Thinking with Images as Continuous Actions: Numerical Visual Chain-of-Thought

Kesen Zhao[1] **Beier Zhu**[2,✉] **Junbao Zhou**[1] **Xingyu Zhu**[3] **Zhongqi Yue**[4] **Hanwang Zhang**[1]
[1]NTU [2]USTC [3]NUS [4]Microsoft Research
Correspondence to: Beier Zhu <beier.zhu@ustc.edu.cn>

Abstract

Recent MLLMs increasingly rely on *visual chain-of-thought* to perform region-grounded reasoning over images. However, existing approaches ground regions via either textified coordinates—causing modality mismatch and semantic fragmentation—or fixed-granularity patches that both limit precise region selection and often require non-trivial architectural changes. In this paper, we propose Numerical Visual Chain-of-Thought (NV-CoT), a framework that enables MLLMs to reason over images using *continuous* numerical coordinates. NV-CoT expands the MLLM action space from discrete vocabulary tokens to a continuous Euclidean space, allowing models to directly generate bounding-box coordinates as actions with only minimal architectural modification. The framework supports both supervised fine-tuning and reinforcement learning. In particular, we replace categorical token policies with a Gaussian (or Laplace) policy over coordinates and introduce stochasticity via reparameterized sampling, making NV-CoT fully compatible with GRPO-style policy optimization. Extensive experiments on three benchmarks against 11 representative visual reasoning baselines demonstrate that NV-CoT significantly improves localization precision and final answer accuracy, while also accelerating training convergence, validating the effectiveness of continuous-action visual reasoning in MLLMs. **Project page:** kesenzhao.github.io/NV-CoT/

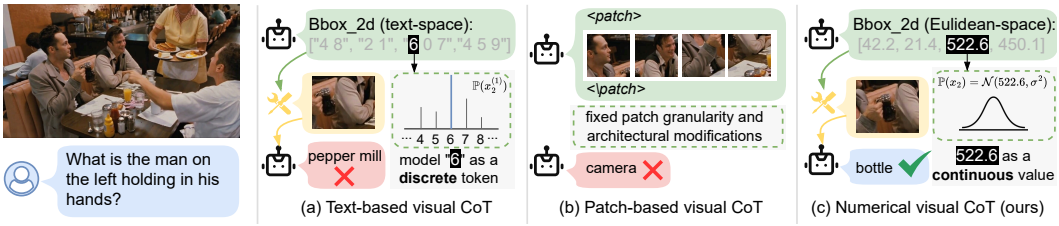
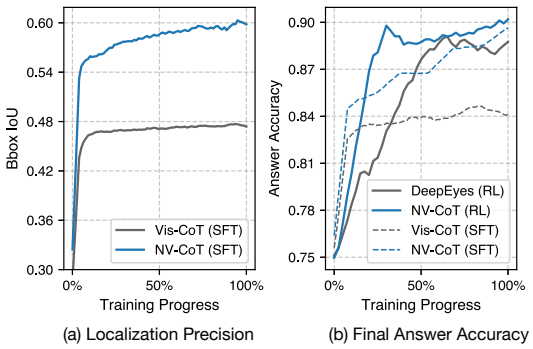


Figure 1: Comparison of different paradigms for thinking with images. (a) Text-based approaches represent localized regions as discrete coordinate tokens, leading to modality mismatch and fragmented semantics. (b) Patch-based approaches reason directly over fine-grained visual tokens but are constrained by the fixed spatial granularity of the vision backbone. (c) Our NV-CoT predicts region coordinates in continuous space, enabling flexible and precise localization.

1 Introduction

Recent advances in multimodal large language models (MLLMs) have enabled deeper reasoning over visual inputs via "thinking with images" (*i.e.*, visual chain-of-thought) [1–5]. In this paradigm, models first localize task-relevant image regions aligned with the textual query. They then conduct multi-step reasoning by explicitly grounding intermediate inferences on these detected regions. Such

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Localization Precision (b) Final Answer Accuracy

Figure 2: Our NV-CoT outperforms text-based visual CoT models (Vis-CoT [2] and DeepEyes [1]) in **localization precision**, **answer accuracy**, and **convergence speed** across both SFT and RL. SFT-based models are evaluated on the Vis-CoT-363K dataset [2], where ground-truth bounding boxes are available, while RL-based models are evaluated on the DeepEyes-47K dataset [1]. We only replace the text-space discrete coordinate objective with our Euclidean-space continuous one, while keeping all other training configurations unchanged for a fair comparison.

region-grounded visual reasoning is a fundamental capability for MLLMs and underpins a broad range of downstream applications.

To align with the text-based output interface, many existing MLLMs [6–8] serialize localized regions as bounding-box coordinates **in text**, *e.g.*, [“x_1”,“y_1”, “x_2”, “y_2”], and then invoke a cropping tool to extract the corresponding patch (Figure 1 (a)). Despite its simplicity, this design faces two fundamental issues. (1) Modality mismatch: coordinates are continuous in the visual world but are predicted as discrete text tokens and typically trained with cross-entropy, which ignores geometric proximity (*e.g.*, predicting x_1=“3.2” may be penalized similarly to x_1=“4.1” when the target is x_1^*=“3.1”, since both predictions differ from the target by only a single discrete token under token-level cross-entropy). (2) Semantic fragmentation: numbers are tokenized into multiple unrelated sub-tokens, making numerical comparison and reasoning brittle and prone to hallucination (*e.g.*, confusing whether “3.11” or “3.9” is larger).

Several recent works attempt to alleviate these issues by enabling MLLMs to operate directly on fine-grained visual patches (Figure 1 (b)). For example, LVR [9] performs reasoning in the visual embedding space, while PaDT [10] generates indices corresponding to original visual patch embeddings. Nevertheless, these approaches are fundamentally constrained by the fixed granularity of the vision backbone, as the predefined patch partitioning limits flexible and precise region selection. In addition, patch-generation or patch-indexing mechanisms often require non-trivial architectural changes, which weakens modularity and limits general applicability across MLLMs.

To avoid brittle textified coordinates with only minimal architectural changes, we propose Numerical Visual Chain-of-Thought (NV-CoT), which expands the MLLM action space from discrete vocabulary tokens to a continuous Euclidean space, enabling the model to directly generate numerical bounding box coordinates as actions (Figure 1 (c)). Specifically, we extend the standard LLM head with four coordinate outputs that directly predict the numerical bounding-box coordinates (*e.g.*, $[x_1, y_1, x_2, y_2]$[1]). Notably, NV-CoT is applicable to both supervised fine-tuning (SFT) and reinforcement learning (RL) stages. In the SFT stage, we replace token-level cross-entropy with a regression loss to supervise continuous bounding-box coordinates. However, extending this idea to the RL stage is non-trivial. Most LLM-oriented policy optimization methods parameterize the policy as a token-level categorical distribution over the vocabulary, which naturally supports discrete actions but does not directly accommodate continuous coordinate outputs. Furthermore, deterministic coordinate prediction lacks the stochasticity required by GRPO-style algorithms for advantage estimation and policy exploration. To address these challenges, we replace the categorical policy with a Gaussian policy that explicitly models continuous coordinate regression. Stochasticity is introduced via the reparameterization trick, where the model predicts both the mean and standard deviation of the bounding box distribution and samples coordinates accordingly. This design enables efficient exploration and stable gradient propagation, while also providing closed-form likelihoods for computing importance ratios and KL regularization, making NV-CoT compatible with GRPO-style RL frameworks. Finally, we also provide a Laplace-policy variant of NV-CoT, motivated by the empirical robustness and favorable performance of ℓ_1-type losses for localization; it serves as a drop-in replacement for the Gaussian policy under the same GRPO pipeline.

