OpenVLThinkerV2: A Generalist Multimodal Reasoning Model for Multi-domain Visual Tasks

Wenbo Hu **Xin Chen** **Yan Gao-Tian** **Nanyun Peng** **Kai-Wei Chang**
University of California, Los Angeles (UCLA)
{whu, kwchang}@cs.ucla.edu
Project Page GitHub

Abstract

Group Relative Policy Optimization (GRPO) has emerged as the de facto Reinforcement Learning (RL) objective driving recent advancements in Multimodal Large Language Models. However, extending this success to open-source multimodal generalist models remains heavily constrained by two primary challenges: the extreme variance in reward topologies across diverse visual tasks, and the inherent difficulty of balancing fine-grained perception with multi-step reasoning capabilities. To address these issues, we introduce **Gaussian GRPO (G^{2}RPO)**, a novel RL training objective that replaces standard linear scaling with non-linear distributional matching. By mathematically forcing the advantage distribution of any given task to strictly converge to a standard normal distribution, $\mathcal{N}(0,1)$, G^{2}RPO promotes advantage-scale equity across tasks, mitigates vulnerabilities to heavy-tail outliers, and offers symmetric update for positive and negative rewards. Leveraging the enhanced training stability provided by G^{2}RPO, we introduce two task-level shaping mechanisms to seamlessly balance perception and reasoning. First, response length shaping dynamically elicits extended reasoning chains for complex queries while enforce direct outputs to bolster visual grounding. Second, entropy shaping tightly bounds the model's exploration zone, effectively preventing both entropy collapse and entropy explosion. Integrating these methodologies, we present **OpenVLThinkerV2**, a highly robust, general-purpose multimodal model. Extensive evaluations across 18 diverse benchmarks demonstrate its superior performance over strong open-source and leading proprietary frontier models.

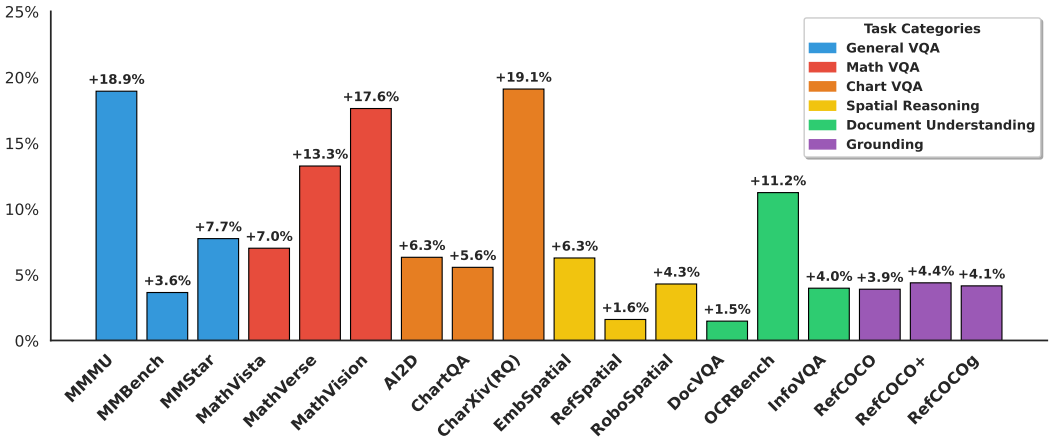

Figure 1: Performance improvement (relative) of OpenVLThinkerV2 over its baseline Qwen3-VL-Instruct-8B across diverse visual tasks.

