CONSCORE-VG: CONSISTENCY-SCORE-GUIDED TEST-TIME REINFORCEMENT LEARNING FOR VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding requires models to localize image regions according to language instructions and is a fundamental capability of vision-language models. Recent approaches based on supervised fine-tuning or reinforcement learning have achieved strong performance, but adapting them to new scenarios without additional annotations remains challenging. Test-time reinforcement learning (TTRL) offers a label-free alternative, yet extending it to visual grounding is challenging because continuous coordinate outputs are not naturally compatible with standard consensus-based pseudo-labeling, and spatial consensus among predictions does not necessarily imply correct localization. To address these challenges, we propose ConScore-VG, a consistency-guided TTRL framework that combines spatial and semantic consistency scores for reliable pseudo-label selection. The Spatial Consistency Score measures how well each candidate localization agrees with the overall spatial consensus, while the Semantic Consistency Score evaluates region-instruction alignment through visual information gain. Extensive experiments on referring expression comprehension (REC) and GUI grounding benchmarks demonstrate that ConScore-VG consistently improves grounding performance across both settings. Notably, ConScore-VG yields consistent gains on REC with only 256 unlabeled samples and improves over the baseline by 6.68% on ScreenSpot-v2. These results demonstrate the potential of TTRL for visual grounding and establish ConScore-VG as a data-efficient approach for adapting vision-language models to downstream grounding tasks without labels.

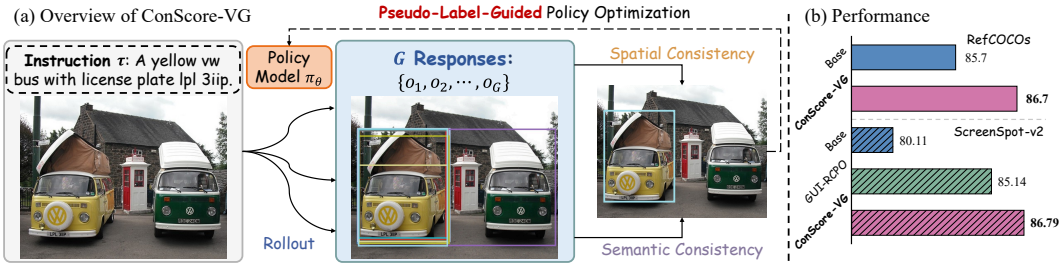


Figure 1: **(a) Overview of ConScore-VG.** On unlabeled test data, ConScore-VG leverages spatial and semantic consistency to select reliable pseudo-labels for policy optimization. **(b) Performance.** Average accuracy evaluated separately on the RefCOCOs validation splits and ScreenSpot-v2, using Qwen2.5-VL-3B as the base model.

1 INTRODUCTION

Visual grounding (Peng et al., 2023; You et al., 2024) aims to localize the image region referred to by a given natural language instruction and is fundamental to fine-grained vision-language understanding. By explicitly linking linguistic expressions to visual elements (Li et al., 2026), it supports a wide range of complex vision-language applications, including robotic manipulation (Huang et al., 2025a; Li et al., 2025c), autonomous driving (Zhang et al., 2026; Liao et al., 2026), and multimodal

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.

reasoning (Cao et al., 2025; Wang et al., 2026b; Wu et al., 2026), all of which benefit from precise localization for reliable perception, reasoning, and decision-making.

Recent progress in visual grounding has been driven primarily by two training paradigms, supervised fine-tuning (SFT) on large-scale annotated datasets (Zhao et al., 2023; Liu et al., 2024; Lai et al., 2024; Sun et al., 2025; Zhu et al., 2026) and reinforcement learning (RL) with task-specific localization rewards (Shen et al., 2025; Liu et al., 2025d; Bai et al., 2025b; Huang et al., 2025b). Despite their effectiveness, both paradigms remain fundamentally limited in two aspects. First, they primarily optimize models during training, leaving the potential for test-time adaptation largely untapped. Second, they depend heavily on labeled data, often requiring precise spatial annotations for each sample, which poses a substantial bottleneck to scaling visual grounding toward new domains and complex scenarios where such annotations are unavailable (Zuo et al., 2025). This motivates a more promising direction in which models can adapt and improve their grounding capabilities at test time in the absence of explicit supervision and ground-truth-derived reward signals.

Test-time reinforcement learning (TTRL) (Zuo et al., 2025; He et al., 2026; Yu et al., 2026) has recently emerged as a promising paradigm for adapting models directly on unlabeled test data using self-generated reward signals. Building on this idea, TTRL has also been extended to multimodal models. For instance, Singh et al. (2026) introduced TTRV, which adapts vision-language models on unlabeled test samples using rewards derived from prediction frequency and output diversity. These advances naturally motivate the application of TTRL to visual grounding, yet its continuous output space introduces distinct challenges. Unlike reasoning tasks with discrete answers, visual grounding predicts continuous spatial coordinates, making methods that construct pseudo-labels or reward signals based on answer frequency difficult to apply directly. Although GUI-RCPO (Du et al., 2026) has explored TTRL for GUI grounding, a specialized form of visual grounding, its reward design is tailored to this setting and does not readily generalize to broader visual grounding scenarios. More critically, it relies solely on region consistency among sampled predictions, which can introduce supervision noise because spatial agreement does not guarantee correct localization and may reinforce consistently incorrect predictions (He et al., 2026; Lin et al., 2026). Additional related work is discussed in Appendix A.1.

To address these challenges, we propose **ConScore-VG**, a TTRL framework for **V**isual **G**rounding guided by spatial and semantic **Con**sistency **Score**s, as illustrated in Figure 1(a). ConScore-VG evaluates each sampled prediction using two complementary consistency signals, namely a **Spatial Consistency Score** that captures the spatial consensus among sampled predictions and a **Semantic Consistency Score** that measures the semantic relevance between the predicted region and the language instruction from an information-gain perspective (Shannon, 1948). The latter helps mitigate the pseudo-label noise caused by relying solely on spatial consistency. By jointly leveraging these two complementary signals, ConScore-VG selects reliable pseudo-labels and constructs self-supervised rewards for Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Without requiring additional annotations, ConScore-VG enables effective test-time adaptation across diverse visual grounding scenarios.

To validate the effectiveness of our proposed ConScore-VG, we evaluate it on the RefCOCOs benchmarks for general visual grounding and the ScreenSpot series benchmarks for GUI grounding. Using only 256 unlabeled test samples, our ConScore-VG achieves consistent performance improvements across the RefCOCOs and reaches performance comparable to larger models fine-tuned with labeled data. On ScreenSpot-v2 (Wu et al., 2025), ConScore-VG achieves a remarkable 6.68% average accuracy improvement over the baseline and outperforms GUI-RCPO (Du et al., 2026). These results demonstrate the effectiveness and data efficiency of ConScore-VG across diverse grounding scenarios. The main contributions of this work are summarized as follows:

- We propose **ConScore-VG**, a consistency-score-guided test-time reinforcement learning framework that enables label-free adaptation for visual grounding.
- We introduce two complementary scores for reliable pseudo-label selection. The **Spatial Consistency Score** measures consensus among continuous coordinate predictions, while the **Semantic Consistency Score** mitigates the pseudo-label noise caused by relying on such consensus alone.
- Extensive experiments demonstrate the effectiveness of ConScore-VG across diverse visual grounding scenarios. Without labeled adaptation data, ConScore-VG achieves consistent gains on the RefCOCOs benchmarks with limited unlabeled test samples and further improves performance on ScreenSpot-v2 by 6.68%.

2 PRELIMINARY: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Group Relative Policy Optimization (GRPO), first introduced in DeepSeekMath (Shao et al., 2024), is a value-free variant of PPO that computes relative advantages by normalizing rewards within each response group. Specifically, in vision-language tasks, given an image-query pair (I, q), GRPO samples a group of G candidate responses $\{o_1, o_2, \ldots, o_G\}$ from the old policy model $\pi_{\theta_{\text{old}}}$. Each complete response o_i is assigned a scalar reward $r_i = R(a_i, a^*)$, where a_i denotes the answer extracted from the response o_i and a^* denotes the ground-truth annotation. GRPO then compares responses within the same group by normalizing their rewards to compute the relative advantage:

$$A_i = \frac{r_i - \text{mean}\{r_1, r_2, \ldots, r_G\}}{\text{std}\{r_1, r_2, \ldots, r_G\}}, \tag{1}$$

where A_i represents the advantage of the candidate response o_i relative to the other responses in the sampled group.

Based on these advantages, GRPO updates the policy π_θ with the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\left[\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q,I)\right]}$$
$$\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min(\phi_{i,t}(\theta)A_i,\ \text{clip}(\phi_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\, A_i) - \beta\mathbb{D}_{KL}[\pi_\theta \,\|\, \pi_{\text{ref}}]\right\}\right], \tag{2}$$
$$\text{with } \phi_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, I, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, I, o_{i,<t})},$$

where θ denotes the parameters of the policy model, ε is the clipping parameter used to limit overly large policy updates and β is a hyperparameter controlling the KL penalty strength. This objective encourages the model to assign higher probability to responses with larger advantages within the group, while constraining policy updates through clipping and KL regularization.