[1]We write text-based coordinates as [“x_1”, “y_1”, “x_2”, “y_2”] and numerical coordinates as $[x_1, y_1, x_2, y_2]$.

In Figure 2, we plot bounding-box precision (IoU under the SFT setting with ground-truth coordinates) and final answer accuracy over the course of training. The results show that `NV-CoT` consistently improves **localization accuracy**, **final answer accuracy**, and **convergence speed**. Extensive experiments on V^* Bench [11], HR-Bench 4K [12], and HR-Bench 8K [12] demonstrate that `NV-CoT` consistently improves visual reasoning performance. For example, on V^* Bench, `NV-CoT-7B` outperforms LVR-7B (patch-based visual CoT) and DeepEyes-7B (text-based visual CoT) by 7.8% and 2.6%, respectively (Table 1). Beyond high-resolution benchmarks, `NV-CoT` also outperforms baselines on general multimodal benchmarks, including MMMU-Pro [13], ChartQA [14], and OCRBenchv2 [15] (Table 2). It also retains performance on pure-text reasoning tasks (Table 3). On grounding and counting benchmarks, `NV-CoT` outperforms both Qwen2.5-VL-7B and DeepEyes-7B across the RefCOCO series, COCO detection, CountBenchQA, and TallyQA (Table 4).

2 Related Work

Multimodal large language models. By integrating language modeling capabilities with visual understanding, MLLMs enable complex vision-language tasks [16–20]. Representative early models such as BLIP-2 [21] and LLaVA [22, 23] align vision and language by projecting image into the latent space of a frozen LLM via a query transformer or a lightweight projector. Subsequent strong open-source models, including the LLaVA family [24–29], Qwen-VL [7, 30, 31], and InternVL [32, 33], further improve visual resolution handling and scalability. However, despite their strong perception and alignment capabilities, these models generally lack explicit reasoning mechanisms.

Thinking with Images (visual CoT). Incorporating CoT reasoning to enhance visual understanding has been actively explored in recent MLLMs [34, 35]. Early text-based approaches [6–8] serialize localized image regions as bounding-box coordinates in text, and subsequently invoke cropping tools to extract the corresponding visual patches. While effective, this paradigm suffers from modality mismatch and semantic fragmentation. More recent patch-based methods [9, 10, 36] instead operate directly on fine-grained visual patches, avoiding explicit textual serialization. However, these approaches are fundamentally constrained by the fixed spatial granularity of the vision backbone, limiting their flexibility in region selection and reasoning precision. In contrast, `NV-CoT` expands the MLLM action space from discrete vocabulary tokens to a continuous Euclidean space, enabling the model to directly predict numerical bounding-box coordinates as continuous actions. This design naturally bridges visual perception and reasoning, without relying on fixed patch partitions.

To incorporate CoT to MLLMs, existing methods adopt either supervised fine-tuning [2, 37–39] or reinforcement learning [1, 3, 5, 35]. Our approach is compatible with both paradigms, providing a unified framework for learning continuous visual actions.

3 Method

Given a query q and an image I_0, a visual CoT MLLM π_θ performs region-grounded reasoning by selecting a region of interest and invoking a zoom-in function for fine-grained inspection. Existing visual CoT methods localize regions by generating **text-form** box coordinates and then cropping the image for feedback. We instead cast localization as a continuous-action problem in Euclidean space: the MLLM directly outputs **numerical coordinates**. Our method proceeds in three steps. First, we expand the MLLM vocabulary with a continuous coordinate action space (§3.1), enabling numerical box generation. Second, we instantiate `NV-CoT` for SFT with coordinate regression when box supervision is available (§3.2). Third, we extend `NV-CoT` to GRPO by introducing a stochastic Gaussian policy with closed-form importance ratios and KL regularization (§3.3). Finally, we provide a Laplace-policy variant as a drop-in replacement, motivated by the empirical preference of ℓ_1-type losses for localization (§3.4).

3.1 Expanded Actions for Continuous Coordinates

To enable numerical coordinate prediction, we augment the standard LLM output head over the vocabulary $\mathcal{V}$ with a continuous action space $\mathcal{A} = \mathbb{R}^4 \times \mathbb{R}_+$: four coordinate heads parameterized by $W = [\mathbf{w}_{x_1}, \mathbf{w}_{y_1}, \mathbf{w}_{x_2}, \mathbf{w}_{y_2}]^\top \in \mathbb{R}^{4\times D}$ to predict the Gaussian mean, and a fifth head $\mathbf{w}_\sigma \in \mathbb{R}^D$ to predict a shared standard deviation, yielding a Gaussian policy for continuous localization. Let $\mathbf{h} \in \mathbb{R}^D$ denote the final hidden representation before the coordinate decoding step. We obtain the

mean and shared standard deviation as

$$\boldsymbol{\mu} = [x_1, y_1, x_2, y_2]^\top = W\mathbf{h}, \qquad \sigma = \max(\mathbf{w}_\sigma^\top \mathbf{h}, \varepsilon), \tag{1}$$

where $\varepsilon > 0$ is a small positive offset to ensure $\sigma > 0$. This forms a Gaussian policy $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$. Accordingly, we augment the parameter set as $\theta \leftarrow \theta \cup \{W, \mathbf{w}_\sigma\}$.

The introduction of σ is crucial: while deterministic coordinate suffices for SFT, GRPO-style RL requires stochastic policies to enable sampling and exploration. Note that our modification to the MLLM is minimal, as we only introduce five lightweight linear heads, in contrast to patch-based visual CoT methods that require more architectural changes.

3.2 `NV-CoT` for Supervised Fine-Tuning

We begin by applying our `NV-CoT` to supervised fine-tuning, where intermediate bounding-box coordinates in the reasoning process are available [2, 39]. Let $\mathbf{b}^* = [x_1^*, y_1^*, x_2^*, y_2^*]^\top$ denote the ground-truth bounding box for a zoom-in step. During SFT, we supervise the coordinate outputs in $\mathcal{A}$ using an ℓ_2^2 regression loss, while optimizing the discrete token outputs in the original LLM head $\mathcal{V}$ with the standard cross-entropy loss.

$$\mathcal{L}_{\mathsf{SFT}}(\theta) = \sum_{t \in \mathcal{T}_\mathcal{V}} \Big(-\log \pi_\theta(o_t \mid o_{<t}, q, I_0) \Big) \; + \; \lambda \sum_{t \in \mathcal{T}_\mathcal{A}} \|\boldsymbol{\mu}_t - \mathbf{b}_t^*\|_2^2,$$

where $\mathcal{T}_\mathcal{V} = \{t : o_t \in \mathcal{V}\}$ and $\mathcal{T}_\mathcal{A} = \{t : o_t \in \mathcal{A}\}$ denote textual token and coordinate-action positions, respectively; o_t is the output of the LLM, and $\lambda > 0$ is a hyperparameter; σ is not supervised in SFT and is only used for RL.

