataset. These results suggest that image–question dependence provides a promising alternative to output consensus for VLM test-time reinforcement learning. Future work may explore adaptively calibrating confidence based on initial rollout quality and extending dependence-based supervision to broader multimodal tasks.

AI USE STATEMENT

We used generative AI tools, primarily OpenAI ChatGPT, to assist with the literature review at the very early stage of the project and code review when implementing the method. We also used these tools to improve clarity and readability of author-written text and to check grammar errors when preparing the paper draft. Research design, methodology, experiments, mathematical derivations, analysis, and final claims were made and verified by the authors. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

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

Shuai Bai, Ke qin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025b. URL https://api.semanticscholar.org/CorpusID:276449796.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *arXiv preprint arXiv:2604.08456*, 2026.

Xinrui He, Yikun Ban, Jiaru Zou, Tianxin Wei, Curtiss Cook, and Jingrui He. LLM-forest: Ensemble learning of LLMs with graph-augmented prompts for data imputation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6921–6936, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.361. URL https://aclanthology.org/2025.findings-acl.361/.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017a.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision*, pp. 2989–2998, 2017b.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

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

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Hongxiang Lin, Zhirui Kuai, Erpeng Xue, and Lei Wang. Detecting and mitigating the correct-answer extinction window in test-time reinforcement learning with majority voting. *arXiv preprint arXiv:2605.19444*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

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.

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. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33153–33163, 2026.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

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

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at test time): RNNs with expressive hidden states. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/sun25h.html`.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. Beyond majority voting: Towards fine-grained and more reliable reward signal for test-time reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37251–37265, 2026.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. 2025. URL https://api.semanticscholar.org/CorpusID:278959690.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pp. 37–53. Springer, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

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 *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

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.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

Dataset	Task focus	Public labeled source split	Unlabeled train	Evaluation
InfoVQA	Infographic reading and reasoning	Official test split	20	3,268
TextVQA	Scene-text question answering	Validation split	20	4,980
GQA-Relation	Visual relations	Test-dev balanced, Relation subset	20	3,873
GQA-Attribute	Visual attributes	Test-dev balanced, Attribute subset	20	1,746
ST-VQA	Scene-text question answering	Public train split	20	26,054
MathVerse-VD	Vision-dominant mathematical reasoning	Test-mini, Vision-Dominant subset	20	768
COCO-QA-N	Counting	Test split, Number subset	20	2,735
CLEVR-Humans	Compositional visual reasoning	Validation split	20	7,182
Total			160	50,606

Table 5: Dataset statistics for the experiments. The 20 training answers are withheld from every test-time training method.

A Dataset Statistics

Evaluation splits. For each dataset, we use the official test split whenever its ground-truth annotations are publicly available. We randomly sample 20 examples from this split for test-time training and hide their answers throughout training; all remaining examples are used for evaluation. Although TTIQ does not access labels during training, ground-truth answers are still required to compute the evaluation metrics. Therefore, for datasets whose test annotations are not publicly released, we instead use the largest publicly available labeled validation or evaluation split and apply the same 20-example sampling protocol. Table 5 reports the exact source split and the resulting training and evaluation sizes for each dataset.

Train n	Method	InfoVQA	TextVQA	GQA-Rel.	COCO-QA-N	Avg.
20	TTRV	58.31	74.73	31.00	69.91	58.49
	TTIQ	**59.69**	**79.45**	**37.12**	**75.03**	**62.82**
50	TTRV	**59.03**	77.12	35.43	73.56	61.29
	TTIQ	58.85	**79.14**	**37.33**	**73.71**	**62.26**
100	TTRV	57.94	77.86	**38.20**	72.81	61.70
	TTIQ	**58.47**	**80.43**	37.36	**74.16**	**62.60**

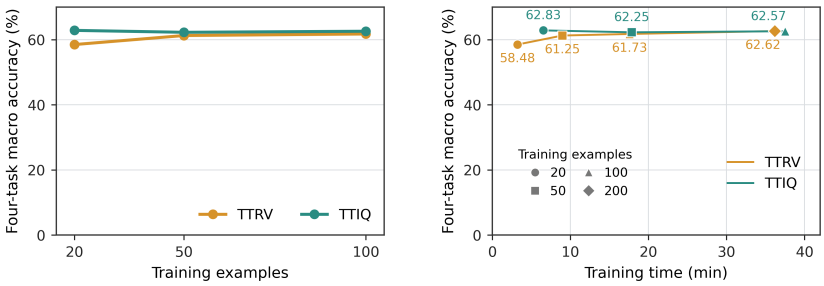
Table 6: Exact match accuracy (%) with 20, 50, and 100 unlabeled training examples. Bold marks the better method for each training size and task.

B Training-Set Size and Efficiency

B.1 Impact of Training-Set Size

We evaluate TTIQ and TTRV using nested sets of 20, 50, and 100 unlabeled examples on four representative tasks with InternVL2.5-4B. All settings use the same held-out split obtained by excluding the full 100-example training pool. We increase the number of updates proportionally with the training-set size, keeping the average number of exposures per example approximately constant.