3 METHOD

In this section, we first formulate GRPO-based test-time reinforcement learning for visual grounding and identify reliable pseudo-label estimation from unlabeled test data as the central challenge (Section 3.1). We then introduce ConScore-VG, which evaluates candidate bounding boxes from complementary spatial and semantic perspectives to select reliable pseudo-labels for policy optimization (Section 3.2).

3.1 PROBLEM FORMULATION

Visual grounding aims to localize the image region referred to by a given language instruction. We formulate the task under GRPO-based policy optimization. Given an image I and a textual instruction τ, we first transform τ into a task query q using a predefined template, yielding an input instance $x = (I, q)$. For each input x, GRPO samples a group of G candidate responses from the old policy model $\pi_{\theta_{\text{old}}}$.

$$o_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid I, q), \qquad i = 1, \ldots, G. \tag{3}$$

The corresponding grounding predictions are then extracted from the sampled responses using a box parser f_{box}.

$$\mathcal{B} = \{b_i = f_{\text{box}}(o_i)\}_{i=1}^{G}. \tag{4}$$

GRPO requires a scalar reward for each sampled response o_i to compute group-relative advantages and optimize the policy. In the standard supervised setting, each input is paired with a ground-truth bounding box b_{gt}, allowing the reward for each response to be computed as the Intersection over Union (IoU) between its predicted box b_i and b_{gt}, i.e., $r_i^{\text{gt}} = \text{IoU}(b_i, b_{\text{gt}})$.

In test-time reinforcement learning, however, the model is adapted on unlabeled test data, making b_{gt} unavailable during optimization. The central challenge is therefore to construct reliable reward signals using only the image I, the query q, and the prediction set $\mathcal{B}$. A natural solution is to estimate a **pseudo-label** $\hat{b}$ as a surrogate localization target and compute the reward as $r_i = \text{IoU}(b_i, \hat{b})$. **This formulation enables GRPO-based optimization for visual grounding without requiring ground-truth annotations**.

3.2 CONSCORE-VG

We propose ConScore-VG, a consistency-guided TTRL framework for adapting vision-language models to visual grounding with unlabeled test data. As shown in Figure 2, ConScore-VG evaluates each candidate bounding box using a Spatial Consistency Score (Section 3.2.1) and a Semantic Consistency Score (Section 3.2.2). For each input, it combines these scores to select a pseudo-label, which is used to compute IoU-based rewards that guide policy optimization with GRPO (Section 3.2.3). Additional method details and PyTorch-style pseudocode are provided in Appendix B.

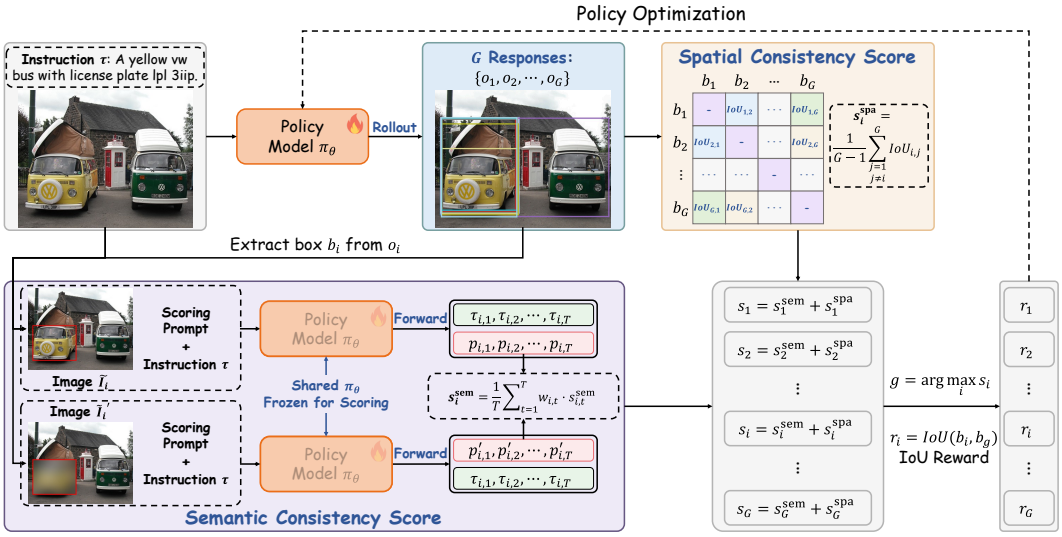


Figure 2: Overview of ConScore-VG. Given an input image I and a grounding instruction τ, we first sample a group of responses from the policy model. For each predicted box b_i, we compute a Spatial Consistency Score s_i^{spa} and a Semantic Consistency Score s_i^{sem}. The former measures the agreement between b_i and the overall spatial consensus, while the latter measures the semantic consistency between the visual content within b_i and τ based on token-level probability changes induced by perturbing that region. Finally, we select the prediction with the highest combined consistency score as the pseudo-label and use it to construct rewards for policy optimization.

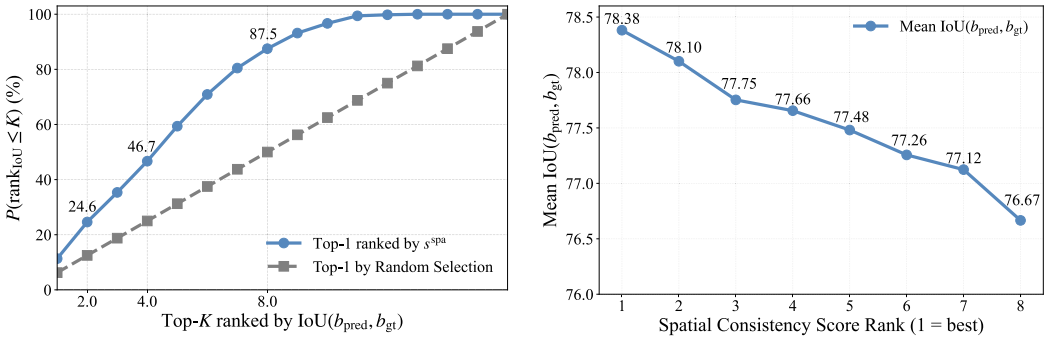
3.2.1 SPATIAL CONSISTENCY SCORE VIA SPATIAL CONSENSUS

Ground-truth bounding boxes are unavailable on unlabeled test data, precluding direct evaluation of localization accuracy. Inspired by TTRL methods that derive pseudo-labels from agreement among the model's own predictions, we assess each predicted bounding box for consistency with the spatial consensus formed by multiple rollouts. Specifically, for each input, we sample G responses from the policy and extract the corresponding grounding predictions $\mathcal{B} = \{b_i\}_{i=1}^{G}$. We define the Spatial Consistency Score of the i-th prediction as its mean IoU with the remaining predictions:

$$s_i^{\text{spa}} = \frac{1}{G-1} \sum_{j=1, j\neq i}^{G} \text{IoU}(b_i, b_j). \tag{5}$$

This score measures the consistency of each predicted bounding box with the spatial consensus reflected in the sampled prediction distribution. Predictions with greater average overlap receive higher scores, while spatially isolated predictions receive lower scores. This formulation extends consensus-based scoring from discrete answers to continuous bounding box predictions.

We further **assess whether the Spatial Consistency Score reliably reflects localization quality, as measured by the ground-truth IoU** $\text{IoU}(b_i, b_{\text{gt}})$. We first compare the cumulative probability that the pseudo-label selected using s_i^{spa} falls within the top-K sampled predictions ranked by the ground-truth IoU with the corresponding probability under random selection. As shown in Figure 3(a), score-based selection achieves substantially higher top-K inclusion probabilities than random selection, indicating that it preferentially selects predictions with higher ground-truth IoU. We then rank the sampled predictions by s_i^{spa} and compute the mean ground-truth IoU at each of the top eight rank positions. Figure 3(b) shows that higher-ranked predictions consistently attain greater



(a) Top-K Recall of Spatial-Score-Based Pseudo-Label Selection

(b) Mean Ground-Truth IoU by Spatial Score Rank

Figure 3: Relationship between the Spatial Consistency Score and localization quality. (**a**) Cumulative probability that the pseudo-label selected by s_i^{spa} is among the top-K predictions ranked by IoU with the ground-truth box, compared with random selection. (**b**) Mean ground-truth IoU of predictions at each spatial-consistency rank. Higher spatial consistency is consistently associated with better localization quality.

mean IoU. Together, these results demonstrate that stronger consistency with the policy's spatial consensus is associated with better localization quality, supporting the Spatial Consistency Score as a reliable criterion for pseudo-label selection.

