Start Small, Think Big: Curriculum-based Relative Policy Optimization for Visual Grounding

Qingyang Yan[1*], **Guangyao Chen**[2*†], **Yixiong Zou**[1†]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

Abstract

Chain-of-Thought (CoT) prompting has recently shown significant promise across various NLP and computer vision tasks by explicitly generating intermediate reasoning steps. However, we find that reinforcement learning (RL)-based fine-tuned CoT reasoning can paradoxically degrade performance in Visual Grounding tasks, particularly as CoT outputs become lengthy or complex. Additionally, our analysis reveals that increased dataset size does not always enhance performance due to varying data complexities. Motivated by these findings, we propose **Curriculum-based Relative Policy Optimization (CuRPO)**, a novel training strategy that leverages CoT length and generalized Intersection over Union (gIoU) rewards as complexity indicators to progressively structure training data from simpler to more challenging examples. Extensive experiments on RefCOCO, RefCOCO+, RefCOCOg, and LISA datasets demonstrate the effectiveness of our approach. **CuRPO** consistently outperforms existing methods, including Visual-RFT, reaching a peak improvement of up to 15.49 mAP on RefCOCO. Moreover, **CuRPO** exhibits exceptional efficiency and robustness, delivering strong localization performance even in few-shot learning scenarios, particularly benefiting tasks characterized by ambiguous and intricate textual descriptions.

Code — https://github.com/qyoung-yan/CuRPO

Introduction

Chain-of-Thought (CoT) prompting has recently garnered significant attention within both natural language processing (NLP) and computer vision research communities due to its capability to enhance model interpretability and performance by explicitly generating intermediate reasoning steps (Wei et al. 2022; Kojima et al. 2022; Ge et al. 2023). These explicit reasoning processes have been successfully integrated with reinforcement learning techniques such as Group Relative Policy Optimization (GRPO) (Liu et al. 2025a), significantly boosting performance across a variety of tasks, including question answering, arithmetic reasoning, and visual reasoning tasks (Chen et al. 2024c; Liu et al. 2025a; Tan et al. 2025). Despite these successes, we observe a somewhat counterintuitive phenomenon in the domain of Visual Grounding: explicitly generating CoT reasoning steps does not consistently lead to improved performance and, in certain cases, may even degrade model accuracy, particularly when CoTs become overly lengthy or complex.

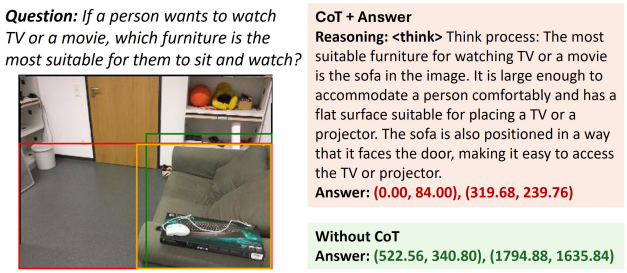

Figure 1: Comparison of visual grounding results with and without CoT. The CoT-guided model produces an incorrect bounding box (red) due to misinterpretation of the textual context. In contrast, the model without CoT successfully identifies the correct furniture for watching TV or a movie (green bounding box).

As illustrated in Figure 1, our experiments reveal that a GRPO-trained model generates incorrect bounding box predictions when explicitly prompted to produce CoT reasoning, while the same model successfully identifies the correct object without generating explicit CoT. This suggests that, within the context of Visual Grounding, explicitly generating CoT may introduce unnecessary complexity, thereby reducing localization accuracy. Further investigation into this phenomenon uncovered another crucial issue: increasing the size of the training dataset for Visual Grounding does not consistently result in improved model performance. Surprisingly, model performance can even deteriorate as more data is added, indicating the presence of varying levels of difficulty within larger datasets. This observation prompts us to reconsider conventional assumptions about data quantity and complexity, motivating a deeper exploration of whether and how the complexity and ordering of training examples might impact the learning dynamics and overall performance of visual grounding models.

[*]These authors contributed equally.
[†]Corresponding author.

Our proposed method, termed **Curriculum-based Relative Policy Optimization (CuRPO)**, explicitly leverages CoT length and reward indicators to structure training in a curriculum manner, progressively increasing task complexity. Specifically, **CuRPO** begins by generating multiple CoT-based reasoning outputs for each training sample, calculating their average length to quantify the inherent complexity of each example. Concurrently, we employ visual reward as an additional reward-based complexity measure, ensuring comprehensive evaluation of each data point's difficulty. By sorting and grouping training examples according to these indicators, **CuRPO** systematically exposes the model first to simpler examples, gradually advancing to more challenging scenarios. This carefully structured progression allows the model to incrementally acquire robust reasoning and precise localization capabilities, significantly enhancing its overall accuracy and training stability, even in settings with limited training data.

Extensive experiments across multiple visual grounding benchmarks, including RefCOCO, RefCOCO+, RefCOCOg, and LISA, substantiate the effectiveness of our proposed method. **CuRPO** consistently demonstrates substantial performance improvements over existing baselines, notably outperforming Visual-RFT (Liu et al. 2025a) by large margins, achieving up to +12.52 mAP on the RefCOCO dataset. Importantly, our approach exhibits remarkable efficiency and effectiveness under few-shot learning scenarios, delivering strong localization performance even with minimal training samples and showing their efficacy in tackling datasets characterized by ambiguous and complex textual descriptions.