3.3 `NV-CoT` for Reinforcement Learning

Bounding-box annotations for SFT are costly and often unavailable at scale. As a result, visual CoT models are commonly trained with RL (GRPO Shao et al. [40]) using task-level feedback (*e.g.*, final answer) as the reward signal, without requiring box-level supervision. Extending `NV-CoT` from SFT to this RL setting, however, is non-trivial: RL needs stochastic sampling and well-defined importance ratios, whereas standard LLM policies are token-level categorical distributions that do not accommodate continuous coordinate actions. To this end, we formulate continuous region localization as a Gaussian policy over the action space $\mathcal{A}$.

Rollouts. GRPO samples a group of G trajectories per prompt. With our Gaussian policy, we independently sample coordinate actions as $\mathbf{b}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I), \; i = 1, \ldots, G$, To enable low-variance gradient estimation, we implement this sampling using the reparameterization trick:

$$\mathbf{b}^{(i)} = \boldsymbol{\mu} + \sigma \boldsymbol{\epsilon}^{(i)}, \quad \boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, I). \tag{2}$$

We then use $\mathbf{b}^{(i)}$ to crop into the image for the i-th trajectory.

Importance ratio. GRPO optimizes the policy via an importance ratio between the current policy π_θ and a reference policy $\pi_{\theta_{\text{old}}}$. For each output $o_t^{(i)}$, the ratio is defined as

$$r_t^{(i)}(\theta) = \frac{\pi_\theta(o_t^{(i)} \mid q, I_0, o_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(o_t^{(i)} \mid q, I_0, o_{<t}^{(i)})}. \tag{3}$$

When $o_t^{(i)} \in \mathcal{V}$ is a text token, $r_t^{(i)}(\theta)$ reduces to the standard likelihood ratio of categorical policies used in GRPO. When $o_t^{(i)} \in \mathcal{A}$ corresponds to a continuous coordinate action $\mathbf{b}_t^{(i)}$, we compute the ratio using Gaussian likelihoods:

$$r_t^{(i)}(\theta) = \frac{\mathcal{N}(\mathbf{b}_t^{(i)} \mid \boldsymbol{\mu}_\theta, \sigma_\theta^2 I)}{\mathcal{N}(\mathbf{b}_t^{(i)} \mid \boldsymbol{\mu}_{\theta_{\text{old}}}, \sigma_{\theta_{\text{old}}}^2 I)}, \text{ where } \mathcal{N}(\mathbf{b} \mid \boldsymbol{\mu}, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\mathbf{b} - \boldsymbol{\mu}\|_2^2}{2\sigma^2} \right), \tag{4}$$

Setting $d = 4$ and substituting this density into the ratio gives the importance ratio in terms of squared Mahalanobis distances:

$$r_t^{(i)}(\theta) = \frac{\sigma_{\theta_{\text{old}}}^4}{\sigma_\theta^4} \exp\left(-\frac{\|\mathbf{b}_t^{(i)} - \boldsymbol{\mu}_\theta\|_2^2}{2\sigma_\theta^2} + \frac{\|\mathbf{b}_t^{(i)} - \boldsymbol{\mu}_{\theta_{\text{old}}}\|_2^2}{2\sigma_{\theta_{\text{old}}}^2} \right). \tag{5}$$

KL penalty. Thanks to the closed-form KL divergence of Gaussian distributions, this penalty is analytically tractable for continuous actions $\mathcal{A}$:

$$\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathsf{ref}}) = \frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathsf{ref}}\|_2^2}{2\sigma_{\mathsf{ref}}^2} + 2\Big(\frac{\sigma_\theta^2}{\sigma_{\mathsf{ref}}^2} - \log \sigma_\theta^2\Big) + C, \tag{6}$$

where C is a constant. However, σ_{ref} is unknown since σ is not supervised in SFT. Therefore, we drop the variance-related term and only constrain the mean: $\mathrm{KL}_{\boldsymbol{\mu}}(\pi_\theta \,\|\, \pi_{\mathsf{ref}}) = \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\mathsf{ref}}\|_2^2$. If SFT stage is unavailable (and thus no reference mean $\boldsymbol{\mu}_{\mathsf{ref}}$ can be obtained), we omit this KL regularization term. For textual token outputs in $\mathcal{V}$, we retain the standard token-level KL divergence.

Reward. Following DeepEyes [1], we adopt an outcome-driven reward that combines answer correctness, format validity, and a conditional zoom-in use bonus. Formally, for a trajectory $o^{(i)}$, the total reward is

$$R(o^{(i)}) \;=\; R_{\mathsf{acc}}(o^{(i)}) \;+\; R_{\mathsf{fmt}}(o^{(i)}) \;+\; R_{\mathsf{zoom}}(o^{(i)}), \tag{7}$$

where the zoom-in bonus is activated only when the trajectory is correct and invokes the visual grounding action at least once, encouraging region of interest identification.

Optimization. Finally, given a question q and an input image I_0, we sample a group of G trajectories $\{o^{(1)}, \ldots, o^{(G)}\}$ from $\pi_{\theta_{\mathsf{old}}}$, each associated with intermediate bounding boxes $\{\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(G)}\}$ (via Eq. (2)). We compute trajectory-level rewards $\mathbf{R} = \{R(o^{(1)}), \ldots, R(o^{(G)})\}$ using Eq. (7) and use the normalized advantage for all steps in the i-th trajectory: $A_t^{(i)} = \frac{R(o^{(i)}) - \mathsf{mean}(\mathbf{R})}{\mathsf{std}(\mathbf{R})}$. Given the importance ratio $r_t^{(i)}(\theta)$ and the KL term $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$ defined above, we optimize the objective:

$$\mathcal{L}_{\mathsf{GRPO}}(\theta) = \mathbb{E}_{q, I_0, \{o^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\mathsf{old}}}} \Big[\frac{1}{GT} \sum_{i=1}^{G} \sum_{t=1}^{T} \min\Big(r_t^{(i)}(\theta) A_t^{(i)}, \mathrm{clip}\big(r_t^{(i)}(\theta), 1-\varepsilon, 1+\varepsilon\big) A_t^{(i)} \Big) - \beta \mathrm{KL}(\pi_\theta \| \pi_{\mathsf{ref}}) \Big], \quad \text{where } \varepsilon > 0 \text{ is the clipping range.} \tag{8}$$

Inference. During inference, we disable stochasticity, crop using the coordinate mean $\boldsymbol{\mu}$, and express all iterative zoom-in boxes in the original image's coordinate frame.

3.4 Extending `NV-CoT` with Laplace Policy

In prediction tasks such as keypoint localization and bounding-box regression, it is well known that ℓ_1 loss often outperforms ℓ_2^2 loss due to its robustness to outliers and sharper error profiles [41, 42]. From a probabilistic perspective, this corresponds to modeling prediction errors with a Laplace distribution rather than a Gaussian distribution. Motivated by this observation, we further extend `NV-CoT` by replacing the Gaussian policy with a Laplace policy for continuous localization.