3.2.2 Semantic Consistency Score via Visual Information Gain

Although the Spatial Consistency Score favors predictions with higher ground-truth IoU, Figure 3 shows that the selected pseudo-labels can still be inaccurate. Relying solely on prediction consensus to construct pseudo-labels can introduce substantial noise, as consistency does not guarantee correctness (He et al., 2026; Lin et al., 2026). In particular, multiple sampled predictions may exhibit strong spatial agreement around a region that is semantically inconsistent with the language instruction, causing consensus-based selection to reinforce an incorrect localization. Inspired by the generation-verification asymmetry discussed by He et al. (2026), we introduce an additional self-verification signal by leveraging visual referring as an inverse task of visual grounding, in which the model is asked to generate a textual description of a given candidate region. Building on this inverse formulation, we introduce the Semantic Consistency Score to evaluate the region's semantic consistency with the language instruction, thereby complementing the spatial consensus signal.

Specifically, given a candidate bounding box b_i, we construct a scoring prompt ψ that asks the policy model to describe the visual content within the box (see Appendix F for details). Inspired by Wang et al. (2026c), we pair an unperturbed visual input with a perturbed counterpart to assess how much the visual evidence within b_i contributes to the likelihood of the original grounding instruction τ. We first mark b_i on the original image I to obtain the unperturbed visual input $\tilde{I}_i$ and then create its perturbed counterpart $\tilde{I}_i'$ by corrupting the visual content inside b_i while leaving the rest of the image unchanged. The perturbation procedure is detailed in Appendix B.1. Both visual inputs are paired with the same scoring prompt ψ, and the original grounding instruction is used as the shared teacher-forced target sequence, denoted by $\tau_i = \tau = \{\tau_{i,t}\}_{t=1}^{T}$, where $T = |\tau_i|$. With the textual context and target sequence fixed, forward passes through the policy model yield two sequences of target-token probabilities, $p_i = \{p_{i,t}\}_{t=1}^{T}$ and $p_i' = \{p_{i,t}'\}_{t=1}^{T}$, corresponding to $\tilde{I}_i$ and $\tilde{I}_i'$, respectively.

$$p_{i,t} = \pi_\theta(\tau_{i,t} \mid \tilde{I}_i, \psi, \tau_{i,<t}), \tag{6}$$

$$p_{i,t}' = \pi_\theta(\tau_{i,t} \mid \tilde{I}_i', \psi, \tau_{i,<t}). \tag{7}$$

We then define the token-level visual information gain as $c_{i,t} = \log p_{i,t} - \log p_{i,t}'$. From an information-gain perspective (Shannon, 1948), $c_{i,t}$ measures how much the likelihood of token $\tau_{i,t}$ changes when the visual content inside b_i is perturbed. A larger $c_{i,t}$ indicates stronger visual support for the token, suggesting greater semantic consistency between the candidate region and the corresponding instruction token. We then map $c_{i,t}$ to a bounded interval to obtain the token-level semantic consistency score, defined as

$$s_{i,t}^{\text{sem}} = \tanh\left(\frac{c_{i,t}}{2}\right). \tag{8}$$

However, when both $p_{i,t}$ and $p'_{i,t}$ are extremely small, a large log-probability difference may reflect a substantial relative change despite low confidence in the target token under both inputs. To suppress such unreliable token-level signals, we first compute the combined probability mass in log space and set an adaptive threshold at the ρ-quantile of the resulting values within each sequence. We then compute the token-level confidence weight $w_{i,t}$ using the sigmoid function $\sigma(\cdot)$.

$$\ell_{i,t} = \log(p_{i,t} + p'_{i,t}), \tag{9}$$

$$\gamma_i = \text{Quantile}_\rho(\{\ell_{i,t}\}_{t=1}^{T}), \tag{10}$$

$$w_{i,t} = \sigma(\ell_{i,t} - \gamma_i). \tag{11}$$

Finally, we compute the Semantic Consistency Score for b_i by first weighting each token-level score with $w_{i,t}$ and then averaging across all tokens.

$$s_i^{\text{sem}} = \frac{1}{T} \sum_{t=1}^{T} w_{i,t} \cdot s_{i,t}^{\text{sem}}. \tag{12}$$

The quantile threshold is computed separately for each candidate region, allowing the confidence gate to adapt to the distribution of token-level confidence values within its sequence instead of relying on a fixed global threshold. With the original grounding instruction serving as the shared teacher-forced target for both visual inputs, s_i^{sem} measures how strongly the visual content within b_i supports the grounding instruction, rather than the region's visual saliency alone.

We further **analyze the relationship between the Semantic Consistency Score and localization quality by controlling the IoU of candidate boxes with the ground truth**. For each sample, we construct candidate boxes b_μ at predefined IoU levels satisfying $\text{IoU}(b_\mu, b_{\text{gt}}) \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and compute their corresponding Semantic Consistency Scores s_μ^{sem}. After normalizing the scores, we examine the frequency of candidates at each IoU level within each score interval and compute the mean normalized score for each IoU level. Figure 4 shows that higher score intervals contain a greater proportion of high-IoU candidates and that the mean normalized score increases consistently with IoU. These results indicate a positive association between semantic consistency and localization quality, supporting the complementary use of s^{sem} with the Spatial Consistency Score for identifying reliable pseudo-label candidates.

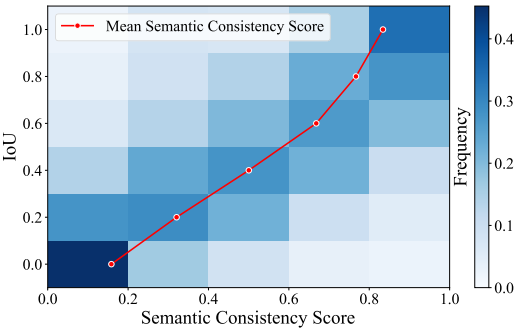


Figure 4: Relationship between the Semantic Consistency Score and localization quality. The heatmap shows the frequency distribution of candidate boxes across IoU levels within each Semantic Consistency Score interval, while the red curve reports the mean score at each IoU level. Higher scores are consistently associated with higher-IoU predictions.

3.2.3 Pseudo-label Construction and Policy Optimization

We combine the Spatial Consistency Score and the Semantic Consistency Score with equal weights to select a pseudo-label that accounts for both spatial consensus and semantic alignment. Specifically, we select the candidate box with the highest combined score as the pseudo-label $\hat{b}$.

$$s_i = s_i^{\text{sem}} + s_i^{\text{spa}}, \quad g = \arg\max_i s_i, \quad \hat{b} = b_g. \tag{13}$$

Using $\hat{b}$ as the reference, we assign each rollout an IoU-based reward $r_i = \text{IoU}(b_i, \hat{b})$. We then compute group-relative advantages from these rewards according to Eq. 1 and update the policy using the GRPO objective defined in Eq. 2.

4 Experiments

We first describe the experimental setup in Section 4.1. Section 4.2 then reports the main results on both general visual grounding and GUI grounding, together with a comparison against the TTRL-

based method GUI-RCPO (Du et al., 2026) on the former. We further ablate the two consistency scores in Section 4.3. Additional experimental results are provided in Appendix E.

4.1 Experimental Setup

Implementation Details. We build ConScore-VG on the VLM-R1 (Shen et al., 2025) framework, using Qwen2.5-VL-3B (Bai et al., 2025a) as the base model. For test-time adaptation on general visual grounding, we randomly sample $N = 256$ unlabeled instances from the test dataset, whereas for GUI grounding we follow the data setting of Du et al. (2026). The policy generates $G = 16$ rollouts per input with a sampling temperature of 1.0, and the quantile parameter ρ of the semantic consistency score is fixed to 0.25 across all datasets. Unless otherwise specified, the model is adapted for 3 epochs with a global batch size of 8, a learning rate of 1×10^{-6}, and a KL coefficient of $\beta = 0.04$. All experiments are conducted on 4 NVIDIA A100 GPUs (80 GB) with DeepSpeed (Rasley et al., 2020) for efficient distributed training. Additional details are provided in Appendix D.

Benchmarks. We evaluate ConScore-VG on benchmarks for both general visual grounding and GUI grounding. For general visual grounding, we use the RefCOCOs (Yu et al., 2016; Mao et al., 2016) benchmarks, comprising RefCOCO, RefCOCO+, and RefCOCOg. To assess its applicability to GUI grounding, we further evaluate ConScore-VG on ScreenSpot (Cheng et al., 2024a), ScreenSpot-v2 (Wu et al., 2025), and ScreenSpot-Pro (Li et al., 2025a), which cover diverse interface layouts and grounding scenarios.