Table 6 and Figure 6 show distinct scaling behavior. TTIQ achieves strong macro result with only 20 examples and remains broadly stable as more examples are added. In contrast, TTRV improves more with training-set size, but its result with 100 examples still remains below that of TTIQ with 20 examples. TTIQ also wins on most individual task–size combinations. This contrast reflects a fundamental difference in their supervision signals. TTRV relies on response agreement as an indirect proxy for correctness, whose useful effects may accumulate over more training examples. In contrast,



(a) (b)

Figure 6: Effect of training-set size with InternVL2.5-4B. **(a)** Macro accuracy over InfoVQA, TextVQA, GQA-Relation, and COCO-QA-N. **(b)** Mean training time on four GH200 GPUs. Training time includes only the update loop; model setup and evaluation are excluded.

TTIQ uses controlled input ablations to expose, within each example, whether a response is grounded in both the image and the question. It therefore directly corrects how the model uses multimodal evidence rather than relying on a useful training signal to emerge across many examples. Once this effective signal is available, increasing the training set mainly broadens example coverage, resulting in limited and task-dependent gains beyond 20 examples.

Method	Extra TF	Sec./update	8 updates	Peak GB/GPU
TTRV	0	21.9	2.92 min	76.84
TTIQ	2	35.8	4.78 min	76.84

Table 7: Matched training-only efficiency on InternVL2.5-4B and InfoVQA with four GH200 GPUs.

B.2 Training Efficiency

Table 7 reports a matched training-only comparison on InfoVQA. The two additional teacher-forced scoring passes increase the update time, but do not increase peak GPU memory. With eight updates, TTIQ finishes in 4.78 minutes, compared with 2.92 minutes for TTRV, while both use the same peak memory.

Figure 6 further shows that training time grows approximately linearly with the number of updates. Because test-time training is performed on a small unlabeled subset, the absolute training cost remains modest, making the additional scoring overhead of TTIQ acceptable in practice. In particular, the 20-example setting provides a favorable accuracy–efficiency trade-off.

C Beyond Exact Match

Task	Question	Base	TTRV	TTIQ	GT
GQA-Rel.	What vehicle is in front of the red vehicle?	A car.	Car.	**SUV**	suv
GQA-Rel.	What are the jars sitting on top of?	Tea cups.	Cups.	**Stove.**	stove
GQA-Attr	How clean are the pots?	The pots are clean.	Clean.	**Dirty.**	dirty
GQA-Attr	Which side of the image is the cyclist on?	Left.	Left	**Right.**	right
ST-VQA	What number is this bus known as?	Private bus.	Private.	**23**	23
ST-VQA	What does the advertisement say behind the woman's head?	Stretchy shery	Stretchy	**Stretch**	Stretch
COCO-QA-N	How many giraffes are standing near each other in an enclosed area?	Three giraffes are standing near each other in an enclosed area.	3	**4**	four
COCO-QA-N	How many people are on snow skis while one also has wheels?	Two people on snow skis, one of whom has a wheelchair.	2	**3**	three
CLEVR-H	What color is the metallic cylinder?	Blue.	The metallic cylinder is blue.	**Red**	red

Table 8: Representative form-to-content correction chains from the final InternVL3-2B outputs. TTRV changes the surface form or length of the Base response but retains the same central incorrect answer, whereas TTIQ changes the answer content to an acceptable ground truth.

Method	InternVL3-2B		InternVL2.5-4B		InternVL3-8B	
	Token-F1	ROUGE-L	Token-F1	ROUGE-L	Token-F1	ROUGE-L
Base	61.65	61.56	59.94	59.87	70.22	70.14
TTRL	62.65	62.58	65.41	65.36	69.88	69.82
SCOPE	62.66	62.59	65.55	65.49	69.75	69.69
TTRV	62.16	62.07	65.23	65.17	69.79	69.73
TTIQ	**63.53**	**63.45**	**66.96**	**66.89**	**70.76**	**70.69**

Table 9: Macro Token-F1 and ROUGE-L F1 over the eight complete evaluation sets. Best results within each backbone are bold.

C.1 Additional Evaluation Metrics

We further report Token-F1 and ROUGE-L F1 alongside exact-match accuracy. These metrics provide a more graded evaluation by measuring token and sequence overlap between the prediction and the ground-truth answer. Let $N(z)$ denote the token sequence obtained after applying the same normalization as the exact-match evaluator.