In summary, our contributions are threefold:

- We identify and empirically verify that explicitly generating CoT reasoning may negatively impact visual grounding performance.
- We introduce CoT length and reward-based sorting as novel indicators of data complexity, forming the basis for an effective curriculum training strategy (**CuRPO**).
- Our **CuRPO** method demonstrates consistent performance gains across multiple visual grounding benchmarks, significantly improving accuracy, training stability, and exhibiting remarkable effectiveness in few-shot settings.

Related Work

Chain-of-Thought Reasoning. Chain-of-Thought (CoT) prompting asks language models to generate explicit intermediate steps before answering (Wei et al. 2022). By decomposing problems into smaller sub-tasks, CoT improves arithmetic, commonsense, and code-generation benchmarks (Zhang et al. 2024b; Ke et al. 2025). These gains, however, depend strongly on reasoning length and token budget. Longer traces help on difficult items (Fu et al. 2022; Jin et al. 2024) but can degrade accuracy or efficiency on simpler ones due to "over-thinking" (Nayab et al. 2024; Chen et al. 2024a,b; Zou et al. 2024b). Excessive length also hurts generalization when such sequences are under-represented in pre-training corpora (Anil et al. 2022). To mitigate these issues, Wu et al. (2025) propose a length-adaptive scheduler that selects a "balanced" CoT depth based on task hardness and model size, recovering efficiency without sacrificing accuracy. Overall, the evidence suggests that CoT is beneficial only when reasoning depth is matched to instance difficulty. Enumeration-based search such as Self-Consistency (Wang et al. 2023) and Tree-of-Thought (ToT) sampling (Yao et al. 2023) improves robustness by exploring multiple reasoning paths and selecting the highest-scoring answer. Recently, Chain-of-Preference Optimization (CPO) aligns each intermediate step with high-quality ToT traces during fine-tuning, yielding notable gains on question answering, fact verification, and arithmetic reasoning (Zhang et al. 2024a,d). Beyond step selection, Wu et al. (2025) show that extremely long chains do not always improve performance and identify task-specific optimal lengths; they advocate letting models decide when to "stop thinking" rather than imposing a fixed budget. Complementary to this, Li et al. (2025) introduce *SelfBudgeter*, which first predicts the minimal token allocation required for a given query and then generates a response that respects either the self-estimated or a user-defined budget, cutting decoding cost by 35–50% with no accuracy loss.

Curriculum Training Curriculum Training (CT) presents data in an easy-to-hard order to speed convergence and improve robustness (Bengio et al. 2009; Hacohen and Weinshall 2019; Zhang et al. 2024c; Zou et al. 2024d). While CT works well in vision, NLP, and multimodal tasks, defining difficulty is challenging when complexity is semantic, perceptual, and relational. FASTCURL (Song et al. 2025) sidesteps manual heuristics by using prompt length as a proxy for reasoning complexity, gradually widening the context window. This staged exposure reduces redundant reasoning and reaches state-of-the-art accuracy on MATH 500 and OlympiadBench with fewer updates. Building on these ideas, Song et al. (2025); Zou et al. (2024c) extend curriculum reinforcement learning to reasoning models: they segment data by input length and progressively enlarge the context window during RLHF, achieving 49.6% on AIME 2024 with a 1.5B-parameter model. Their ablation shows that both length-aware segmentation and progressive window extension are necessary; removing either halves the gain. Apart from prompt length, other works estimate difficulty by model-based uncertainty (Liu, Feng, and Schütze 2024), latent semantic novelty (Xu, Li, and Tao 2024; Zou et al. 2024a, 2021; Zhang et al. 2024c; Xiao et al. 2025; Zeng et al. 2025; Chen et al. 2025), or retrieval distance (Chen, Zhao, and Chang 2024). Adaptive schedulers that jointly tune pacing and difficulty metric—e.g., Dynamic Temperature Scaling (DTS) (Kim et al. 2025)—achieve faster convergence and better out-of-distribution robustness. However, sub-optimal curricula can still stall learning or trap models in local minima, underscoring the need for principled theory that couples data, capacity, and task structure.

Role of CoT in Visual Grounding

A *Chain-of-Thought* (CoT) explicitly guides large language models (LLMs) through intermediate reasoning steps, which

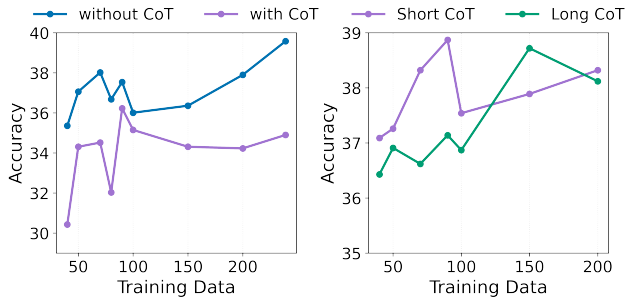

Figure 2: (a) Accuracy comparison with and without CoT. (b) Performance comparison between short and long CoTs.