Evaluation Metrics. For general visual grounding, we report Accuracy@0.5, defined as the percentage of samples for which the predicted bounding box has an IoU greater than 0.5 with the ground-truth box. For GUI grounding, we follow standard evaluation protocols and report grounding accuracy, defined as the proportion of samples for which the center of the predicted bounding box lies within the corresponding ground-truth box (Cheng et al., 2024a; Du et al., 2026).

4.2 Main Results

Table 1: Main results on the RefCOCOs. ConScore performs selection without updating the policy model, while ConScore-GRPO adapts on 256 unlabeled samples. ConScore-GRPO performs comparably to the larger VisionReasoner-7B trained with labeled GRPO. * denotes results reproduced using the provided model weights. **Bold**: best, underline: second best.

Method	RefCOCO		RefCOCO+		RefCOCOg		Sum
	val	testA	val	testA	val	test	
VGTR (Du et al., 2022)	79.3	82.2	64.4	70.9	66.8	67.3	430.9
TransVG (Deng et al., 2021)	81.0	82.7	64.8	70.7	68.7	67.7	435.6
RefTR (Li & Sigal, 2021)	85.7	88.7	77.6	82.3	79.3	80.0	493.6
MDETR (Kamath et al., 2021)	86.8	89.6	79.5	84.1	81.6	80.9	502.5
GLIP-T (Li et al., 2022)	50.4	54.3	49.5	52.8	66.1	66.9	340.0
G-DINO-T (Liu et al., 2024)	89.2	91.9	81.1	87.4	84.2	84.9	518.7
DQ-DETR (Liu et al., 2023)	88.6	91.0	81.7	86.2	82.8	83.4	513.7
Shikra-7B (Chen et al., 2023)	87.0	90.6	81.6	87.4	82.3	82.2	511.1
InternVL2-8B (Chen et al., 2024)	87.1	91.1	79.8	87.9	82.7	82.7	511.3
VisionReasoner-7B (Liu et al., 2025c)	89.1	91.0	**85.0**	87.6	**87.6**	**88.5**	**528.8**
Qwen2.5-VL-3B*	89.2	92.1	82.9	88.7	85.1	85.5	523.5
w/ ConScore	89.5	92.2	83.3	89.0	85.6	85.6	525.2 ↑+1.7
w/ ConScore-GRPO	**90.0**	**92.4**	83.9	**89.4**	86.1	86.0	527.8 ↑+4.3

4.2.1 Results on General Visual Grounding

Referring Expression Comprehension (REC) is a representative task in general visual grounding, with the RefCOCOs serving as widely used benchmarks. We evaluate ConScore-VG on the RefCOCOs benchmarks against a broad range of methods, including specialized grounding models such as VGTR (Du et al., 2022), TransVG (Deng et al., 2021), RefTR (Li & Sigal, 2021), MDETR (Kamath et al., 2021), GLIP (Li et al., 2022), GroundingDINO (Liu et al., 2024), and DQ-DETR (Liu

et al., 2023), as well as methods based on multimodal large language models (MLLMs), such as Shikra (Chen et al., 2023), InternVL2-8B (Chen et al., 2024), Qwen2.5-VL-3B (Bai et al., 2025a), and VisionReasoner (Liu et al., 2025c). We further evaluate two complementary test-time variants of ConScore-VG, namely ConScore and ConScore-GRPO. ConScore samples multiple responses at inference time and directly selects the highest-scoring box as the final prediction without parameter updates, whereas ConScore-GRPO uses the selected box as a pseudo-label to compute IoU-based rewards for all sampled boxes and performs GRPO-based test-time adaptation.

The results in Table 1 show that ConScore improves grounding performance over greedy decoding without updating model parameters, supporting the use of joint spatial and semantic consistency for candidate selection. ConScore-GRPO further improves performance through GRPO-based policy optimization on only 256 unlabeled samples using ConScore-selected pseudo-labels, exceeding Qwen2.5-VL-3B by 4.3 points under greedy decoding. It also achieves performance comparable to the larger VisionReasoner-7B, which is optimized through GRPO using labeled data. Overall, these results demonstrate the effectiveness and data efficiency of ConScore-VG for test-time adaptation, achieving competitive grounding performance without relying on labeled data.

4.2.2 Results on GUI Grounding

GUI grounding is a specialized visual grounding task that enables agents to interact with digital interfaces through natural language instructions. Whereas general visual grounding is commonly evaluated using IoU-based localization accuracy, GUI grounding typically considers a prediction correct when the center of the predicted box falls within the ground-truth box. To evaluate the applicability of ConScore-VG to GUI grounding, we compare it with the recent TTRL-based method GUI-RCPO (Du et al., 2026) on ScreenSpot, ScreenSpot-v2, and ScreenSpot-Pro. Following GUI-RCPO, we perform test-time reinforcement learning on ScreenSpot-v2 with a sampling temperature of 1.0 while retaining the remaining training hyperparameters under a 4-GPU configuration. Because ConScore-VG selects pseudo-labels without defining a task-specific reward function, we adopt the reward formulation of GUI-G1 (Zhou et al., 2025) to derive rewards from the selected pseudo-labels, as detailed in Appendix C. We also design a task-specific semantic consistency prompt ψ for GUI grounding, as provided in Appendix F.2.

As shown in Table 2, ConScore-VG improves over the base model by 6.68% on ScreenSpot-v2 and also consistently outperforms GUI-RCPO, with accuracy gains of 1.41%, 1.65%, and 6.51% on ScreenSpot, ScreenSpot-v2, and ScreenSpot-Pro, respectively. The consistent improvements, particularly the substantial gain on ScreenSpot-Pro, support the effectiveness of combining spatial and semantic consistency for reliable pseudo-label selection in GUI grounding. These results also demonstrate the robustness of ConScore-VG across different visual grounding benchmarks and interface scenarios, supporting efficient adaptation to diverse downstream visual grounding tasks.

Table 2: Comparison of ConScore-VG and GUI-RCPO in GUI Grounding Accuracy on ScreenSpot-v2 (**SSv2**), ScreenSpot (**SSv1**), and ScreenSpot-Pro (**SSPro**). **Bold**: best, underline: second best.

Method	**SSv2 Mobile**		**SSv2 Desktop**		**SSv2 Web**		**SSv2 avg**	**SSv1 avg**	**SSPro avg**
	Text	Icon	Text	Icon	Text	Icon			
Qwen2.5-VL-3B	97.67	75.92	85.56	59.52	84.64	65.82	80.11	76.97	20.18
GUI-RCPO	98.06	**81.68**	91.11	65.08	90.71	**73.42**	85.14	82.47	24.67
ConScore-VG	**98.97**	79.62	**95.36**	**74.29**	**91.03**	72.41	**86.79**	**83.88**	**31.18**

4.2.3 Comparison with GUI-RCPO on General Visual Grounding

GUI-RCPO (Du et al., 2026) is the TTRL-based GUI grounding method most closely related to ConScore-VG. To provide further evidence of the effectiveness of ConScore-VG in general visual grounding, we compare it with GUI-RCPO under the same training setup on the validation splits of the RefCOCOs.

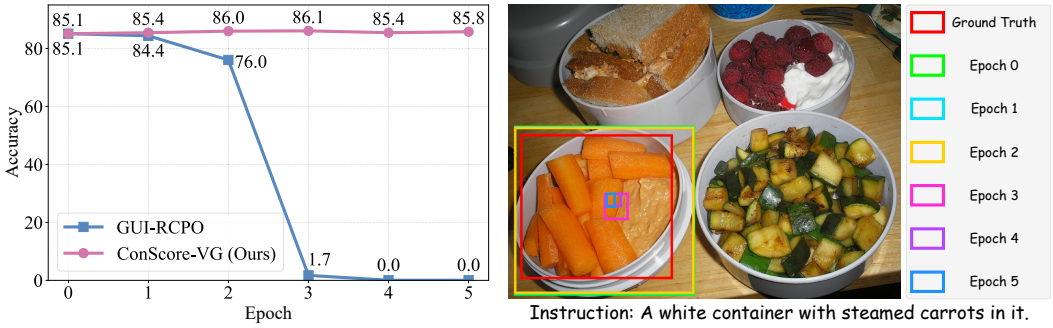
Table 3: Comparison between GUI-RCPO and ConScore-VG on General Visual Grounding.

Method	**RefCOCO**	**RefCOCO+**	**RefCOCOg**
Base	89.2	82.9	85.1
w/ GUI-RCPO	18.9	14.5	1.7
w/ ConScore-VG	**90.0**	**83.9**	**86.1**

As shown in Table 3, directly optimizing the policy with the GUI-RCPO reward substantially degrades performance on general visual grounding. This degradation may stem from a mismatch

between the GUI-RCPO reward design and the objective of general visual grounding. By rewarding overlap with intersection regions across multiple rollouts, GUI-RCPO may favor overly compact boxes that satisfy the center-based criterion of GUI grounding but fail to capture the full target extent required for general visual grounding. In contrast, ConScore-VG jointly considers spatial and semantic consistency, making it better suited to the requirements of general visual grounding.