Laplace policy. Specifically, we model the coordinate action as $\mathbf{b} \sim \mathrm{Laplace}(\boldsymbol{\mu}, \alpha)$, whose probability density function is

$$p(\mathbf{b} \mid \boldsymbol{\mu}, \alpha) = \frac{1}{(2\alpha)^4} \exp\left(-\frac{\|\mathbf{b} - \boldsymbol{\mu}\|_1}{\alpha} \right), \tag{9}$$

where $\boldsymbol{\mu} \in \mathbb{R}^4$ denotes the predicted coordinate mean and $\alpha > 0$ is a shared scale parameter. Equivalently, this corresponds to assuming independent dimensions, under which the Laplace likelihood factorizes across coordinates.

SFT objective. Under the Laplace assumption, maximizing the log-likelihood of ground-truth coordinates is equivalent to minimizing an ℓ_1 loss, since $-\log p(\mathbf{b}^* \mid \boldsymbol{\mu}, \alpha) = \frac{1}{\alpha}\|\boldsymbol{\mu} - \mathbf{b}^*\|_1 + \mathrm{const}$. Thus, in the SFT stage, we replace the ℓ_2^2 loss Eq. (2) with

$$\mathcal{L}_{\mathsf{SFT}}^{\mathsf{Lap}}(\theta) = \sum_{t \in \mathcal{V}_{\mathcal{A}}} \|\boldsymbol{\mu}_t - \mathbf{b}_t^*\|_1, \tag{10}$$

which encourages sharper and more robust localization.

RL with Laplace policy. Analogous to the Gaussian case, we parameterize the Laplace policy with a location parameter $\boldsymbol{\mu}$ and a scale parameter α, and employ reparameterized sampling to enable low-variance gradient estimation. Specifically, a coordinate action is sampled as

$$\mathbf{b} = \boldsymbol{\mu} + \alpha\, \mathbf{s} \odot \boldsymbol{\epsilon}, \quad \mathbf{s} \sim \mathrm{Rademacher}(\pm 1),\ \boldsymbol{\epsilon} \sim \mathrm{Exp}(\mathbf{1}), \tag{11}$$

where $\odot$ denotes element-wise multiplication. Here, $\mathbf{s}$ is a random sign vector with i.i.d. entries in $\{\pm 1\}$, and $\boldsymbol{\epsilon}$ has i.i.d. unit-rate exponential entries. For a sampled action $\mathbf{b}_t^{(i)}$, the corresponding importance ratio used in GRPO is

$$r_t^{(i)}(\theta) = \frac{\alpha_{\theta_{\text{old}}}^4}{\alpha_\theta^4} \exp\left(-\frac{\|\mathbf{b}_t^{(i)} - \boldsymbol{\mu}_\theta\|_1}{\alpha_\theta} + \frac{\|\mathbf{b}_t^{(i)} - \boldsymbol{\mu}_{\theta_{\text{old}}}\|_1}{\alpha_{\theta_{\text{old}}}} \right), \tag{12}$$

which can be computed analytically. The closed-form KL penalty can be derived analogously, and the overall GRPO procedure follows the same pipeline as in the Gaussian-policy case.

4 Experiments

In §4.1, we describe the experimental setup. In §4.2, we report the main results on high-resolution benchmarks. In §4.3, we present results beyond high-resolution benchmarks, including general multimodal, pure-text, grounding, and counting tasks. In §4.4, we present ablation studies and analyses of uncertainty, the training coefficient, training dynamics, and qualitative grounding.

4.1 Setup

Baselines. We compare against 11 state-of-the-art MLLM baselines, grouped into four categories. (1) Open-source MLLMs: Qwen2.5-VL-{7B, 32B} [7] and LLaVA-OneVision [29]. (2) Training-free visual CoT: DyFo-7B [43] (3) SFT-based visual CoT: Vis-CoT-7B [2], ZoomEye-7B [44] and RAP-7B [45]. (4) RL-based visual CoT: UV-CoT-7B [3], DeepEyes-7B [1], PaDT [10] and LVR [9]. Our `NV-CoT` is applicable to both SFT and RL settings. In this work, we build the SFT-based model upon Vis-CoT [2] and the RL-based model upon DeepEyes [1]

Datasets. For fair comparison, we use the *same training data* as the corresponding backbones (*e.g.*, DeepEyes-7B and Vis-CoT-7B). For evaluation, we consider four categories: *(1) High-resolution visual reasoning.* We evaluate on V^* Bench [11], which measures attribute (Attr) and spatial reasoning (Spatial), and HR-Bench 4K/8K [12], which assesses fine-grained spatial perception (FSP) and fine-grained comparison perception (FCP) across resolutions. *(2) General multimodal understanding.* We use MMMU-Pro [13] for multidisciplinary understanding, ChartQA [14] for chart reasoning, and OCRBenchv2 [15] for visual text reasoning. *(3) Grounding and counting.* We evaluate referring-expression grounding on RefCOCO, RefCOCO+, and RefCOCOg [46, 47], object detection on COCO [48], and visual counting on CountBenchQA [49] and TallyQA [50]. *(4) Pure-text reasoning.* We evaluate language and mathematical reasoning on MMLU [51] and GSM8K [52], respectively.

Evaluation protocols. For a fair comparison under identical inference scaling, we evaluate all models with a single zoom-in per query, since some baselines (*e.g.*, Vis-CoT [2] and UV-CoT [3]) do not support iterative tool invocations. For methods that do support multiple tool calls (*e.g.*, DeepEyes [1] and our `NV-CoT` (RL)), we additionally report results under multiple tool use.

Implementation details. To ensure a fair comparison, we *keep the training configurations identical* to those of the backbone methods (DeepEyes-7B [1] and Vis-CoT-7B [2]), and only replace text-form coordinate prediction with our continuous action formulation. Specifically, for SFT, we fine-tune LLaVA for one epoch with a batch size of 128. We use the Adam optimizer with zero weight decay, a learning rate of 2×10^{-5}, and a cosine learning-rate scheduler. λ in Eq. (2) is set to 0.3. For RL, we train the RL-based model on top of Qwen2.5-VL-7B using GRPO for 80 iterations on 32 A100 GPUs. We set the batch size to 256 and the GRPO group size to $G = 16$. The KL coefficient is set to $\beta = 0$.

4.2 Main Results on High-Resolution Benchmarks

The overall performance comparisons are reported in Table 1, leading to the following observations:

Table 1: Results on high-resolution benchmarks. **Best** results are in bold. Our NV-CoT consistently improves performance under both SFT and RL settings. We use NV-CoT (SFT) with an ℓ_1 loss and NV-CoT (RL) with a Laplace policy. **DeepEyes-7B**† and NV-CoT† denote results under multi-step tool use (multiple zoom-in calls). For a fair comparison, all other visual CoT methods are evaluated with a single zoom-in call, since some baselines do not support multiple tool invocations.