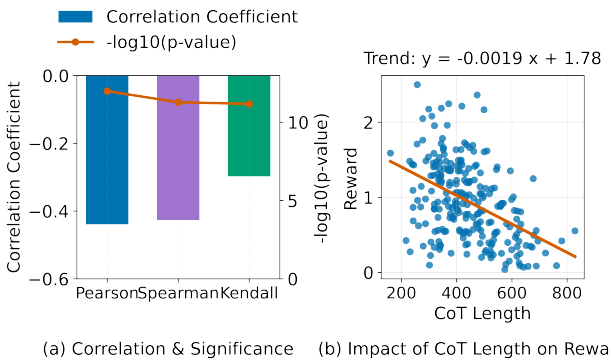

Figure 3: (a) Correlation coefficients between CoT length and reward. (b) Negative impact of increased CoT length on reward.

significantly improves their accuracy in arithmetic, commonsense reasoning, and symbolic reasoning tasks (Wei et al. 2022; Kojima et al. 2022). Recent vision-language models (VLMs) adapt CoT via prompt tuning or supervised fine-tuning methods, demonstrating effectiveness in multi-modal scenarios (Ge et al. 2023; Chen et al. 2024c). Beyond supervised methods, reinforcement learning (RL) provides another mechanism for enhancing reasoning capabilities. Specifically, *Group Relative Policy Optimization* (GRPO) (Shao et al. 2024) modifies standard Proximal Policy Optimization by employing variance-normalized relative advantage comparisons within generated response groups, removing the dependency on a learned critic, thereby achieving superior training stability and data efficiency (Guo et al. 2025; Shao et al. 2024). GRPO has successfully enhanced reasoning performance in mathematical tasks (Guo et al. 2025) and universal visual grounding (Bai et al. 2025), highlighting its potential for broader visual reasoning contexts.

Although CoT prompting generally benefits reasoning tasks, recent evidence suggests this improvement may not extend straightforwardly to visual grounding scenarios (Wu et al. 2025). Specifically, longer reasoning chains may introduce excessive complexity or noisy reasoning steps, adversely impacting localization performance. To investigate this counterintuitive phenomenon, we systematically analyze how generating CoT affects visual grounding accuracy and explore whether the length of the CoT correlates with task difficulty.

Do CoTs Help Visual Grounding?

In our experiments, we fine-tune the visual grounding model using the GRPO algorithm guided by a reward function from (Liu et al. 2025a). To evaluate the impact of explicitly outputting Chain-of-Thought (CoT) reasoning steps, we compare two experimental conditions (Table 1): (1) a scenario where the model directly outputs the final grounding results based on the given image and textual query (the "no CoT" condition), and (2) a scenario where the model explicitly generates intermediate CoT reasoning steps before producing the final prediction. As shown in Figure 1, our findings reveal that the model consistently achieves superior performance when it is not required to explicitly output CoT. As shwon in Figure 2(a), at a small dataset size of 40 samples, the model without CoT generation achieves an mIoU of 35.6, outperforming the CoT-outputting counterpart, which only reaches an mIoU of 34.3 even with a significantly larger dataset of 239 samples. Moreover, as the dataset expands, the performance gap widens further, with the direct-output model attaining a maximum mIoU of 39.6. These observations challenge the prevailing assumption that explicit intermediate reasoning steps invariably enhance model performance. Our empirical results suggest that, in visual grounding tasks, requiring explicit CoT generation can be unnecessary or even detrimental, particularly at smaller data scales, indicating potential inefficiencies or noise introduced by overly verbose reasoning chains.

Does Longer CoT Mean Higher Difficulty?

To understand whether CoT length significantly influences visual grounding performance, we empirically evaluate how variations in reasoning chain length correlate with model accuracy. Specifically, we generate multiple candidate CoTs for each visual grounding sample, compute their average length, and subsequently group samples into subsets of shorter and longer CoTs. Models trained on shorter-CoT subsets consistently achieve higher grounding accuracy (see Figure 2(b)), suggesting that simpler reasoning chains are more beneficial during early learning stages. Moreover, models initially struggle with longer CoT samples—reflected by reduced accuracy—indicating increased task complexity associated with extended reasoning chains. Nonetheless, after continued training on larger datasets containing longer-CoT data, the model progressively improves and eventually surpasses performance obtained from shorter-CoT data. This finding clearly suggests that CoT length acts as an implicit indicator of task difficulty, directly impacting model learning dynamics and overall performance.

To theoretically underpin this empirical observation, we posit that the complexity of reasoning tasks increases with the number of required reasoning steps. Consider a simplified probabilistic model, where the probability of successfully completing a reasoning chain consisting of C indepen-