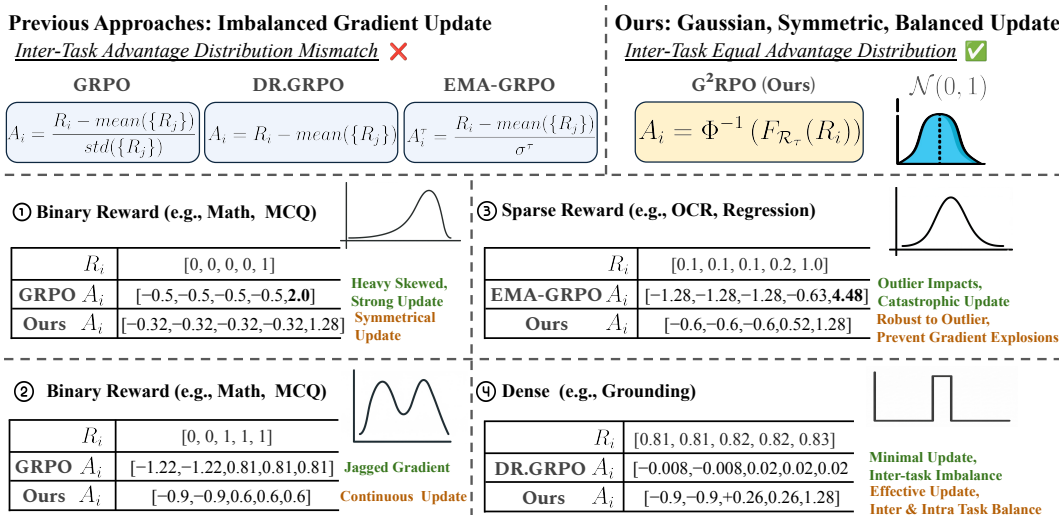

Figure 2: Comparison of advantage formulations against previous methods. By enforcing a Gaussian topology, G^{2}RPO provides 1) intrinsic robustness to outliers, 2) symmetric updates for positive and negative rewards, and 3) uniform variance across diverse tasks.

1 Introduction

Reinforcement Learning (RL) has emerged as a primary driver of recent advancements in Multimodal Large Language Models (MLLMs), significantly enhancing performance across domains ranging from complex visual reasoning to fine-grained object detection (Bai et al., 2025; Comanici et al., 2025; Singh et al., 2026; Liu et al., 2025c; Feng et al., 2025b; Team, 2026; Seed, 2026), encompassing the diverse spectrum of tasks illustrated in Figure 1. However, the vast diversity of visual tasks imposes a significant challenge when optimizing them jointly during the MLLM post-training stage. The extreme variance in reward topologies—ranging from sparse, binary signals in math visual question answering (VQA) to dense, continuous Intersection-over-Union (IoU) scores in grounding tasks—creates significant intra- and inter-task update imbalances. This instability is particularly detrimental to the Group Relative Policy Optimization (GRPO) (Guo et al., 2025) algorithm, rendering it highly susceptible to gradient explosion during large-scale training.

Standard GRPO suffers from intra-task imbalance because its sample-wise standard deviation normalization disproportionately favors low-variance rollouts (Liu et al., 2025b; Bereket & Leskovec, 2025; Chu et al., 2025; Huang et al., 2025). Dr.GRPO (Liu et al., 2025b) removes this normalization but inevitably causes inter-task imbalance, where high variance tasks dominate gradient update and low variance ones are suppressed. While recent methods like EMA-GRPO (Feng et al., 2025b) mitigate this using task-wise moving averages of reward variance, they fundamentally rely on linear transformations. Since linear scaling merely matches the first two statistical moments (mean and variance) while preserving higher-order distributional shapes, it fails to guarantee true inter-task gradient equity and leaves the optimization vulnerable to structural pathologies like heavy-tail outliers.

To overcome the statistical fragility of linear normalization, we propose **Gaussian GRPO (G^{2}RPO)**, which replaces scalar standardization with non-linear distributional matching. By utilizing 1D Optimal Transport—which admits a highly efficient closed-form solution via Cumulative Distribution Functions (CDFs)—G^{2}RPO strictly maps the empirical reward distribution of any given task directly to the standard normal distribution, $\mathcal{N}(0,1)$. As illustrated in Figure 2, enforcing this strict Gaussian topology mathematically caps outliers, smooths bimodal step-functions into symmetrically balanced tails, and promotes advantage-scale equity across tasks.

Another major challenge in MLLM RL post-training is preserving robust capabilities in both fine-grained perception and multi-step reasoning (Liu et al., 2025a; Tian et al., 2025; Tu et al., 2025; Yang et al., 2025b). While recent works (Wang et al., 2025d; Tian et al., 2025; Zhang et al., 2025a) attempt to enhance perception by perturbing visual inputs for auxiliary training objectives or by inserting explicit vision anchors, these approaches suffer from significant drawbacks. They require costly data annotations or incur substantial

computational overhead from additional modules, severely limiting their scalability and leaving their efficacy unverified across diverse, multi-domain multimodal benchmarks.