To further investigate the limitations of the GUI-RCPO reward for general visual grounding, we track the performance of ConScore-VG and GUI-RCPO across training epochs on the RefCOCOg validation split and visualize the evolution of GUI-RCPO predictions, as shown in Figure 5. As training progresses, GUI-RCPO rapidly collapses from 85.1% to nearly 0.0% accuracy, with its predicted boxes progressively shrinking toward compact regions. These results show that the GUI-RCPO reward is misaligned with the general visual grounding objective, progressively favoring overly compact boxes and thereby degrading localization performance.



(a) Accuracy across Training Epochs

(b) Grounding Prediction Evolution of GUI-RCPO across Epochs

Figure 5: Visualization of the limitations of the GUI-RCPO reward in general visual grounding. (**a**) Accuracy of GUI-RCPO and ConScore-VG across training epochs on the RefCOCOg validation split. (**b**) Evolution of GUI-RCPO grounding predictions during training.

4.3 Ablation Study on Semantic and Spatial Consistency Scores

Table 4: Ablation of Semantic and Spatial Consistency Scores. Combining both scores achieves the best performance, confirming the effectiveness of our design.

Method	Semantic score	Spatial score	RefCOCO	RefCOCO+	RefCOCOg
Base			89.2	82.9	85.1
Semantic only	✓		89.4	83.7	85.4
Spatial only		✓	89.7	83.4	85.9
ConScore-VG	✓	✓	**90.0**	**83.9**	**86.1**

We conduct ablation studies on the validation splits of the RefCOCOs benchmarks to assess the contributions of the Spatial Consistency Score s^{spa} and the Semantic Consistency Score s^{sem}. As shown in Table 4, using either score alone consistently improves over the baseline. The stronger overall performance of s^{spa} suggests that agreement with the spatial consensus among sampled predictions provides an informative signal for identifying reliable candidates. Combining both scores in the full ConScore-VG achieves the best performance, highlighting their complementary contributions to reliable pseudo-label selection.

5 Conclusion

This work investigates test-time reinforcement learning for visual grounding, where continuous coordinate outputs make majority-vote-based pseudo-label selection challenging, while spatial consensus alone can reinforce semantically incorrect predictions. To address these challenges, we introduce ConScore-VG, which jointly leverages spatial and semantic consistency to select reliable pseudo-labels for policy optimization from unlabeled test data. Experiments across general visual grounding and GUI grounding benchmarks show that ConScore-VG consistently improves performance using only a limited number of unlabeled samples and generalizes effectively across datasets and tasks. These findings demonstrate the potential of consistency-guided TTRL as a data-efficient framework for adapting vision-language models to visual grounding without labeled supervision.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, including grammatical correction, sentence refinement, and improvements to readability. We also used generative AI tools to assist in refining the wording of task-specific prompts for the Semantic Consistency Score across different visual grounding tasks and to support minor code modifications. We did not use generative AI tools to develop the research methodology, design experiments, analyze or interpret experimental results, or derive scientific conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work develops a test-time reinforcement learning framework for visual grounding based on spatial and semantic consistency. All training data are derived from previously released datasets that are properly cited, no new data are collected, and no human subjects are involved. The proposed method is intended solely for visual grounding research and is not designed to generate harmful content or support malicious use. We do not identify specific concerns related to discrimination, bias, fairness, privacy, or security within the scope of this work. We follow standard research practices for data use, attribution, and experimental reproducibility.

REPRODUCIBILITY STATEMENT

The main text and appendix provide the implementation details and experimental settings necessary for reproducing our results. The complete source code is included in the supplementary material, and we will additionally release the training data, experimental scripts, and model checkpoints to support full reproducibility.

REFERENCES

Shuai Bai, Keqin 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, 2025a. URL https://arxiv.org/abs/2502.13923.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025b.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9313–9332, 2024a.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911. IEEE, 2024b.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1769–1779, 2021.

Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2022.

Yong Du, Yuchen Yan, Fei Tang, Zhengxi Lu, Chang Zong, Weiming Lu, Shengpei Jiang, and Yongliang Shen. Test-time reinforcement learning for gui grounding via region consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30593–30601, 2026.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, et al. How far can unsupervised rlvr scale llm training? In *International Conference on Learning Representations*, volume 2026, pp. 14823–14865, 2026.

Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. Roboground: Robotic manipulation with grounded vision-language priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22540–22550. IEEE, 2025a.

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. SAM-r1: Leveraging SAM for reward feedback in multimodal segmentation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=dHOSTp8MBl.

Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 01–10, 2023. doi: 10.1109/CVPR52729.2023.00263.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8778–8786, 2025a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

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

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9845–9867, 2025b.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6657–6678, 2024.

Zhuoling Li, LiangLiang Ren, Jinrong Yang, Yong Zhao, Xiaoyang Wu, Zhenhua Xu, Xiang Bai, and Hengshuang Zhao. VIP: Vision instructed pre-training for robotic manipulation. In *Forty-second International Conference on Machine Learning*, 2025c. URL https://openreview.net/forum?id=ccUNMIbpcf.

Haicheng Liao, Huanming Shen, Bonan Wang, Yongkang Li, Yihong Tang, Chengyue Wang, Dingyi Zhuang, Kehua Chen, Hai Yang, Chengzhong Xu, and Zhenning Li. Think before you drive: World model-inspired multimodal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3677–3687, June 2026.

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.

Changyu Liu, Yiyang Liu, Taowen Wang, Qiao Zhuang, James Chenhao Liang, Wenhao Yang, Renjing Xu, Qifan Wang, Dongfang Liu, and Cheng Han. On-the-fly vla adaptation via test-time reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40107–40125, 2026a.

Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1728–1736, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, and Tieniu Tan. Rethinking the role of prompting strategies in llm test-time scaling: A perspective of probability theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27962–27994, 2025a.

Yizhou Liu, Fei Tang, Yuchen Yan, Zhengxi Lu, Songqin Nong, Tao Jiang, Wenhao Xu, Wenqi Zhang, Weiming Lu, Jun Xiao, et al. Learning from reliable negatives: Confidence-anchored test-time adaptation for gui grounding. In *European Conference on Computer Vision*, pp. 227–244. Springer, 2026b.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv e-prints*, pp. arXiv–2505, 2025c.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044. IEEE, 2025d.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

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.

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.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogério 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.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5198–5215, 2022.

Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2512–2522. IEEE, 2025.

Fei Tang, Zhangxuan Gu, Zhengxi Lu, Xuyang Liu, Shuheng Shen, Changhua Meng, Wen Wang, Wenqi Zhang, Yongliang Shen, Weiming Lu, et al. Gui-g^2: Gaussian reward modeling for gui grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33214–33222, 2026.

Aozhe Wang, Zhengxi Lu, Jianze Wang, Shangke Lv, Ying Liu, Weiming Lu, Jun Xiao, Yueting Zhuang, Hua Yang, Qianglong Chen, and Yongliang Shen. Ttpo: Test-time policy optimization, 2026a. URL https://arxiv.org/abs/2608.27448.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, et al. Vgr: Visual grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026b.

Song Wang, Gongfan Fang, Lingdong Kong, Xiangtai Li, Jianyun Xu, Sheng Yang, Qiang Li, Jianke Zhu, and Xinchao Wang. Pixelthink: Towards efficient chain-of-pixel reasoning. *arXiv preprint arXiv:2505.23727*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 80502–80530, 2026c.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33577–33587, June 2026.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. OS-ATLAS: Foundation action model for generalist GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=n9PDaFNi8t.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pp. 57153–57180, 2024.

En Yu, Kangheng Lin, Liang Zhao, jisheng yin, Yana Wei, 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BeXcXrXetA.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Zhaoning Yu, Zhaolun Su, Leitian Tao, Haozhu Wang, Aashu Singh, Hanchao Yu, Jianyu Wang, Hongyang Gao, Weizhe Yuan, Jason E Weston, Ping Yu, and Jing Xu. RESTRAIN: From spurious votes to signals — self-training RL with self-penalization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=87ySF7viys.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, LeiZhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Zhenguo Zhang, Haohan Zheng, Yishen Wang, Le Xu, Tianchen Deng, Xuefeng Chen, Qu Chen, Bo Zhang, and Wuxiong Huang. Omnidrive-r1: Reinforcement-driven interleaved multi-modal chain-of-thought for trustworthy vision-language autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1106–1116, June 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

Yuqi Zhou, Sunhao Dai, Shuai Wang, Kaiwen Zhou, Qinglin Jia, and Jun Xu. GUI-g1: Understanding r1-zero-like training for visual grounding in GUI agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=1XLjrmKZ4p.

Xiuyuan Zhu, Ke Lu, Hao Wu, Zijin Du, Dongming Zhang, and Jian Xue. Ioupd: Iou-aware privileged distillation for visual grounding with multimodal large language models. *arXiv preprint arXiv:2607.15732*, 2026.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VuVhgEiu20.