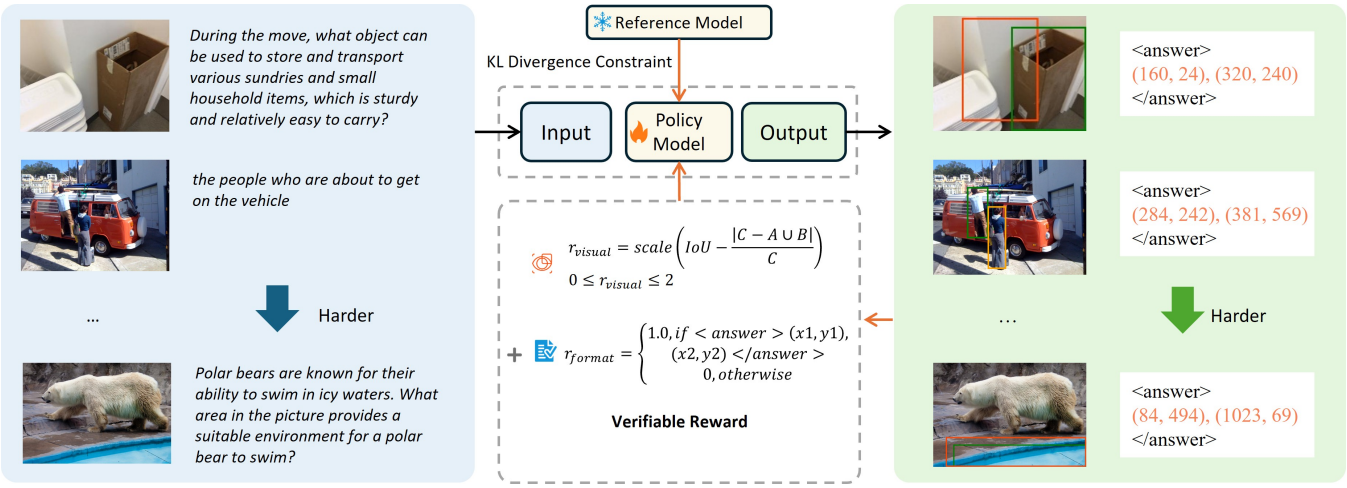

Figure 4: Overview of curriculum-based GRPO training framework for visual grounding. We first sort training examples by the complexity indicated by their CoT length, from simplest (shortest CoT) to hardest (longest CoT). Each query-image pair is fed into a policy model constrained by KL-divergence with a reference model. The model outputs bounding boxes and receives a combined reward incorporating visual accuracy (scaled gIoU) and format correctness. This curriculum strategy progressively guides the model from simpler to increasingly complex visual grounding tasks.

dent steps, each with success probability p_c, is given by: $\Pr(\text{success}) = \prod_{c=1}^{C} p_c$. Assuming $p_c < 1$, increasing the number of reasoning steps C (i.e., CoT length) exponentially decreases the probability of overall success, making longer reasoning chains inherently more challenging. Empirically, we verify this theoretical assumption by analyzing the correlation between CoT length and the generalized Intersection-over-Union (gIoU)-based reward. As shown in Figure 3(a), our statistical analysis reveals a clear negative correlation between CoT length and reward: the Pearson correlation coefficient is -0.4395 ($p = 1.04 \times 10^{-12}$), Spearman's rank correlation coefficient is -0.4268 ($p = 5.34 \times 10^{-12}$), and Kendall's Tau coefficient is -0.2981 ($p = 6.81 \times 10^{-12}$). These results robustly confirm that increased CoT length systematically corresponds to decreased reward, reflecting higher task difficulty. Additionally, Figure 3(b) visually confirms this inverse relationship: longer CoTs are systematically associated with lower rewards, reflecting increased difficulty. Taken together, these theoretical and empirical analyses affirm that CoT length serves as a robust signal of task complexity in visual grounding tasks.

Does More Data Always Mean Better?

Further investigation into the impact of training data scale on visual grounding performance reveals intriguing trends. As depicted in Figure 2(a), simply increasing the size of the training dataset does not guarantee consistent performance improvements. Surprisingly, the model explicitly generating CoT demonstrates significant performance instability, with accuracy initially increasing but subsequently dropping and remaining stagnant even as more data is introduced. Conversely, the direct-output model without CoT consistently improves with increasing data size, highlighting the influence of output strategies on performance stability. Furthermore, Figure 2(b) examines performance variations in terms of CoT length, indicating that models trained on shorter-CoT subsets initially exhibit higher accuracy compared to those trained on longer-CoT subsets. Nevertheless, with larger training datasets, the performance gap diminishes, suggesting that longer reasoning chains, despite their inherent complexity, become manageable as the model gradually adapts. These observations collectively emphasize the importance of data complexity and suggest that merely enlarging the dataset size is insufficient. Instead, careful consideration of data ordering and complexity progression may enhance the effectiveness of training visual grounding models.

Direct prompt	Question Output your grounding box. Following "`<answer>(x1,y1),(x2,y2)</answer>`" format.
CoT prompt	Question Output the thinking process in "`<think>...</think>`" and then the grounding box, following the format: "`<think>reasoning chain</think><answer>(x1,y1),(x2,y2)</answer>`".

Table 1: Prompts used to construct the dataset. Prompts are shown separately for direct output and CoT-based output.

Method

Motivated by our key empirical observation that longer Chain-of-Thought (CoT) reasoning is associated with increased task difficulty and lower rewards, we propose a novel training framework termed Curriculum-based Relative Policy Optimization (CuRPO). CuRPO leverages a curriculum learning paradigm to progressively guide the

model from simpler visual grounding tasks, characterized by shorter CoTs, to more complex tasks involving longer and more intricate reasoning processes. By systematically organizing and gradually introducing training examples based on CoT length and their corresponding reward signals, CuRPO effectively enhances the model's reasoning capacity and localization accuracy, while maintaining training stability and efficiency.

Reward Function Design

In visual grounding tasks, a standard metric for assessing localization accuracy is Intersection over Union (IoU). However, IoU cannot provide useful feedback when predicted and ground-truth bounding boxes do not overlap, as the intersection area becomes zero. To resolve this limitation, we adopt Generalized IoU (gIoU) (Rezatofighi et al. 2019), which considers both overlap and spatial relationships between bounding boxes. Formally, gIoU is defined as:

$$\text{gIoU}(A, B) = \text{IoU}(A, B) - \frac{C - (A \cup B)}{C}, \qquad (1)$$

where C denotes the area of the smallest enclosing box containing predicted box A and ground-truth box B. This allows meaningful feedback even without overlap.