To address this with a simple and scalable approach, we frame this challenge as a multi-task balancing optimization between vision-centric tasks (e.g., OCR, visual grounding) and reasoning-centric tasks (e.g., math and science VQA). Alongside G^2RPO, we systematically analyze task-level performance, focusing specifically on the dynamics of response length and entropy loss during training. Observing a stark divergence in the optimization trajectories of distinct task types, we introduce task-level *response length and entropy shaping* to encourage stable and accelerated convergence. For response length, we explicitly elicit extended reasoning chains for complex queries while enforcing concise, direct outputs for vision-centric tasks, which effectively solves complex questions, mitigates hallucinations and bolsters visual grounding. Concurrently, our entropy shaping mechanism confines the model to an optimal exploration zone, preventing both entropy collapse (premature over-reliance on high-probability tokens) and entropy explosion (generation of incoherent text).

Integrating these methodologies, we introduce **OpenVLThinkerV2**. We evaluate our model across 18 benchmarks spanning six major task categories: general science knowledge, mathematics, chart and document understanding, spatial reasoning, and visual grounding. Extensive experiments demonstrate that OpenVLThinkerV2 consistently achieves robust performance, establishing new state-of-the-art (SOTA) results among open-source models. For instance, OpenVLThinkerV2 achieves 71.6% on MMMU and 79.5% on MathVista, surpassing GPT-4o by a significant margin. Furthermore, across six distinct benchmarks evaluating document understanding and spatial reasoning, OpenVLThinkerV2 significantly outperforms proprietary frontier models, including GPT-5 and Gemini 2.5 Pro.

Our main contributions are summarized as follows:

- We propose **G^2RPO**, a novel RL training objective that replaces linear scaling with non-linear distributional matching. By mathematically forcing each task's advantage distribution to converge to $\mathcal{N}(0,1)$, G^2RPO promotes advantage-scale equity across tasks and systematically mitigates vulnerabilities to structural pathologies like heavy-tail outliers.
- We introduce **task-level response length and entropy shaping** mechanisms to balance perception and multi-step reasoning. These dynamic bounds encourages early response length convergence and effectively preventing both entropy collapse and explosion.
- We present **OpenVLThinkerV2**, a robust, general-purpose multimodal model. Extensive evaluations across 18 benchmarks demonstrate its superior performance, establishing new SOTA results and consistently outperforming leading proprietary frontier models.

2 Method

2.1 Preliminary

Group Relative Policy Optimization (GRPO). We model an autoregressive language model as a stochastic policy π_θ. For a given query $q \sim \mathcal{D}$, GRPO samples a group of G responses $\mathcal{G} = \{y_1, \ldots, y_G\}$ from the behavior policy $\pi_{\theta_{\text{old}}}$ and computes their scalar rewards $\{R_1, \ldots, R_G\}$. It then maximizes the following token-level objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(r_{i,t}(\theta)\widehat{A}_i,\ \text{clip}\left(r_{i,t}(\theta),1-\varepsilon,1+\varepsilon\right)\widehat{A}_i\right)\right] \tag{1}$$

where $r_{i,t}(\theta)$ is the probability ratio between the new and old policy, and ε is the clipping range. The advantage $\widehat{A}_i$ is defined by standardizing the rewards within the prompt group:

$$\widehat{A}_i^{\text{GRPO}} = \frac{R_i - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}} + \epsilon} \tag{2}$$

where $\mu_{\mathcal{G}}$ and $\sigma_{\mathcal{G}}$ are the group's empirical mean and standard deviation.

Multi-Task Normalization. While GRPO is highly effective for single-task alignment, its localized standard deviation $\sigma_{\mathcal{G}}$ destabilizes optimization in multi-task scenarios where

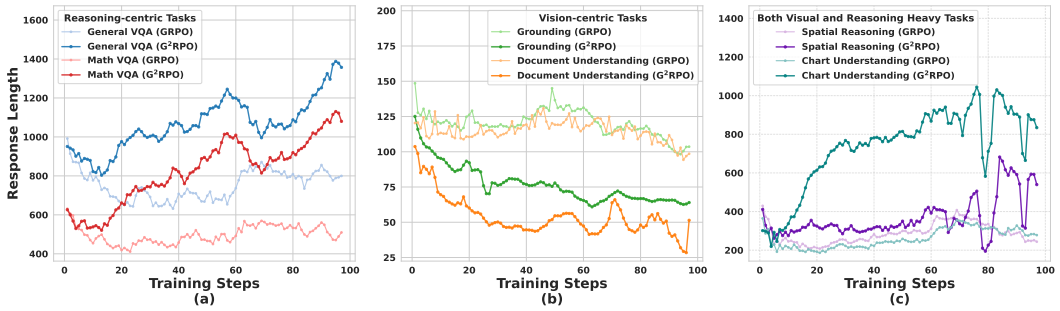

Figure 3: Comparison of response length dynamics during training. G^{2}RPO effectively encourage early convergence. a) It scales up reasoning length for complex question. b) It reduces overthinking for visual-centric tasks, enhancing perceptual grounding and mitigating hallucinations. c) For both reasoning and perception heavy tasks, the generation length stabilizes within an optimal range, effectively balancing both capabilities.