Model	V^* Bench			HR Bench 4K			HR Bench 8K		
	Attr	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
LLaVA-OneVision-7B	75.7	75.0	75.4	71.7	53.3	62.5	69.1	52.1	60.6
Qwen2.5-VL-7B	73.9	67.1	71.2	84.7	52.9	68.8	78.7	51.8	65.3
Qwen2.5-VL-32B	87.8	88.1	87.9	89.4	58.3	73.9	83.6	56.9	70.3
DyFo-7B	80.0	82.9	81.2	-	-	-	-	-	-
PaDT-7B	75.7	73.6	74.9	86.0	54.8	70.4	79.0	53.8	66.4
LVR-7B	84.4	76.3	81.2	88.3	56.8	72.5	82.5	54.5	68.5
UV-CoT-7B	82.6	76.3	80.1	79.8	54.5	67.1	76.8	54.3	65.5
ZoomEye-7B	93.9	85.5	90.6	84.3	55.0	69.6	88.5	50.0	69.3
RAP-7B	87.0	78.9	83.8	76.3	50.5	63.4	69.0	46.8	57.9
Vis-CoT-7B	80.9	75.0	78.5	79.0	54.3	66.6	75.5	53.0	64.3
+ NV-CoT (SFT)	84.3	78.9	82.2	82.0	55.3	68.6	77.3	54.5	65.9
Δ	+3.5	+3.9	+3.7	+3.0	+1.0	+2.0	+1.8	+1.5	+1.6
DeepEyes-7B	87.8	84.2	86.4	89.3	57.8	73.5	84.3	56.5	70.4
+ NV-CoT (RL)	90.4	86.8	89.0	91.0	58.5	74.8	85.8	58.3	72.0
Δ	+2.6	+2.6	+2.6	+1.7	+0.7	+1.3	+1.5	+1.8	+1.6
DeepEyes-7B†	91.3	88.2	90.1	91.3	59.0	75.2	86.8	58.5	72.7
+ NV-CoT† (RL)	**93.0**	**89.5**	**91.6**	**91.8**	**60.3**	**76.0**	**87.5**	**60.0**	**73.8**
Δ	+1.7	+1.3	+1.5	+0.5	+1.3	+0.8	+0.7	+1.5	+1.1

Model	MMMU-Pro	ChartQA	OCRBenchv2
Qwen2.5-VL-7B	38.3	58.9	44.7
DeepEyes-7B	44.6	62.5	46.1
NV-CoT -7B	**46.7**	**65.1**	**47.6**

Table 2: Performance comparison on general multimodal benchmarks. NV-CoT achieves the best performance across MMMU-Pro, ChartQA, and OCRBench v2.

(1) Our NV-CoT consistently outperforms the backbone models across all benchmarks under both SFT and RL. Specifically, under SFT, NV-CoT improves over Vis-CoT-7B by $+3.7\%$ overall on V^* Bench, $+2.0\%$ on HR-Bench 4K, and $+1.6\%$ on HR-Bench 8K. Under RL, NV-CoT achieves further gains over DeepEyes-7B, with overall improvements of $+2.6\%$, $+1.3\%$, and $+1.6\%$ on the three benchmarks, respectively. These results suggest that replacing text-based bounding-box prediction with our numerical one benefits both SFT and RL training paradigms.

(2) We observe a clear paradigm advantage of zoom-in visual CoT over patch-based reasoning. The backbone DeepEyes-7B and our NV-CoT (RL) consistently outperform patch-based baselines such as PaDT-7B and LVR-7B across all benchmarks, suggesting that region selection via zoom-in is more effective than reasoning over fixed-granularity patches.

(3) Methods that perform thinking with images substantially outperform basic MLLMs such as LLaVA-OneVision-7B and Qwen2.5-VL-7B. Notably, despite being built on a 7B backbone, NV-CoT achieves performance that surpasses Qwen2.5-VL-32B across all benchmarks. This highlights the importance of region-grounded visual reasoning in improving multimodal understanding.

(4) Allowing multiple zoom-in tool calls further improves "thinking with images" performance, indicating that iterative region selection and refinement is beneficial for fine-grained visual reasoning. Notably, DeepEyes-7B† achieves consistently higher scores than its single-call counterpart. Building on this stronger multi-step setting, NV-CoT† further boosts performance by replacing textified coordinates with continuous coordinate actions, enabling more precise region localization.

Model	MMLU	GSM8K
Qwen2.5-VL-7B	70.88	73.84
NV-CoT-7B	71.07	73.54

Table 3: Performance on pure-text reasoning benchmarks. NV-CoT maintains comparable performance to Qwen2.5-VL-7B on MMLU and GSM8K.

Table 4: Performance comparison on grounding and counting benchmarks. NV-CoT consistently outperforms strong baselines across RefCOCO series, COCO detection, and counting tasks.

Model	RefCOCO	RefCOCO+	RefCOCOg	COCO	CountBenchQA	TallyQA
Qwen2.5-VL-7B	87.1	79.7	83.0	19.5	77.4	71.2
DeepEyes-7B	88.5	81.5	85.1	22.8	81.9	75.9
NV-CoT -7B	**91.7**	**85.0**	**87.4**	**27.4**	**86.2**	**80.5**

Table 5: Ablation on coordinate policy (Gaussian (ℓ_2^2) vs. Laplace (ℓ_1)) and uncertainty parameterization (shared σ/α vs. per-coordinate $\boldsymbol{\sigma}/\boldsymbol{\alpha}$). Non-shared $\boldsymbol{\sigma}/\boldsymbol{\alpha}$: per-coordinate scale (standard deviation) predicted by four heads. Laplace policy consistently performs better than Gaussian policy, while using a shared versus per-coordinate $\boldsymbol{\sigma}/\boldsymbol{\alpha}$ yields comparable results.

Model		V^* Bench			HR Bench 4K			HR Bench 8K		
		Attr	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
SFT	ℓ_2^2 loss	83.5	76.3	80.6	81.5	53.5	67.5	76.0	54.3	65.1
	ℓ_1 loss	84.3	78.9	82.2	82.0	55.3	68.6	77.3	54.5	65.9
RL	Gaussian policy	88.7	85.5	87.4	90.3	58.3	74.3	85.0	58.0	71.5
	+ Independent $\boldsymbol{\sigma}$	87.8	86.8	87.4	90.0	58.3	74.1	85.5	57.8	71.6
	Laplace policy	90.4	86.8	89.0	91.0	58.5	74.8	85.8	58.3	72.0
	+ Independent $\boldsymbol{\alpha}$	89.6	86.8	88.5	90.8	58.8	74.8	86.0	58.3	72.1

4.3 Results Beyond High-Resolution Benchmarks

General multimodal benchmarks. To evaluate generality beyond grounding-specific benchmarks, we conduct experiments on general multimodal benchmarks, including MMMU-Pro, ChartQA, and OCRBenchv2. As shown in Table 2, NV-CoT consistently improves performance over both the base VLM and DeepEyes. This suggests that our method does not degrade general reasoning ability, but instead provides complementary gains.

Pure-text reasoning benchmarks. To test whether the coordinate branch affects language reasoning, we evaluate the base Qwen2.5-VL-7B and NV-CoT on MMLU and GSM8K (with strict-match). NV-CoT scores 71.07% versus 70.88% on MMLU and 73.54% versus 73.84% on GSM8K, indicating comparable performance on these pure-text tasks (Table 3).