We define our overall reward function as:

$$R_d = R_{\text{visual}} + R_{\text{format}}, \qquad (2)$$

where R_{visual} is derived from the gIoU metric, and R_{format} ensures compliance with output formatting requirements. To stabilize training, we linearly rescale gIoU values from their original range $[-1, 1]$ to $[0, 2]$, reducing overly negative feedback and enhancing the visibility of spatial cues.

Curriculum Training Process

Given our observation that CoT length influences task difficulty, we propose a curriculum training strategy based on sorting examples by their CoT lengths. Specifically, we first instruct a pretrained VLM to generate multiple CoTs (typically 8 per sample) for each data point, explicitly capturing its reasoning process.

Next, we sort training examples by their average CoT length (shortest to longest). Shorter CoTs typically correspond to simpler reasoning tasks and thus provide a natural ordering of task complexity. We then sequentially introduce samples into training:

1. Initial training phase: the model sees only samples with short CoTs, learning fundamental visual reasoning patterns.
2. Intermediate phase: gradually introduce medium-length CoTs, exposing the model to moderately complex reasoning scenarios.
3. Advanced training phase: incorporate longer CoTs, challenging the model to reason through increasingly complex visual grounding tasks.

This progressive approach ensures the model acquires robust reasoning capabilities before tackling the most difficult examples.

Algorithm 1: Curriculum-based Relative Policy Optimization (CuRPO)

Require: Dataset $\mathcal{D}$; initial policy π_θ; reference policy π_{ref}; **SortCriterion** $\in$ {*Length*, *Reward*}; total training steps T; group size G; clip ratio ϵ; KL weight β
Ensure: Optimized policy π_θ

```
Compute complexity score s(x) for each sample x ∈ D according to SortCriterion
Sort D in ascending order of s(x) and split into curriculum phases {D_1, ..., D_M}
for phase m = 1 to M do
    for t = 1 to T/M do
        Sample mini-batch B ⊂ D_m
        for all (q, I) ∈ B do
            Generate G candidate outputs {o_i}_{i=1}^G ~ π_θ(o | q, I)
            Compute reward r'_i = R_visual(o_i) + R_format(o_i)
            Compute μ' = (1/G) Σ_{j=1}^G r'_j,  σ' = sqrt((1/G) Σ_{j=1}^G (r'_j − μ')^2)
            A_i ← (r'_i − μ') / σ'  {group-normalized advantage}
            c_i ← π_θ(o_i | q, I) / π_old(o_i | q, I)
            L_i ← min(c_i A_i, clip(c_i, 1 − ε, 1 + ε) A_i)
        end for
        θ ← θ + η ∇_θ ( (1/|B|G) Σ_{(q,I)∈B} Σ_{i=1}^G L_i − β D_KL(π_θ ‖ π_ref) )
    end for
end for
return π_θ
```

GRPO-based Training Objective

Inspired by the Visual-RFT (Liu et al. 2025a), we employ Group Relative Policy Optimization (GRPO) (Shao et al. 2024) as our reinforcement learning backbone. GRPO optimizes relative advantages within groups of generated responses, significantly enhancing sample efficiency and stability compared to standard PPO variants.

Given a query q, the current policy π_θ generates a group of G candidate outputs $\{o_i\}$, each associated with a modified reward $r'_i = R_d(o_i|q)$ that implicitly penalizes long CoTs through lower gIoU scores. We first compute group-normalized advantages:

$$A_i = \tfrac{r'_i - \mu'}{\sigma'}, \quad \mu' = \tfrac{1}{G}\textstyle\sum_{j=1}^G r'_j, \quad \sigma' = \sqrt{\tfrac{1}{G}\textstyle\sum_{j=1}^G (r'_j - \mu')^2}. \quad (3)$$

The final GRPO objective is defined as a clipped surrogate loss plus KL-divergence regularization to maintain policy stability:

$$L_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} \min\big(c_i A_i,\ \text{clip}(c_i, 1-\epsilon, 1+\epsilon) A_i\big) - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot|q) \,\|\, \pi_{\text{ref}}(\cdot|q)\big), \qquad (4)$$

where $c_i = \frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}$ represents the probability ratio between new and old policies, and π_{ref} serves as a stable reference policy.

Method	Model	#Train	mIoU
SFT (Liu et al. 2025b)	Qwen2-VL-2B	239	29.7
OV-Seg (Liang et al. 2023)	OV-Seg	239	30.5
GroundingDINO (Liu et al. 2024)	X-Decoder	239	28.5
GroundedSAM (Zou et al. 2023)	X-Decoder	239	28.6
Visual-RFT (Liu et al. 2025a)	Qwen2-VL-2B	239	34.4
CuRPO (Ours)	Qwen2-VL-2B	50	37.4 (+3.0)
CuRPO (Ours)	Qwen2-VL-2B	200	**38.7** (+4.3)
CuRPO (Ours)	Qwen2-VL-2B	239	**38.4** (+4.0)

Table 2: Results on LISA. Visual grounding performance for different methods and training-data sizes. To save space, the citation for each baseline is placed on a second line within the same cell. Each experiment is repeated three times with different random seeds.

Algorithm 1 condenses the entire CuRPO pipeline. We first assign each image–query pair a *complexity score*, either the average CoT length or its corresponding reward, and sort the dataset from easiest to hardest before splitting it into curriculum phases. Within every phase, the policy π_θ generates G candidate responses, receives a combined reward that blends scaled gIoU and format-correctness, and converts these scores into *group-normalised advantages*. CuRPO then maximises a clipped surrogate loss while regularising with a KL term that keeps the updated policy close to a reference model π_{ref}. By gradually unlocking harder phases, the policy is forced to master short, easy reasoning chains first and then adapt to longer, more complex CoTs.