reward scales differ drastically. EMA-GRPO addresses this inter-task imbalance by scaling the advantage using a historically tracked, task-specific standard deviation $\sigma_\tau^{(t)}$:

$$\sigma_\tau^{(t)} = \alpha\sigma_\tau^{(t-1)} + (1-\alpha)\sigma_{\mathcal{G}}, \qquad \widehat{A}_i^{\text{EMA-GRPO}} = \frac{R_i - \mu_{\mathcal{G}}}{\sigma_\tau^{(t)} + \epsilon} \tag{3}$$

However, these approaches do not fundamentally resolve the issue of imbalanced gradient updates across diverse tasks. Because standard normalization relies on a *linear transformation*, it strictly preserves the original shape of the reward distribution. This limitation leaves the optimization process vulnerable to severe topological pathologies. For example, as illustrated in Figure 2's scenarios 1 and 3, it is susceptible to *heavy-tail outliers*, where a single anomalously high reward artificially inflates the exponential moving average (EMA) variance over time, thereby suppressing the learning signals for subsequent normal responses. In addition, EMA-GRPO requires extra hyperparameter tuning and introduces a *momentum lag* where the fixed decay rate α struggles to adapt to sudden policy breakthroughs, inadvertently feeding the model stale advantage signals.

2.2 Gaussian GRPO (G^{2}RPO)

To overcome the statistical fragility of linear moment-matching, we abandon scalar standardization in favor of non-linear distributional matching. Intuitively, an optimal advantage distribution should be intrinsically robust to outliers, symmetric across positive and negative rewards, and maintain uniform variance across highly diverse tasks. These structural requirements naturally motivate the adoption of a Gaussian distribution. Therefore, we propose **Gaussian GRPO (G^{2}RPO)**, which formulates advantage estimation as an Optimal Transport problem. Specifically, our objective is to find a transport map that strictly maps the empirical reward distribution of any task directly to a well-behaved target distribution, namely the standard normal distribution $P_{\mathcal{N}} \equiv \mathcal{N}(0,1)$.

Given a set of empirical rewards $\mathcal{R}_\tau = \{R_1, \ldots, R_N\}$ for a specific task τ, let $P_{\mathcal{R}_\tau}$ denote its empirical distribution, our goal is to find a mapping function Ψ that transports this empirical distribution to $\mathcal{N}(0,1)$. This can be achieved by minimizing the Wasserstein-2 (W_2) distance between $P_{\mathcal{R}_\tau}$ and $P_{\mathcal{N}}$, defined under the squared Euclidean cost function:

$$W_2^2(P_{\mathcal{R}_\tau}, P_{\mathcal{N}}) = \inf_{\gamma \in \Pi(P_{\mathcal{R}_\tau}, P_{\mathcal{N}})} \int_{\mathbb{R}\times\mathbb{R}} |x-y|^2 \, d\gamma(x,y) \tag{7}$$

where $\Pi(P_{\mathcal{R}_\tau}, P_{\mathcal{N}})$ denotes the collection of all joint distributions with marginals $P_{\mathcal{R}_\tau}$, $P_{\mathcal{N}}$.

In 1-dimensional space, this unique optimal transport map admits a highly efficient closed-form solution via the Cumulative Distribution Functions (CDFs). By evaluating this map, G^{2}RPO mathematically neutralizes extreme skewness, bimodal splits, and outliers. Instead of standardizing via mean and variance, G^{2}RPO assigns advantages by mapping the relative rank of the response directly to the inverse CDF of the target normal distribution:

$$\widehat{A}_i^{\text{Ours}} = \Psi(R_i, \mathcal{R}_\tau) = \Phi^{-1}\left(F_{\mathcal{R}_\tau}(R_i)\right) \tag{8}$$

Algorithm 1: G^{2}RPO: Distributional Matching via 1D Optimal Transport

Objective: Map empirical task rewards $\mathcal{R}_\tau = \{R_1, \ldots, R_N\}$ to $\mathcal{N}(0,1)$.