Appendix **15**

A Discussion **15**

A.1 Related Work 15

A.2 Limitations and Future Work 16

B Additional Method Details **16**

B.1 Visual Perturbation for Semantic Consistency Scoring 16

B.2 PyTorch-style Pseudocode for ConScore-VG 17

C Reward Functions Adopted for GUI Grounding **18**

D Detailed Training Configurations **18**

E More Experiments **19**

E.1 Cross-Dataset Generalization 19

E.2 Ablation Study on Training Epochs 19

E.3 Ablation Study on the Number of Adaptation Samples 20

E.4 Qualitative Comparison for GUI Grounding 20

E.5 Additional Qualitative Results 20

F Prompt for Semantic Consistency Scoring **20**

F.1 Scoring Prompt for General Visual Grounding 23

F.2 Scoring Prompt for GUI Grounding 23

A Discussion

A.1 Related Work

Visual Grounding. Visual grounding aims to localize a target region in an image conditioned on a textual instruction. As a fundamental capability of vision-language models, it underpins a broad range of downstream applications (Huang et al., 2025a; Zhang et al., 2026; Wang et al., 2026b; Tang et al., 2026). Early approaches (Subramanian et al., 2022; Li et al., 2022; Jin et al., 2023; Cheng et al., 2024b) build on pretrained vision-language models such as CLIP (Radford et al., 2021) to align candidate regions with textual queries. With the rapid development of multimodal large language models (MLLMs), recent studies leverage their enhanced cross-modal understanding to handle more complex grounding instructions, typically through supervised fine-tuning (SFT) (Zhao et al., 2023; Zhang et al., 2024; Li et al., 2024; Lai et al., 2024; Li et al., 2025b; Sun et al., 2025; Zhu et al., 2026). However, SFT-based methods remain heavily dependent on large-scale annotated data and often exhibit limited generalization to out-of-distribution data. To alleviate this limitation, a growing body of work adopts reinforcement learning for visual grounding, achieving substantial gains with limited training data while preserving generalization (Liu et al., 2025d; Bai et al., 2025b; Liu et al., 2025c; Cao et al., 2025; Wang et al., 2025; Huang et al., 2025b; Yu et al., 2025). For instance, Shen et al. (2025) introduced VLM-R1, which employs rule-based reinforcement learning with verifiable grounding rewards, while Liu et al. (2025b) extended this framework from box-level to pixel-level grounding with Seg-Zero.

Test-Time Reinforcement Learning. Existing SFT- and RL-based methods primarily optimize models at training time and rely on annotated data for supervision or verifiable rewards, hindering

adaptation to new domains where annotations are difficult or costly to obtain. Test-time scaling offers an orthogonal direction by allocating additional computation at inference time, with methods such as self-consistency aggregating multiple sampled predictions without updating model parameters (Wang et al., 2023; Liu et al., 2025a). Building on this paradigm, Test-Time Reinforcement Learning (TTRL) integrates test-time adaptation with reinforcement learning, enabling policy optimization directly on unlabeled test data using self-generated reward signals (Zuo et al., 2025; He et al., 2026; Yu et al., 2026; Lin et al., 2026; Singh et al., 2026; Liu et al., 2026a; Wang et al., 2026a; Liu et al., 2026b). For instance, Zuo et al. (2025) first formulated the TTRL setting by constructing pseudo-labels through majority voting and demonstrated its effectiveness on unlabeled test data. Singh et al. (2026) extended this paradigm to VLMs by combining frequency-based rewards with entropy-based diversity control. Subsequent studies by He et al. (2026) and Lin et al. (2026) identified a key limitation of consensus-based supervision, showing that model agreement does not necessarily imply correctness and can instead reinforce erroneous predictions. To mitigate the resulting pseudo-label noise, Liu et al. (2026b) identified trajectories inconsistent with the pseudo-label as relatively reliable negative samples and proposed applying GRPO only to these negative trajectories. In parallel, Wang et al. (2026a) proposed TTPO, which asymmetrically optimizes sampled trajectories by distilling pseudo-label-consistent trajectories through On-Policy Self-Distillation (OPSD) (Zhao et al., 2026) while penalizing inconsistent ones with GRPO. This design reduces the risk of an erroneous pseudo-label uniformly steering all trajectories toward the same incorrect prediction. Among existing TTRL methods, GUI-RCPO (Du et al., 2026) is most closely related to our work. It extends TTRL to GUI grounding by deriving self-supervised rewards from region consistency across sampled predictions. However, GUI-RCPO is specifically designed for GUI grounding, and its region-consistency reward can bias optimization toward overly compact predictions. More importantly, relying solely on spatial consensus makes it difficult to distinguish spatially consistent but incorrect predictions from genuinely correct ones. To address these limitations, we propose ConScore-VG, a TTRL framework designed for broader visual grounding tasks, which jointly leverages spatial consistency across sampled predictions and semantic consistency with the grounding instruction for more reliable pseudo-label selection and reward estimation.

A.2 Limitations and Future Work

Limitations. This work presents an initial study of consistency-guided test-time reinforcement learning for visual grounding. Despite the promising results, two limitations remain. First, ConScore-VG currently focuses on single-image grounding with bounding-box outputs, while its extension to multi-image grounding, referring expression segmentation, and video grounding remains unexplored. Second, the Semantic Consistency Score relies on task-specific scoring prompts, making adaptation to new visual grounding settings dependent on manual prompt design and additional engineering effort.

Future Work. Future work will explore two directions.

- **Extension to Broader Visual Grounding Settings.** We plan to extend ConScore-VG to multi-image grounding, referring expression segmentation, and video grounding by adapting the spatial and semantic consistency mechanisms to these more complex output spaces.
- **Automated Scoring Prompt Design.** We aim to reduce reliance on manual prompt design by automatically generating or optimizing task-specific scoring prompts, thereby facilitating adaptation to diverse visual grounding settings.

B Additional Method Details

B.1 Visual Perturbation for Semantic Consistency Scoring

To score semantic consistency, we construct paired visual inputs that differ only in the visual evidence within each candidate region. Specifically, for the i-th response trajectory o_i, we first extract its predicted bounding box $b_i = [x_1, y_1, x_2, y_2]$ and mark the box on the original image I to obtain $\tilde{I}_i$. We then construct its perturbed counterpart $\tilde{I}'_i$ by modifying only the visual content within b_i while keeping the rest of the image unchanged. During implementation, we define the set of pixels covered by b_i as $\Omega_{b_i} = \{(x, y) \mid x_1 \le x \le x_2,\ y_1 \le y \le y_2\}$ and compute the mean pixel value of

each channel within this region as:

$$\mu_{i,c} = \frac{1}{|\Omega_{b_i}|} \sum_{(x,y)\in\Omega_{b_i}} I(x, y, c). \tag{14}$$

For each channel c, we replace the corresponding pixel values within b_i with $\mu_{i,c}$ while leaving all pixels outside the region unchanged. This perturbation suppresses fine-grained visual evidence within the candidate region while preserving the surrounding context and overall image structure. The resulting pair $(\tilde{I}_i, \tilde{I}'_i)$ is subsequently used to score the semantic consistency between the predicted bounding box b_i and the grounding instruction τ.

B.2 PyTorch-style Pseudocode for ConScore-VG

As shown below, we provide PyTorch-style pseudocode for computing the Spatial Consistency Score and Semantic Consistency Score, together with the complete ConScore-VG optimization procedure for a single unlabeled test instance. ConScore-VG first samples a group of rollouts, evaluates their predicted bounding boxes using the two consistency scores, selects a pseudo-label, and then performs GRPO-based policy optimization.