Experiments

Experimental Setup

Implementation Details. Building on preliminary insights, we refine our curriculum training by using a more granular sorting strategy. Specifically, we sort training examples first by CoT length and then within each CoT-length bin (interval = 50 tokens) by reward values. This ensures that the model sees easiest examples (short CoTs and high reward) first, and gradually transitions to more complex ones. By progressively introducing longer CoTs and lower-reward samples, the curriculum training better scaffolds the model's visual reasoning development. We conduct training using the Qwen2-VL-2B model (Wang et al. 2024) fine-tuned with GRPO. Our baseline is the zero-curriculum fine-tuned Qwen-VL-2B model in the "with CoT" setting: the model is required to generate a full Chain-of-Thought before outputting its final bounding box. We evaluate its mIoU and mAP performance to compare against our curriculum-trained versions.

Datasets and Metrics. We conduct experiments on two standard vision-language grounding datasets: RefCOCO and LISA. **RefCOCO** (Yu et al. 2016) is a widely used referring-expression comprehension benchmark, where the model must localize objects referenced by natural language. **LISA** (Lai et al. 2024) is a multi-scene visual grounding benchmark requiring robust localization under complex layouts. We report class-wise Average Precision (AP) and mean Average Precision (mAP) following standard evaluation protocols (Padilla, Netto, and Da Silva 2020).

Dataset	Method	mAP
RefCOCO (val)	Qwen2-VL-2B	11.57
	Visual-RFT (Liu et al. 2025a)	21.28
	CuRPO (Random)	33.09 (+11.81)
	CuRPO (Length)	**33.80 (+12.52)**
	CuRPO (Reward)	32.64 (+11.36)
RefCOCO (test)	Qwen2-VL-2B	10.70
	Visual-RFT (Liu et al. 2025a)	20.38
	CuRPO (Random)	29.92 (+9.54)
	CuRPO (Length)	**31.42 (+11.04)**
	CuRPO (Reward)	27.89 (+7.51)
RefCOCO+ (val)	Qwen2-VL-2B	13.72
	Visual-RFT (Liu et al. 2025a)	18.41
	CuRPO (Random)	26.82 (+8.41)
	CuRPO (Length)	26.18 (+7.77)
	CuRPO (Reward)	**26.85 (+8.44)**
RefCOCO+ (test)	Qwen2-VL-2B	16.11
	Visual-RFT (Liu et al. 2025a)	20.90
	CuRPO (Random)	24.34 (+3.44)
	CuRPO (Length)	**25.10 (+4.20)**
	CuRPO (Reward)	23.55 (+2.65)
RefCOCOg (val)	Qwen2-VL-2B	14.89
	Visual-RFT (Liu et al. 2025a)	23.39
	CuRPO (Random)	27.98 (+4.59)
	CuRPO (Length)	29.27 (+5.88)
	CuRPO (Reward)	**32.65 (+9.26)**

Table 3: Comparison results (mAP) of different methods on RefCOCO, RefCOCO+ and RefCOCOg datasets. Each CuRPO experiment is repeated 30 times with six different random seeds.

Main Results

Results on LISA. The experimental results on the LISA dataset are summarized in Table 2. Our proposed **CuRPO** consistently outperforms all baseline methods. Notably, with only 50 training examples, **CuRPO** achieves an mIoU of 37.4, exceeding the strongest baseline Visual-RFT (Liu et al. 2025a) trained on all 239 examples by +3.0 points. Increasing the training data to 200 examples further boosts performance to 38.7 mIoU, with gains remaining above +4.0 points over Visual-RFT when using all 239 training examples. These results support our hypothesis that curriculum learning based on CoT length substantially enhances visual grounding performance, especially in low-data regimes, with improvements that grow and then saturate as more training data becomes available.

Results on RefCOCO, RefCOCO+ and RefCOCOg. Table 3 summarizes the experimental results on RefCOCO, RefCOCO+, and RefCOCOg. Across all splits, our **CuRPO** variants consistently outperform the baseline models (Qwen2-VL-2B and Visual-RFT (Liu et al. 2025a)),

highlighting the effectiveness of curriculum-based training. For the RefCOCO dataset, **CuRPO (Length)** achieves the highest mAP of 33.80 and 31.42 on the validation and test sets, respectively, surpassing Visual-RFT by +12.52 and +11.04. Even **CuRPO (Random)** already provides large gains (33.09/29.92), achieving a peak improvement of up to 15.49 mAP, which indicates that curriculum RL itself is highly beneficial, while ordering examples by CoT length further improves performance. On the RefCOCO+ dataset, all CuRPO variants bring clear improvements over Visual-RFT. **CuRPO (Reward)** attains the best validation mAP of 26.85 (+8.44), whereas **CuRPO (Length)** yields the best test mAP of 25.10 (+4.20). The close performance of length-based and reward-based curricula suggests that structured difficulty signals become increasingly useful as textual descriptions grow more complex. For RefCOCOg, which contains longer and more ambiguous descriptions, **CuRPO (Reward)** achieves the highest mAP of 32.65, a substantial +9.26 improvement over Visual-RFT, while the length-based curriculum also performs strongly. These results collectively confirm that curriculum learning based on CoT length and reward signals substantially boosts visual grounding performance across datasets of varying difficulty.