Step 1: Rank Rewards. Compute the uniform probability p_i based on the relative rank of each reward.

$$p_i = \frac{\text{rank}(R_i) - 0.5}{N} \tag{4}$$

```
# Sort empirical rewards and track original indices
sorted_rewards, indices = torch.sort(rewards)
ranks = torch.arange(1, N + 1, device=device, dtype=torch.float32)
probabilities = (ranks - 0.5) / N
```

Step 2: Quantile Mapping. Map p_i to the inverse CDF (quantile function) of $\mathcal{N}(0,1)$, denoted as Φ^{-1}.

$$\Psi(R_i, \mathcal{R}_\tau) = \Phi^{-1}(p_i) = \sqrt{2}\,\text{erfinv}(2p_i - 1) \tag{5}$$

```
# Generate target quantiles from Standard Normal N(0,1)
target_quantiles = math.sqrt(2.0) * torch.erfinv(2.0 * probabilities -
    1.0)
target_quantiles = target_quantiles.to(dtype)
```

Step 3: Tie-Breaking Strategy. To ensure identical behaviors receive identical learning signals, assign the mean of the target quantiles to all identical values, where $\mathcal{K}_{R_i}$ is the set of indices j such that $R_j = R_i$.

$$\widehat{A}_i^{\text{Ours}} = \Psi_{\text{tied}}(R_i, \mathcal{R}_\tau) = \frac{1}{|\mathcal{K}_{R_i}|} \sum_{j \in \mathcal{K}_{R_i}} \Psi(R_j, \mathcal{R}_\tau) \tag{6}$$

```
# Handle Ties: Average the quantiles for identical rewards
unique_rewards, inverse_indices = torch.unique(sorted_rewards,
    return_inverse=True)
if unique_rewards.shape[0] < N:
    tied_quantiles = torch.zeros_like(unique_rewards, dtype=dtype)
    tied_quantiles.scatter_add_(0, inverse_indices, target_quantiles)
    counts = torch.bincount(inverse_indices).to(dtype)
    tied_quantiles = tied_quantiles / counts
    target_quantiles = tied_quantiles[inverse_indices]
# Scatter advantages back to original tensor ordering
advantages = torch.zeros_like(rewards)
advantages[indices] = target_quantiles
```

where Φ^{-1} is the quantile function (inverse CDF) of $\mathcal{N}(0,1)$, and $F_{\mathcal{R}_\tau}$ is the empirical CDF of the task's rewards. In practice, this mapping is computed online and independently within each prompt-level rollout group of size G (i.e., $N = G$); we do not aggregate rewards over the minibatch, a moving window, or the full dataset. Task identity is read directly from each training example's source label to select the corresponding shaping thresholds. Because we adopt mid-rank probabilities $p_i = (\text{rank}(R_i) - 0.5)/G$, the mapped advantages avoid the degenerate endpoints 0 and 1 and remain bounded even for small groups (e.g., the largest magnitude is 1.28 for $G = 5$ and 1.53 for $G = 8$). We provide a step-by-step pytorch-style pseudocode implementation of G^{2}RPO, alongside its underlying mathematical formulations, in Algorithm 1.

Discussion. By enforcing a strict Gaussian topology, G^{2}RPO isolates the policy from anomalous reward spikes by mathematically capping outliers at the highest quantile. As illustrated in Figure 2, for binary rewards, it gracefully converts bimodal step-functions into smooth, symmetrically balanced Gaussian tails. Because every task's advantage distribution is mathematically forced to converge to $\mathcal{N}(0,1)$, G^{2}RPO promotes advantage-scale equity across tasks without requiring rigid momentum hyperparameters. We emphasize that

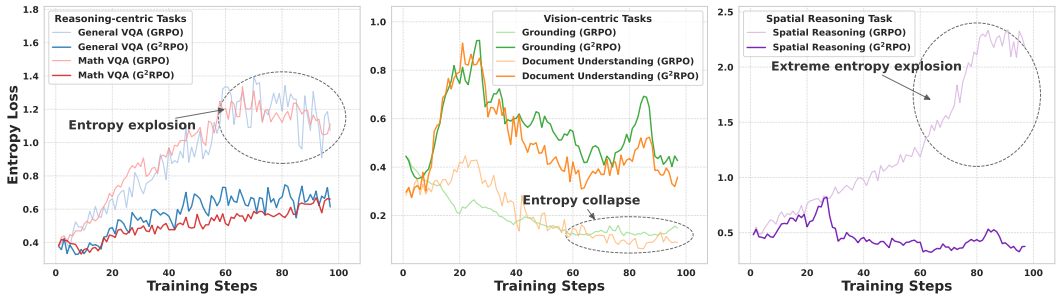

Figure 4: Effect of task-level entropy shaping. G^2RPO effectively prevents entropy explosion for reasoning-centric tasks and OOD task (spatial reasoning) while concurrently mitigating entropy collapse in vision-centric tasks.