Listing 1: PyTorch-style Pseudocode for ConScore-VG on a single unlabeled test sample.

```
def spatial_consistency_score(B, G):
    """Spatial consistency score of each box via mean IoU with the
    others."""
    # calculate IoU matrix [G, G]
    iou_matrix = box_iou(B, B)
    # calculate spatial consistency score
    s_spa = (iou_matrix.sum(dim=-1) - 1) / (G - 1)
    return s_spa

def semantic_consistency_score(b, tau, I, psi, model, rho=0.25):
    """Semantic consistency score of a box via visual information
    gain."""
    # mark the box on I
    I_box = draw_box(I, b)
    # perturb pixels inside b
    I_pert = mean_fill(I_box, b)
    # teacher-forced P(tau | I_box, psi)
    p = model.token_probs(I_box, psi, tau)
    # teacher-forced P(tau | I_pert, psi)
    p_pert = model.token_probs(I_pert, psi, tau)
    # token-level information gain
    c_gain = p.log() - p_pert.log()
    # token-level semantic score
    s_sem_tok = torch.tanh(c_gain / 2)
    # weight for token-level semantic consistency score
    ell = (p + p_pert).log()
    gamma = torch.quantile(ell, rho)
    w = torch.sigmoid(ell - gamma)
    # semantic consistency score
    s_sem = (s_sem_tok * w).mean()
    return s_sem

# ConScore-VG on one unlabeled test instance (I, tau)
# instruction -> task query
q = build_query(tau)
# group rollouts
O = old_policy.sample(I, q, G=16, temperature=1.0)
# extract predicted boxes, [G, 4]
B = torch.stack([parse_box(o) for o in O])
# spatial consistency score
s_spa = spatial_consistency_score(B, G)
# semantic consistency score
s_sem = torch.stack([
```

```
    semantic_consistency_score(b, tau, I, psi, policy) for b in B
])
# combined consistency score
s = s_spa + s_sem
# selected pseudo-label
b_hat = B[s.argmax()]
# IoU-based rewards
r = box_iou(B, b_hat.unsqueeze(0)).squeeze(-1)
# group-relative advantage
A = (r - r.mean()) / (r.std() + 1e-8)
loss = grpo_loss(policy, old_policy, O, A, beta=0.04)
loss.backward()
```

C Reward Functions Adopted for GUI Grounding

For the GUI grounding experiments in Section 4.2.2, we adopt the reward functions introduced in GUI-G1 (Zhou et al., 2025) without modification, including the hit-based reward r_{hit}, the IoU-based reward r_{IoU}, and the box-size regularization reward r_{box}. Notably, during unlabeled adaptation, we replace the ground-truth box b_{gt} in the original reward formulation with the selected pseudo-label $\hat{b}$. Accordingly, all reward terms are computed with respect to $\hat{b}$ rather than b_{gt}.

Hit-based Reward. The hit-based reward r_{hit} evaluates whether the center of the predicted bounding box $b_{\text{pred}} = [x_{p1}, y_{p1}, x_{p2}, y_{p2}]$ falls inside the ground-truth box b_{gt}. Specifically, the center of b_{pred} is computed as $(x_p, y_p) = \left(\frac{x_{p1}+x_{p2}}{2}, \frac{y_{p1}+y_{p2}}{2}\right)$, and the reward is defined as

$$r_{\text{hit}} = \begin{cases} 1 & \text{if } (x_p, y_p) \in b_{\text{gt}}, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

IoU-based Reward. The IoU-based reward r_{IoU} evaluates the spatial overlap between the predicted bounding box b_{pred} and the ground-truth box b_{gt}, and is defined as

$$r_{\text{IoU}} = \text{IoU}(b_{\text{pred}}, b_{\text{gt}}). \tag{16}$$

Box-size Regularization Reward. The box-size regularization reward r_{box} encourages the predicted bounding box to remain comparable in size to the ground-truth box, thereby mitigating degenerate predictions such as overly small boxes favored by r_{hit} and overly large boxes favored by r_{IoU}. Specifically, given the predicted box $b_{\text{pred}} = [x_{p1}, y_{p1}, x_{p2}, y_{p2}]$ and the ground-truth box $b_{\text{gt}} = [x_1, y_1, x_2, y_2]$, the reward r_{box} is defined as

$$r_{\text{box}} = \frac{4}{s_{x_1} + s_{y_1} + s_{x_2} + s_{y_2}}. \tag{17}$$

The four denominator terms are computed as

$$\begin{aligned} s_{x_1} &= \frac{1}{1 - |x_{p1} - x_1|/W}, & s_{y_1} &= \frac{1}{1 - |y_{p1} - y_1|/H}, \\ s_{x_2} &= \frac{1}{1 - |x_{p2} - x_2|/W}, & s_{y_2} &= \frac{1}{1 - |y_{p2} - y_2|/H}, \end{aligned} \tag{18}$$

where W and H denote the image width and height, respectively.

Final Reward. The final reward r combines the hit-based reward r_{hit}, the IoU-based reward r_{IoU}, and the box-size regularization reward r_{box} with fixed weights, as defined by

$$r = r_{\text{hit}} + 0.25 r_{\text{IoU}} + 0.125 r_{\text{box}}. \tag{19}$$

D Detailed Training Configurations

Table 5 summarizes the detailed training configurations of ConScore-VG for both general visual grounding and GUI grounding.

Table 5: Detailed training configurations of ConScore-VG for both general visual grounding and GUI grounding.

Parameter	General Visual Grounding	GUI Grounding
General		
Optimizer	AdamW	AdamW
Learning Rate	1×10^{-6}	1×10^{-6}
Effective Batch Size	8	64
Training Steps	96	40
Number of Samples (N)	256	1272
KL Coefficient (β)	0.04	0.04
Generation		
Number of Train Rollouts (G)	16	16
Sampling Temperature	1.0	1.0
Max Completion Length	2048	128
ConScore-VG		
Quantile Parameter (ρ)	0.25	0.25

E More Experiments

E.1 Cross-Dataset Generalization

We evaluate the cross-dataset generalization of ConScore-VG by adapting the model on one dataset and evaluating it on the remaining datasets. Notably, although these datasets partially overlap at the image level, they exhibit different instruction distributions. As shown in Figure 6, adapting ConScore-VG on a single dataset consistently improves its performance on the other datasets. These results suggest that the benefits of adaptation are not confined to the source dataset but transfer across datasets, reflecting broader improvements in visual grounding capability.

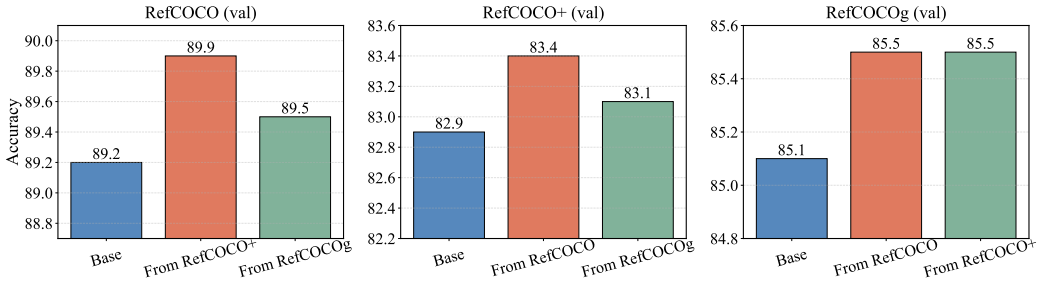


Figure 6: Cross-dataset generalization of ConScore-VG on the validation splits of the RefCOCOs. Each subplot corresponds to an evaluation dataset and compares the base model with models adapted on each of the other two datasets.

E.2 Ablation Study on Training Epochs

Table 6: Ablation Study on Training Epochs. The highlighted column indicates the setting used in the main experiments.

Dataset	0	1	2	3	4	5
RefCOCO (val)	89.2	89.6	89.6	90.0	90.1	90.0
RefCOCO+ (val)	82.9	83.5	83.8	83.9	84.0	84.0
RefCOCOg (val)	85.1	85.4	86.0	86.1	85.4	85.8

We conduct an ablation study on the validation splits of the RefCOCOs benchmarks to examine the effect of the number of TTRL training epochs. As shown in Table 6, performance consistently im-

proves during the first three epochs across all three datasets, indicating that ConScore-VG can adapt effectively with only a few epochs of optimization. Beyond three epochs, the gains largely saturate and become less consistent, with only marginal improvements on RefCOCO and RefCOCO+ and a slight performance drop on RefCOCOg. We therefore use three training epochs in the main experiments, which provides a favorable trade-off between adaptation performance and training efficiency.

E.3 Ablation Study on the Number of Adaptation Samples

We further examine the effect of the number of adaptation samples on the validation splits of the RefCOCOs benchmarks under a fixed budget of 96 optimization steps. As shown in Figure 7, using 256 samples yields the best overall performance across all three datasets. With only 128 samples, repeated optimization over a smaller adaptation set may lead to overfitting. In contrast, using 512 samples results in fewer training epochs under the same optimization budget, which may limit the extent of adaptation. Considering both adaptation performance and optimization efficiency, we use 256 samples and train for three epochs in the main experiments.

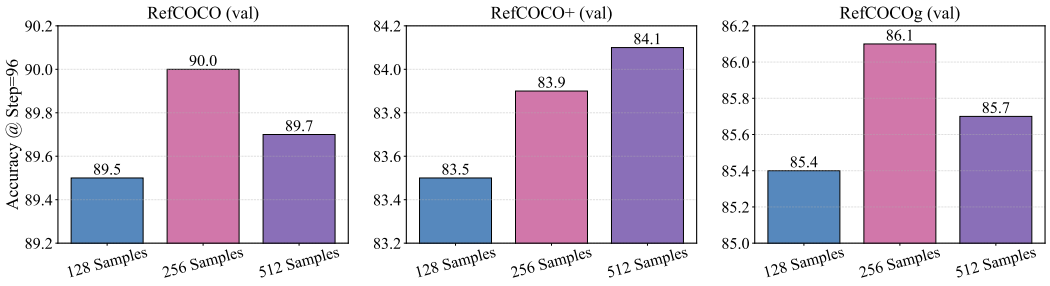


Figure 7: Ablation on the Number of Adaptation Samples. Accuracy under different adaptation sample counts with a fixed budget of 96 optimization steps.