Ablation studies

Impact of Sorting Strategies. As shown in Figure 5 and Table 3, all three sorting strategies yield substantial gains over the non-curriculum baseline. On RefCOCO, **CuRPO (Length)** consistently outperforms both **CuRPO (Random)** and **CuRPO (Reward)**, indicating that ordering samples by CoT length is particularly effective on this relatively simple dataset. For RefCOCO+, **CuRPO (Reward)** achieves the best validation mAP, while **CuRPO (Length)** performs best on the test split, showing that both length- and reward-based curricula provide strong and complementary benefits. On the more challenging RefCOCOg dataset, **CuRPO (Reward)** clearly dominates, suggesting that directly exploiting reward signals becomes more important when descriptions are long and ambiguous. Notably, even **CuRPO (Random)**—which does not explicitly encode task difficulty—still brings large improvements over the baseline, implying that curriculum RL without explicit CoT generation is already beneficial. Overall, leveraging CoT length and reward signals to structure training provides robust gains, with reward-based ordering particularly advantageous on the most difficult dataset.

Effect of CoT across Data Scales. In Figure 5, we further investigate model performance across different numbers of training samples. Interestingly, results indicate that the model explicitly generating CoTs (Visual-RFT) consistently underperforms the same model trained without explicit CoT generation. This aligns with our earlier hypothesis that explicit CoT generation in visual grounding introduces unnecessary complexity, potentially confusing the model and negatively affecting localization accuracy. In contrast, models trained without explicit CoT generation consistently achieve better accuracy, verifying that direct learning reduces reasoning ambiguity, thereby simplifying task acquisition. Moreover, Figure 5 demonstrates that CuRPO-Length

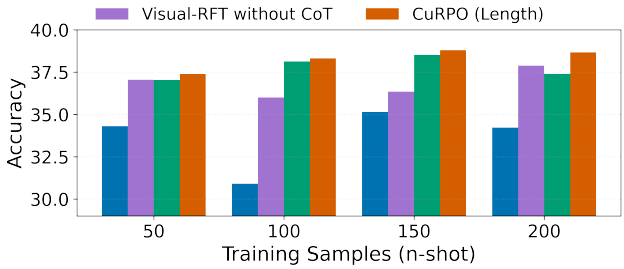

Figure 5: (Comparison of visual grounding accuracy across different training sample sizes and sorting methods.

consistently outperforms both baseline conditions (Visual-RFT with and without CoT) across varying amounts of data. Notably, CuRPO methods show steady improvements as the dataset size increases, highlighting the advantages of curriculum-based approaches. This suggests that structured task complexity progression significantly enhances model generalization, especially as more data becomes available.

Superior Few-Shot Performance. Our analysis emphasizes CuRPO's remarkable effectiveness in few-shot scenarios (50 samples). At this scale, CuRPO methods substantially outperform baselines, demonstrating that explicitly removing the requirement for CoT generation allows the model to rapidly learn visual reasoning patterns without unnecessary intermediate reasoning complexity. The rapid performance gain observed in few-shot conditions underscores CuRPO's capacity for efficiently leveraging limited data. Thus, our proposed framework proves particularly advantageous when faced with limited data resources, quickly capturing essential grounding capabilities by reducing the cognitive overhead introduced by unnecessary reasoning steps.

Conclusion

In this paper, we have systematically investigated the role of CoT prompting in visual grounding tasks, revealing that explicitly generating CoT does not universally benefit model performance and may even hinder accuracy, especially with longer reasoning chains. Motivated by our empirical analysis, we introduced CuRPO, a reinforcement learning method that leverages CoT length and reward signals to progressively structure the training complexity. CuRPO effectively guides the model from simpler reasoning tasks to increasingly challenging ones, resulting in substantial improvements in localization accuracy, enhanced reasoning capabilities, and stable training behavior even under few-shot conditions. Extensive experiments across diverse visual grounding benchmarks validate that our curriculum-driven strategy consistently outperforms existing approaches, demonstrating the importance of carefully ordered training data for optimizing reasoning-intensive vision-language tasks. Future work will focus on evaluating and adapting this curriculum–GRPO paradigm to other vision–language and multimodal reasoning tasks to assess its generalization capacity beyond visual grounding.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grants 62206102, 62436003, 62376103, 62302184, and 62402015; the National Key Research and Development Program of China under grant 2024YFC3307900; the Major Science and Technology Project of Hubei Province under grants 2025BAB011 and 2024BAA008; the Hubei Science and Technology Talent Service Project under grant 2024DJC078; Ant Group through the CCF–Ant Research Fund; the Postdoctoral Fellowship Program of the China Postdoctoral Science Foundation under grant GZB20230024; and the China Postdoctoral Science Foundation under grant 2024M750100. Computations were performed on the HPC Platform of Huazhong University of Science and Technology.

References

Anil, C.; Wu, Y.; Andreassen, A.; Lewkowycz, A.; Misra, V.; Ramasesh, V.; Slone, A.; Gur-Ari, G.; Dyer, E.; and Neyshabur, B. 2022. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35: 38546–38556.

Bai, S.; Li, M.; Liu, Y.; Tang, J.; Zhang, H.; Sun, L.; Chu, X.; and Tang, Y. 2025. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Chen, G.; Horstmann, K.; Wang, Z.; and You, F. 2025. Automated Essential Concept Discovery for Few-Shot Out-of-Distribution Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3964–3974.

Chen, L.; Davis, J. Q.; Hanin, B.; Bailis, P.; Stoica, I.; Zaharia, M. A.; and Zou, J. Y. 2024a. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Advances in Neural Information Processing Systems*, 37: 45767–45790.