Object detection benchmarks. In Table 4, we evaluate NV-CoT on standard grounding benchmarks and grounding-sensitive counting tasks, including the RefCOCO series, COCO detection, CountBenchQA, and TallyQA. For single-object grounding benchmarks, we report accuracy. For COCO, which involves multi-object detection, we report mAP following the standard evaluation protocol. Overall, the results further demonstrate the strong grounding capability of our method.

4.4 Ablation Studies and Analyses

We ablate the coordinate policy and uncertainty parameterization, and analyze uncertainty behavior, the effect of the training coefficient, RL training dynamics, and qualitative grounding.

Regression loss and policy distribution. We jointly analyze the choice of regression loss in SFT and policy distribution in RL, as both stages reflect continuous modeling of coordinates. In the SFT stage, the ℓ_1 loss consistently outperforms the ℓ_2^2 loss across all benchmarks, improving overall accuracy on V^* Bench by $+1.6\%$ and yielding consistent gains on HR-Bench 4K ($+1.1\%$) and 8K ($+0.8\%$). These trends carry over to RL: the Laplace policy consistently outperforms the Gaussian policy on all benchmarks (Table 5). This is consistent with prior observations in bounding-box and keypoint regression that ℓ_1-type objectives are more robust for localization. Under SFT, we further measure

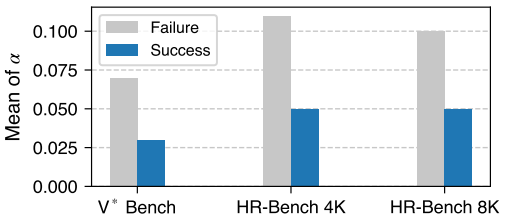


Figure 3: Behavior of α. Successful trajectories exhibit smaller α than failed ones, reflecting higher confidence.

Table 6: Bounding-box IoU on the Vis-CoT-363K dataset. `NV-CoT` substantially improves localization precision over Vis-CoT, with the ℓ_1 loss outperforming the ℓ_2^2 loss.

	Vis-CoT	NV-CoT (ℓ_2^2)	NV-CoT (ℓ_1)
Bbox IoU	47.3	57.4	59.5

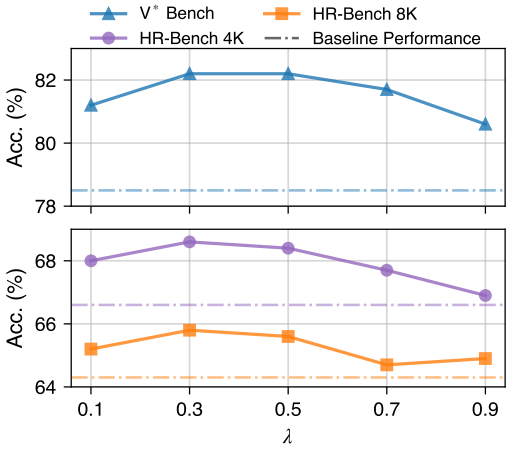


Figure 4: Effect of λ. Performance peaks at $\lambda = 0.3$ and `NV-CoT` consistently outperforms the baseline across all values.

bounding-box IoU on the Vis-CoT-363K validation set. Table 6 shows that `NV-CoT` significantly improves localization accuracy over Vis-CoT, with the ℓ_1 loss achieving the best IoU.

Shared vs. independent σ/α parameterization. We further investigate whether the dispersion parameter should be shared across coordinates or predicted independently. Taking the Gaussian policy as an example, we extend the model with four additional heads $\Sigma = [\mathbf{w}_{\sigma,x_1}, \mathbf{w}_{\sigma,y_1}, \mathbf{w}_{\sigma,x_2}, \mathbf{w}_{\sigma,y_2}]^\top$ to predict per-coordinate standard deviations. Specifically, the independent standard deviations are

$$\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3, \sigma_4]^\top = \max(\Sigma \mathbf{h}, \varepsilon)\,, \tag{13}$$

which generalizes the shared-variance formulation in Eq. (1). Now, the importance ratio changes to:

$$r(\theta) = \prod_{j=1}^{4} \frac{\sigma_{\theta_{\text{old}},j}}{\sigma_{\theta,j}} \exp\left(-\frac{(b_j - \mu_{\theta,j})^2}{2\sigma_{\theta,j}^2} + \frac{(b_j - \mu_{\theta_{\text{old}},j})^2}{2\sigma_{\theta_{\text{old}},j}^2} \right). \tag{14}$$

The Laplace policy follows an analogous procedure. As shown in Table 5, the performance gap between shared and independent parameterizations is marginal across all benchmarks. Given the comparable results, we adopt a shared σ/α in our final model for simplicity and efficiency.

Behavior analysis of α. In Figure 3, we analyze the behavior of the Laplace scale parameter α by comparing its average values across success and failure cases. Results across different image resolutions are normalized. Successful trajectories consistently exhibit smaller α than failed ones. This observation indicates that accurate localization is associated with lower predictive uncertainty, leading the policy to produce more concentrated bounding-box samples.

Effect of λ. In Figure 4, we study the effect of λ in Eq. (2) by varying it from 0.1 to 0.9 and reporting overall accuracy on three benchmarks. `NV-CoT` achieves the best performance at $\lambda = 0.3$, while maintaining relatively stable accuracy across a wide range of values. Notably, across all λ settings, `NV-CoT` consistently outperforms the text-based baselines by a clear margin, indicating that the performance gains are robust and not sensitive to the choice of λ.

RL training diagnostics. To characterize training progress and assess whether either action type dominates optimization, Figure 5 tracks mean reward and tool calls per trajectory, which shows that mean reward rises from 0.2 to 1.4, while tool calls peak at 1.8 per trajectory at 60% of training before declining to 1.3. Table 7 compares their optimization statistics, which shows similar importance-ratio means and clipping fractions for text and coordinate actions, with normalized gradient-norm contributions of 54% and 46%, respectively. Together, these results show no steadily increasing tool use alongside reward gains and no marked optimization imbalance between action types.

Qualitative grounding visualization. In Figure 6, we compare bounding-box localization between `NV-CoT` and the backbone model. The backbone often produces coarse or loosely aligned boxes that include excess background or miss parts of target objects. In contrast, `NV-CoT` produces tighter boxes that better align with target extents, illustrating improved localization.

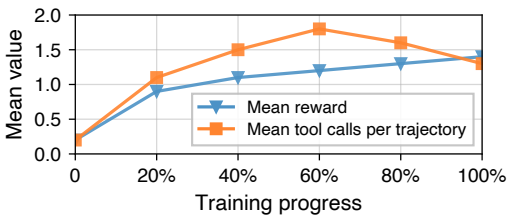


Figure 5: RL training dynamics. Mean reward increases throughout training, while tool calls per trajectory peak at 60% and decline thereafter.

Table 7: RL policy statistics for text and coordinate actions. Clipping uses $\varepsilon = 0.2$; gradient-norm shares are normalized across branches.