E.4 Qualitative Comparison for GUI Grounding

For GUI grounding, we further qualitatively compare ConScore-VG with the Qwen2.5-VL-3B (Bai et al., 2025a) baseline and GUI-RCPO (Du et al., 2026) on ScreenSpot-v2 (Wu et al., 2025). As the GUI-RCPO checkpoint is not publicly available, we implement GUI-RCPO following the reported experimental setup. As shown in Figure 8, ConScore-VG produces bounding boxes that better match the ground-truth regions in both location and extent, whereas GUI-RCPO tends to predict overly compact boxes. Moreover, the Semantic Consistency Score helps improve instruction-region alignment. For example, in the second row of Figure 8, both Qwen2.5-VL-3B and GUI-RCPO incorrectly localize the "`Play`" button, while ConScore-VG correctly identifies the "`Spider`" game specified by the instruction.

E.5 Additional Qualitative Results

In this subsection, we present additional qualitative results of ConScore-VG across GUI grounding benchmarks (Figures 9 and 10) and general visual grounding benchmarks (Figure 11).

F Prompt for Semantic Consistency Scoring

As described in Section 3.2.2, inspired by the generation-verification asymmetry discussed by He et al. (2026), we assess the semantic consistency between each predicted bounding box b_i and the corresponding grounding instruction τ by reformulating visual grounding as its inverse task, visual referring. Specifically, we construct scoring prompts that transform each predicted region and its instruction into a visual referring task, allowing the model to evaluate their semantic correspondence. Given the distinct characteristics of general visual grounding and GUI grounding, we design task-specific scoring prompts for each setting, as detailed in Sections F.1 and F.2.



Figure 8: Comparing methods for GUI grounding on ScreenSpot-v2, including Qwen2.5-VL-3B (Bai et al., 2025a), GUI-RCPO (Du et al., 2026), and ConScore-VG (Ours).

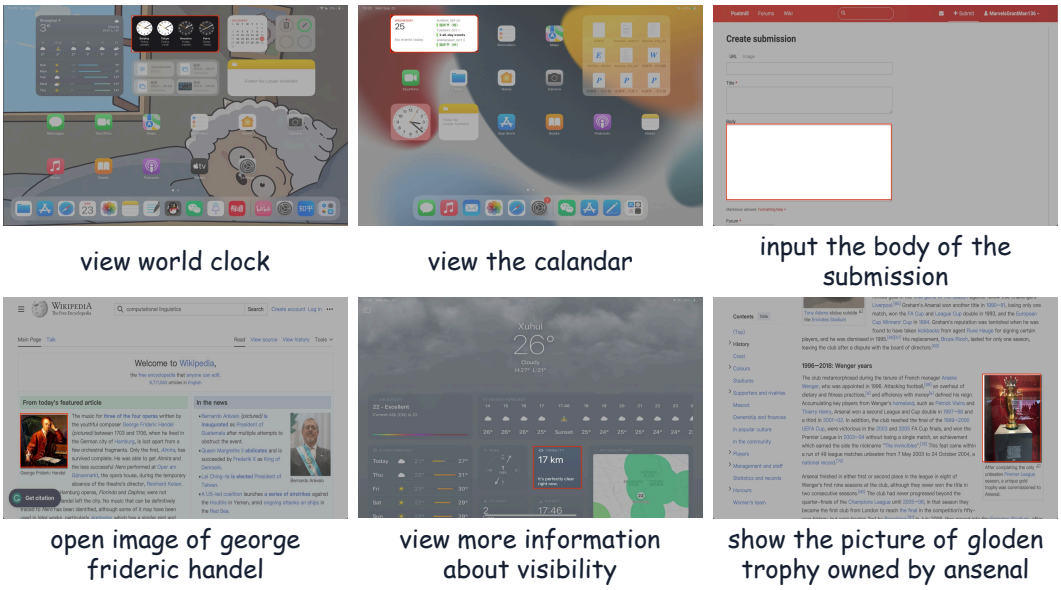


Figure 9: Additional visualization results of ConScore-VG on ScreenSpot-v2. The spelling errors in the instructions are inherited from the original dataset and are left unchanged.

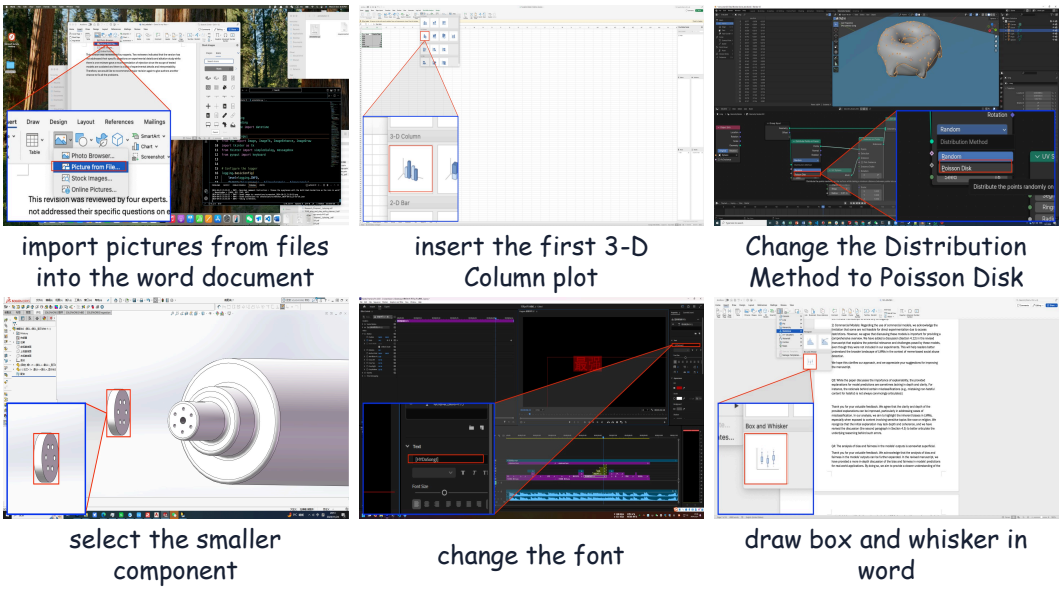


Figure 10: Additional visualization results of ConScore-VG on ScreenSpot-Pro.



Figure 11: Additional visualization results of ConScore-VG on RefCOCOs.

F.1 Scoring Prompt for General Visual Grounding

For general visual grounding, we use Referring Expression Comprehension (REC) as the representative setting and design the scoring prompt accordingly. Given an image and a referring expression, REC requires the model to localize the described object with a bounding box. We leverage its inverse task, Referring Expression Generation (REG), which maps a localized region back to a referring expression. Specifically, the predicted bounding box is used as a visual prompt, and the model is asked to describe the enclosed visual content with a concise referring expression. The complete prompt is provided in Table 7, where "`Instruction`" is replaced by the original referring expression and used as the teacher-forced target sequence.

Table 7: Semantic Consistency Scoring Prompt for General Visual Grounding, Illustrated with Referring Expression Comprehension.

USER: The red box marks a candidate region. Ignore the red border itself. **Describe only the visible content inside the red box using a short referring phrase**. Do not infer content that is masked, filled, blurred, or corrupted.
ASSISTANT: {`Instruction`}.

F.2 Scoring Prompt for GUI Grounding

For GUI grounding, natural language instructions typically specify an intended interaction with a target interface element rather than directly describing its visual appearance. Following the task formulation of ScreenSpot-v2 (Wu et al., 2025), where each instruction is associated with a target UI region, we define the inverse task as generating the corresponding user instruction from a candidate region. Specifically, we use the predicted bounding box as a visual prompt and ask the model to generate a concise instruction that could be fulfilled by clicking or tapping the enclosed UI element based solely on its visible content. This formulation encourages the generated instruction to capture the functional semantics of the candidate region rather than its visual appearance alone, providing a more suitable signal for assessing the semantic consistency between the candidate region and the original instruction. The complete prompt is provided in Table 8, where "`Instruction`" is replaced by the original GUI instruction and used as the teacher-forced target sequence.

Table 8: Semantic Consistency Scoring Prompt for GUI Grounding.

USER: The red box marks a candidate UI region. Ignore the red border itself. **Based only on the visible content inside the red box, write the most likely short user instruction that clicking/tapping this region would fulfill**. Do not infer masked, filled, blurred, or corrupted content. Output only one imperative instruction phrase.
ASSISTANT: {`Instruction`}.