Chen, X.; Xu, J.; Liang, T.; He, Z.; Pang, J.; Yu, D.; Song, L.; Liu, Q.; Zhou, M.; Zhang, Z.; et al. 2024b. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Chen, Z.; Zhao, W.; and Chang, K. 2024. Retrieval-Distance Guided Curriculum for Efficient Large Language Model Pretraining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 1234–1248.

Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Sun, Z.; Gutfreund, D.; and Gan, C. 2024c. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1254–1262.

Fu, Y.; Peng, H.; Sabharwal, A.; Clark, P.; and Khot, T. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Ge, J.; Luo, H.; Qian, S.; Gan, Y.; Fu, J.; and Zhang, S. 2023. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hacohen, G.; and Weinshall, D. 2019. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, 2535–2544. PMLR.

Jin, M.; Yu, Q.; Shu, D.; Zhao, H.; Hua, W.; Meng, Y.; Zhang, Y.; and Du, M. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Ke, Z.; Jiao, F.; Ming, Y.; Nguyen, X.-P.; Xu, A.; Long, D. X.; Li, M.; Qin, C.; Wang, P.; Savarese, S.; et al. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Kim, T.; Kim, G.; Cho, C.; and Lee, Y. H. 2025. Naturalness-Aware Curriculum Learning with Dynamic Temperature for Speech Deepfake Detection. *arXiv preprint arXiv:2505.13976*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, Z.; Dong, Q.; Ma, J.; Zhang, D.; and Sui, Z. 2025. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. *arXiv preprint arXiv:2505.11274*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7061–7070.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Liu, Y.; Feng, Q.; and Schütze, H. 2024. PACEDD: Paced-Curriculum Distillation with Prediction and Difficulty Dynamics. *arXiv preprint arXiv:2406.12345*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025a. Visual-RFT: Visual Reinforcement Fine-Tuning. *CoRR*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025b. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Nayab, S.; Rossolini, G.; Simoni, M.; Saracino, A.; Buttazzo, G.; Manes, N.; and Giacomelli, F. 2024. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*.

Padilla, R.; Netto, S. L.; and Da Silva, E. A. 2020. A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*, 237–242. IEEE.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Song, M.; Zheng, M.; Li, Z.; Yang, W.; Luo, X.; Pan, Y.; and Zhang, F. 2025. Fastcurl: Curriculum reinforcement learning with progressive context extension for efficient training r1-like reasoning models. *arXiv e-prints*, arXiv–2503.

Tan, H.; Ji, Y.; Hao, X.; Lin, M.; Wang, P.; Wang, Z.; and Zhang, S. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Y.; Wang, Y.; Ye, Z.; Du, T.; Jegelka, S.; and Wang, Y. 2025. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.

Xiao, D.; Chen, G.; Peng, P.; Huang, Y.; Zhao, Y.; Dai, Y.; and Tian, Y. 2025. When Every Millisecond Counts: Real-Time Anomaly Detection via the Multimodal Asynchronous Hybrid Network. In *Forty-second International Conference on Machine Learning*.

Xu, Y.; Li, M.; and Tao, C. 2024. Latent Semantic Novelty as a Difficulty Signal for Curriculum Learning. *arXiv preprint arXiv:2402.09876*. To appear in ACL 2024.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European conference on computer vision*, 69–85. Springer.

Zeng, Y.; Wu, H.; Nie, W.; Chen, G.; Zheng, X.; Shen, Y.; Peng, J.; Tian, Y.; and Ji, R. 2025. From Objects to Events: Unlocking Complex Visual Understanding in Object Detectors via LLM-guided Symbolic Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 24380–24391.

Zhang, X.; Du, C.; Pang, T.; Liu, Q.; Gao, W.; and Lin, M. 2024a. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37: 333–356.

Zhang, Y.; Du, L.; Cao, D.; Fu, Q.; and Liu, Y. 2024b. An examination on the effectiveness of divide-and-conquer prompting in large language models. *arXiv preprint arXiv:2402.05359*.

Zhang, Z.; Chen, G.; Zou, Y.; Huang, Z.; Li, Y.; and Li, R. 2024c. Micm: Rethinking unsupervised pretraining for enhanced few-shot learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7686–7695.

Zhang, Z.; Chen, G.; Zou, Y.; Li, Y.; and Li, R. 2024d. Learning Unknowns from Unknowns: Diversified Negative Prototypes Generator for Few-Shot Open-Set Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6053–6062.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15116–15127.

Zou, Y.; Liu, Y.; Hu, Y.; Li, Y.; and Li, R. 2024a. Flatten Long-Range Loss Landscapes for Cross-Domain Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23575–23584.

Zou, Y.; Ma, R.; Li, Y.; and Li, R. 2024b. Attention Temperature Matters in ViT-Based Cross-Domain Few-Shot Learning. *Advances in Neural Information Processing Systems 37*.

Zou, Y.; Yi, S.; Li, Y.; and Li, R. 2024c. A Closer Look at the CLS Token for Cross-Domain Few-Shot Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zou, Y.; Zhang, S.; Chen, G.; Tian, Y.; Keutzer, K.; and Moura, J. M. 2021. Annotation-efficient untrimmed video action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 487–495.

Zou, Y.; Zhang, S.; Zhou, H.; Li, Y.; and Li, R. 2024d. Compositional Few-Shot Class-Incremental Learning. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 62964–62977. PMLR.