Action type	Text tokens	Coordinate actions
Mean ratio	1.012	1.018
Std. ratio	0.071	0.093
Clipped fraction	4.6%	6.8%
Gradient share	54%	46%

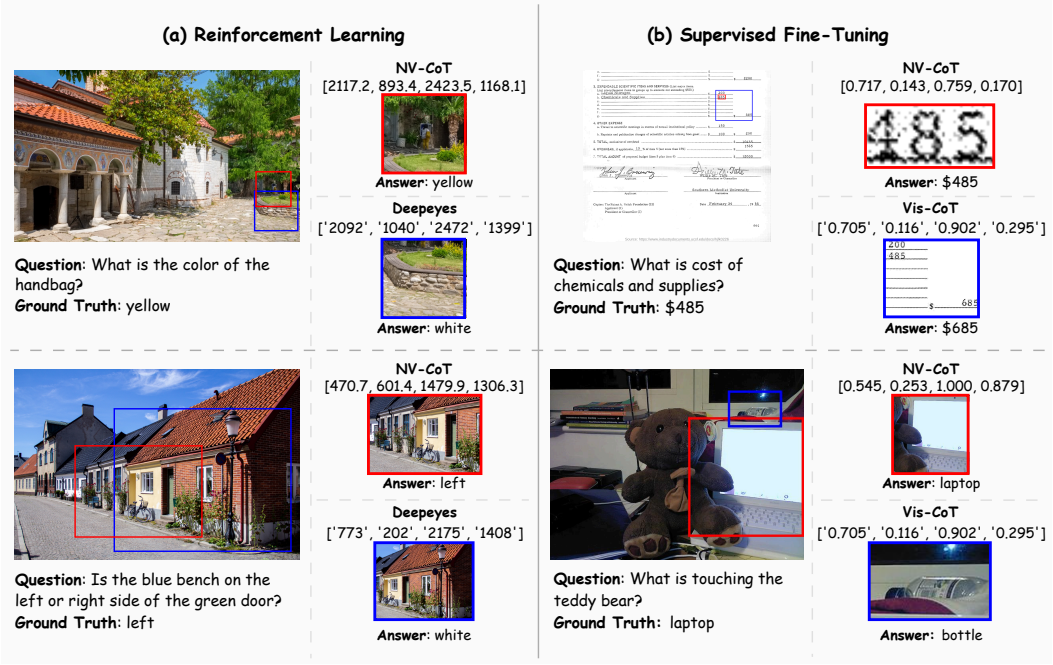


Figure 6: Visualization of bounding boxes. NV-CoT produces more accurate bounding boxes (shown in **red**) compared to the backbone model (shown in **blue**), showing improved localization capability.

5 Conclusion and limitation

We present NV-CoT, a plug-and-play framework that enables MLLMs to directly generate continuous bounding-box coordinates instead of textual tokens. By formulating localization as a continuous-action problem with Gaussian/Laplace policies, NV-CoT supports both supervised fine-tuning and reinforcement learning. NV-CoT mitigates the modality mismatch and semantic fragmentation of text-based visual CoT, while avoiding the fixed-granularity limitation of patch-based designs, with only minimal architectural modifications. Experiments across high-resolution, general multimodal, grounding, and counting benchmarks show consistent gains in visual reasoning and localization, while performance on pure-text tasks remains comparable to the base model. These gains are accompanied by improved final-answer accuracy and faster convergence.

We acknowledge that NV-CoT does not address cases where the underlying MLLM fails to correctly understand the image or the task, as our method focuses on improving visual grounding capability. However, in scenarios where the model has a reasonable understanding but produces inaccurate localization, NV-CoT consistently improves precision through continuous coordinate prediction.

Acknowledgements and Disclosures

Funding. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-030). This research is also supported by National Research Foundation, Singapore, NRF-NRFI10-2024-0004.

Competing Interests. The authors declare no competing interests.

References

[1] Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *ICLR*, 2026.

[2] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*, 2024.

[3] Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. Unsupervised visual chain-of-thought reasoning via preference optimization. In *ICCV*, 2025.

[4] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *ICML*, 2025.

[5] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

[6] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *ICLR*, 2026.

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[8] 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.

[9] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *ICLR*, 2026.

[10] Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, Nancy Chen, Shuicheng Yan, Xulei Yang, and Xun Xu. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. In *ICLR*, 2026.

[11] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

[12] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *AAAI*, 2025.

[13] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025.

[14] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL Findings*, 2022.

[15] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *NeurIPS*, 2025.

[16] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[17] Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, et al. On the generalization of sft: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, volume 2026, pages 27550–27571, 2026.

[18] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943, 2023.

[19] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2815–2826, 2025.

[20] Maolin Wang, Yao Zhao, Jiajia Liu, Jingdong Chen, Chenyi Zhuang, Jinjie Gu, Ruocheng Guo, and Xiangyu Zhao. Large multimodal model compression via iterative efficient pruning and distillation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 235–244, 2024.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

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

[24] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *ECCV*, 2024.

[25] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *ECCV*, 2024.

[26] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *ICLR*, 2025.

[27] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024.

[28] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023.

[29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2025.

[30] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[32] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[33] Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*, 2025.

[34] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. In *NeurIPS*, 2025.

[35] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025.

[36] Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Learning to select visual tools from experience. In *CVPR*, 2026.

[37] Guanghao Zhang, Tao Zhong, Yan Xia, Mushui Liu, Zhelun Yu, Haoyuan Li, Wanggui He, Dong She, Yi Wang, and Hao Jiang. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. In *AAAI*, 2026.

[38] Jiacong Wang, Zijian Kang, Haochen Wang, Liang Xiao, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, and Jun Xiao. Vgr: Visual grounded reasoning. In *ICLR*, 2026.

[39] Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. *arXiv preprint arXiv:2505.18842*, 2025.

[40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *ICLR*, 2025.

[41] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[42] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.

[43] Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *CVPR*, 2025.

[44] Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *EMNLP*, 2025.

[45] Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. Retrieval-augmented perception: High-resolution image perception meets visual RAG. In *ICML*, 2025.

[46] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[47] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[49] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[50] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, 2019.

[51] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

[52] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[53] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

[54] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

[55] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *ICCV*, 2023.

[56] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, 2019.

[57] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022.

[58] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[59] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016.

[60] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[61] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *ACL*, 2024.

[62] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. In *NeurIPS*, 2025.

A Implementation Details

A.1 Datasets

For fair comparison with prior visual chain-of-thought methods, we adopt exactly the same training data as the corresponding backbone models.