For a prediction $\hat{y}$ and a ground-truth answer a, we define their multiset token overlap as

$$O(\hat{y}, a) = \sum_{w} \min(c_w(N(\hat{y})), c_w(N(a))), \tag{13}$$

where $c_w(\cdot)$ denotes the number of occurrences of token w. Token-level precision, recall, and F1 are then defined as

$$\begin{aligned} P_{\text{tok}} &= \frac{O(\hat{y}, a)}{|N(\hat{y})|}, \qquad R_{\text{tok}} = \frac{O(\hat{y}, a)}{|N(a)|}, \\ F_{\text{tok}} &= \frac{2P_{\text{tok}}R_{\text{tok}}}{P_{\text{tok}} + R_{\text{tok}}}. \end{aligned} \tag{14}$$

For ROUGE-L, let

$$L(\hat{y}, a) = \text{LCS}(N(\hat{y}), N(a)) \tag{15}$$

denote the length of the longest common subsequence between the normalized prediction and ground-truth answer. ROUGE-L precision, recall, and F1 are defined as

$$\begin{aligned} P_{\text{L}} &= \frac{L(\hat{y}, a)}{|N(\hat{y})|}, \qquad R_{\text{L}} = \frac{L(\hat{y}, a)}{|N(a)|}, \\ F_{\text{L}} &= \frac{2P_{\text{L}}R_{\text{L}}}{P_{\text{L}} + R_{\text{L}}}. \end{aligned} \tag{16}$$

When an example contains multiple ground-truth answers, we take the maximum score over the ground-truth answer set $\mathcal{G}$:

$$\begin{aligned} S_{\text{tok}}(\hat{y}, \mathcal{G}) &= \max_{a \in \mathcal{G}} F_{\text{tok}}(\hat{y}, a), \\ S_{\text{L}}(\hat{y}, \mathcal{G}) &= \max_{a \in \mathcal{G}} F_{\text{L}}(\hat{y}, a). \end{aligned} \tag{17}$$

Table 9 reports macro averages over the eight complete evaluation sets. TTIQ achieves the highest average Token-F1 and ROUGE-L F1 for every backbone. Since these metrics place greater emphasis on content overlap and are less sensitive to minor differences in answer form, the results suggest that TTIQ's gains extend beyond answer-form changes to better alignment with the ground-truth content.

D Additional Qualitative Examples

D.1 Additional Content Corrections

Table 8 supplements the first analysis in Section 3.1 with cases not shown in the main paper. In every row, Base and TTRV retain the same incorrect content, whereas TTIQ changes the predicted object, attribute, count, text, or mathematical decision to an acceptable answer. The table covers all eight evaluation tasks and seven further cases.

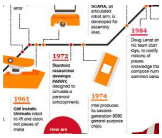


Image salience without question alignment (InfoVQA)
Q: When was the second-generation 8080 chip produced?
× **Image-only:** "1972"
$I = +1.00,\ Q = -1.00$
✓ **Joint:** "1974"
$I = -.33,\ Q = +.33$



Image salience without question alignment (TextVQA)
Q: What is the name of this issue?
× **Image-only:** "99 DEC"
$I = +1.00,\ Q = -1.00$
✓ **Joint:** "Temptation: Lust"
$I = .00,\ Q = .00$



Image salience without question alignment (ST-VQA)
Q: What is the brand of mustard?
× **Image-only:** "DeliStyle Spanish"
$I = +1.00,\ Q = -1.00$
✓ **Joint:** "Publix"
$I = -.33,\ Q = +.33$



Image salience without question alignment (COCO-QA)
Q: How many trains are on the multiple tracks side by side?
× **Image-only:** "5"
$I = +.33,\ Q = -1.00$
✓ **Joint:** "1"
$I = +1.00,\ Q = +1.00$



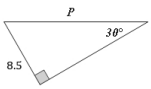
Weak image dependence (TextVQA)
Q: What is the number on the last player in blue?
× **Question-only:** "16"
$I = -1.00,\ Q = -1.00$
✓ **Joint:** "19"
$I = +.67,\ Q = +.67$



Weak image dependence (GQA)
Q: What is the color of the soft blanket?
× **Question-only:** "green"
$I = -1.00,\ Q = +.33$
✓ **Joint:** "white"
$I = +1.00,\ Q = +1.00$



Weak image dependence (COCO-QA)
Q: How many large giraffes are standing next to the barn?
× **Question-only:** "two"
$I = -.33,\ Q = -.33$
✓ **Joint:** "three"
$I = +.67,\ Q = +.67$



Weak image dependence (MathVerse)
Q: Solve for the unknown P in the adjacent figure.
× **Question-only:** "10"
$I = -1.00,\ Q = +.33$
✓ **Joint:** "17"
$I = +1.00,\ Q = +1.00$

Figure 7: Qualitative examples of image–question dependence before test-time training. Each example compares two responses sampled for the same image–question input. Incorrect responses either follow visual content without sufficient question alignment (top) or exhibit weak image dependence (bottom), while the correct responses depend more strongly on both inputs.

D.2 Examples of Joint Image–Question Dependence

Figure 7 presents eight additional image–question examples, each with two responses sampled from InternVL2.5-4B before test-time training. The displayed I and Q values measure how strongly each response depends on the image and the question, respectively.

In the top four examples, the incorrect response has strong image dependence but weak question dependence. It follows salient visual content without correctly addressing the question. In the bottom four examples, the incorrect response has weak image dependence and relies more heavily on textual cues or language priors, resulting in an answer that is not sufficiently supported by the image. These examples show that strong dependence on either input alone is insufficient. TTIQ therefore rewards responses only when they depend strongly on both the image and the question. Across all examples, the response with stronger joint dependence gives the correct answer.