Supervised Fine-Tuning (SFT). For supervised fine-tuning, we adopt the Visual-CoT dataset constructed following the protocol of Vis-CoT [2]. The dataset is built by collecting images and annotations from twelve publicly available datasets spanning five domains, including text/document understanding, fine-grained visual recognition, general visual question answering, chart understanding, and relation reasoning. These source datasets mainly consist of VQA and image captioning benchmarks, from which images and useful annotations such as question-answer pairs, captions, object bounding boxes, and object relations are reused. The data construction process involves both linguistic and visual annotation stages. For linguistic annotation, a large language model (GPT-4) is employed to generate or refine question-answer pairs and, where applicable, to produce intermediate reasoning steps. For visual annotation, task-relevant regions are identified to serve as visual chain-of-thought (CoT) grounding signals. In text-centric domains, such as TextVQA [53], DocVQA [54], DUDE [55], SROIE [56], and InfographicsVQA [57], OCR-based tools (e.g., PaddleOCR) are used to detect text regions, and the CoT bounding boxes are defined as regions containing words or sentences that directly support the answer. In fine-grained recognition tasks, bounding boxes corresponding to discriminative object parts or attributes are used to guide fine-grained visual reasoning. For general VQA and relation reasoning tasks, including Flickr30k [58], Visual7W [59], VSR [22], and GQA [60], object-level annotations and scene graph information are leveraged. The visual CoT bounding boxes correspond to objects or regions that are explicitly involved in answering the question, such as spatially related objects or attributes specified in the query. In particular, for GQA, structured semantic annotations are used to identify the target objects and relations, enabling precise grounding of intermediate reasoning steps.

During SFT, each training example consists of an image, a question, a sequence of intermediate visual grounding steps represented by bounding boxes, and the final answer. The model is trained to jointly generate textual outputs and predict continuous bounding-box coordinates, allowing it to learn region-grounded visual reasoning under explicit supervision.

Reinforcement Learning (RL). For reinforcement learning, we adopt the same training data and protocol as DeepEyes [1]. Unlike supervised fine-tuning, the RL setting does not rely on any explicit intermediate bounding-box annotations. Instead, the model is optimized solely through outcome-level supervision, which better reflects practical scenarios where fine-grained visual grounding labels are unavailable.

The RL training data is constructed to emphasize tool-aware visual reasoning and consists of three complementary subsets: fine-grained visual search data, chart understanding data, and general reasoning data. The fine-grained subset is sampled from a portion of the V^* training set [11], focusing on high-resolution images and detailed perception tasks where localized visual inspection is critical for correct answering. The chart subset is drawn from ArxivQA [61], which contains synthetic plots, diagrams, and structured graphical representations, significantly expanding the diversity of visual modalities beyond natural images. To further enhance reasoning diversity, some samples are incorporate from ThinkLite-VL [62], which include multimodal questions requiring arithmetic reasoning, commonsense inference, and multi-step problem solving. Then a dedicated data selection pipeline is applied. First, question difficulty is estimated by prompting a strong vision-language model to generate multiple responses per question and measuring answer consistency; overly trivial or extremely difficult samples are filtered out. Second, original questions are reformulated into open-ended formats to better align with generative training objectives, and samples that cannot be reliably converted are discarded. Third, unverifiable instances are removed, including those with ambiguous questions, incorrect annotations, or unreadable visual content. Finally, chart data is exempted from this filtering step due to its inherent reliance on visual structure, while reasoning data is retained in its original form.

A.2 Evaluation Protocol

Inference Prompts. We employ two different prompt templates for reinforcement learning and supervised fine-tuning, respectively, reflecting their distinct training objectives and supervision signals. For reinforcement learning, we use a structured inference prompt that explicitly encourages step-by-step reasoning and conditional tool usage. The prompt instructs the model to first reason about the question, then optionally invoke the image zoom-in tool when fine-grained visual inspection is necessary, and finally produce the answer in a strictly defined format. This structured output separates reasoning, tool calls, and final answers, enabling reliable trajectory parsing and outcome-driven

reward computation during policy optimization. For supervised fine-tuning, we adopt a localization-oriented prompt that explicitly asks the model to identify a task-relevant image region by predicting a bounding-box coordinate. The model is instructed to answer the question based on both the original image and the localized region. This prompt design encourages the model to learn explicit visual grounding under bounding-box supervision, strengthening the alignment between region selection and downstream reasoning.

RL Evaluating Prompt Template

Think first, call image_zoom_in_tool if needed, then answer.
Format strictly as:
`<think>...</think>`
`<tool_call>...</tool_call>` *(if tools needed)*
`<answer>...</answer>`

SFT Evaluating Prompt Template

Please provide the bounding box coordinate of the region that can help you answer the question better.
Answer the question based on the original image and local image.

A.3 Coordinate and Crop Handling

For iterative zoom-in, predicted bounding boxes use the original image's global coordinate frame. Before cropping, normalized coordinates are converted to pixel coordinates when needed and adjusted for image padding; the final crop window is clipped to the image boundaries. Coordinates with invalid corner ordering are not reordered: the crop is formed around the averaged center with a minimum crop size and then clipped. Invalid coordinate predictions occur more often early in RL training and decrease rapidly; none were observed during inference.

A.4 More Experimental Results



Figure 7: More visualization results of bounding boxes. NV-CoT produces more accurate bounding boxes (shown in **red**) compared to the backbone model (shown in **blue**), demonstrating improved localization capability.

Additional visualization. In Figure 7, we provide additional qualitative examples of bounding-box localization. As in the main-text visualization, the backbone model (blue) tends to produce coarse or loosely aligned boxes, while `NV-CoT` (red) more closely matches object extents, illustrating improved localization quality.

B Impact Statements

This paper presents work whose primary goal is to advance the capabilities of multimodal large language models in visual reasoning, particularly through more accurate and efficient region grounding. By enabling models to reason over images using continuous numerical representations, our approach improves localization precision and training efficiency, which may benefit a wide range of downstream applications such as visual question answering and optical character recognition. At the same time, improved visual localization and reasoning capabilities could be applied in sensitive domains. However, our work does not introduce new data sources, deployable systems, or application-specific pipelines, and it relies on standard benchmark datasets commonly used in the research community. As such, we do not believe that this work introduces novel ethical risks beyond those already well recognized in the field of MLLMs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe continuous numerical box prediction for visual CoT, and §§3 and 4 evaluate the proposed SFT and RL variants on the stated benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §5 discusses the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper presents method derivations and experimental results, but no theorem or formal theoretical result requiring proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §§3 and 4.1 describe the method and training configuration; the appendix provides dataset, evaluation-protocol, and prompt details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The project page links to the public `NV-CoT` code repository. The paper and appendix describe the method, datasets, training configuration, and evaluation prompts needed to reproduce the reported experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: §4.1 reports the SFT and RL training settings and evaluation protocol, while the appendix documents datasets and evaluation prompts.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The reported tables and figures provide point estimates but do not include error bars, confidence intervals, or statistical significance tests.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: §4.1 reports that RL training uses 32 A100 GPUs for 80 iterations, with batch size 256 and GRPO group size 16, specifying the compute hardware and scale used for the main RL experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The experiments use existing benchmark datasets and do not recruit human participants; the paper reports no deviation from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The Impact Statements section discusses potential benefits for visual question answering and OCR and notes that improved visual reasoning could also be used in sensitive domains.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not describe releasing a high-risk dataset, pretrained model, or deployed system that would require release safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper cites the datasets and prior methods it uses, including in §4.1 and the appendix, but does not explicitly list their licenses or terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The manuscript does not describe releasing a new dataset, code package, or model asset; it presents a method and experimental results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study uses existing datasets and does not conduct crowdsourcing or collect new data from human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The study does not involve human-subject experiments or recruit participants, so IRB review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: §A.1 states that GPT-4 is used during linguistic annotation to generate or refine training questions, answers, and reasoning steps; the proposed method itself is described in §3.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.