aligning the marginal advantage distributions does not by itself equalize the full per-task gradient magnitudes, which additionally depend on factors such as token-length distributions, sampling frequency, and clipping; rather, G^2RPO removes one major source of imbalance—reward-topology-induced advantage-scale mismatch.

2.3 Task-level Length and Entropy Shaping

A primary challenge in advancing RL post-training for MLLMs lies in preserving robust capabilities across both fine-grained perception and multi-step reasoning (Liu et al., 2025a; Tian et al., 2025; Tu et al., 2025; Yang et al., 2025b). We frame this as a multi-task balancing optimization, distinguishing between vision-centric tasks (e.g., OCR, visual grounding), reasoning-centric tasks (e.g., math and science VQA), and hybrid tasks requiring both modalities (e.g., chart reasoning). To this end, we first analyze the task-level performance dynamics of these distinct categories, with a specific focus on the evolution of response length and entropy loss during training.

Task-Level Response Length Shaping. As illustrated in Figure 3, we observe distinct trajectories in response length dynamics during GRPO training. For reasoning-centric tasks, the model initially decreases response length as it adapts to the training data distribution, before ultimately converging to longer, more elaborate reasoning chains. Conversely, for vision-centric tasks, the model consistently reduces response length, suggesting an optimization toward concise outputs that directly enhance perceptual grounding without generating unnecessary, potentially hallucinated reasoning steps.

Motivated by these divergent trends, we explicitly shape the expected response length on a per-task basis to accelerate convergence and jointly enhance both perception and reasoning capabilities. We propose a task-level length shaping mechanism that applies a customized trapezoidal reward envelope, rewarding the model when its generation length falls within an optimal target range while softly penalizing excessively short or long responses.

For a given task, let $|y|$ denote the token length of the model's response y. We define four task-specific threshold hyperparameters: the absolute minimum valid length L_{min}, the start of the optimal length plateau L_{low}, the end of the optimal plateau L_{high}, and the absolute maximum valid length L_{max}. The length reward $R_{\text{length}}(y)$ is mathematically formulated as:

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| < L_{\text{min}} \text{ or } |y| > L_{\text{max}} \\ \frac{|y| - L_{\text{min}}}{L_{\text{low}} - L_{\text{min}}}, & L_{\text{min}} \leq |y| < L_{\text{low}} \\ 1, & L_{\text{low}} \leq |y| \leq L_{\text{high}} \\ \frac{L_{\text{max}} - |y|}{L_{\text{max}} - L_{\text{high}}}, & L_{\text{high}} < |y| \leq L_{\text{max}} \end{cases} \tag{9}$$

Task-Level Entropy Shaping. In a multi-task reinforcement learning setting, varying task difficulties inevitably induce divergent exploration patterns. As illustrated in Figure 4, we observe that for reasoning-centric tasks, the model tends to artificially inflate entropy, leading to *entropy explosion* where it explores incoherent tokens. Conversely, vision-centric tasks are highly susceptible to *entropy collapse*, wherein the model over-exploits high-probability tokens and prematurely abandons necessary exploration. Furthermore, the phenomenon of

Model	General VQA			Math VQA			Chart VQA		
	MMMU	MMBench	MMStar	MathVista	MathVerse	MathVision	AI2D	ChartQA	CharXiv(RQ)
GPT-4o (Hurst et al., 2024)	70.7	84.3	65.1	63.8	41.2	30.4	84.9	86.7	47.1
Gemini 2.5 Pro (Comanici et al., 2025)	81.7	90.1	77.5	82.7	82.9	73.3	88.4	83.3	67.9
MM-Eureka-7B (Meng et al., 2025a)	57.3	87.7	64.4	73.0	50.3	26.9	84.1	77.3	39.5
OpenVLThinker-7B (Deng et al., 2025)	55.8	87.8	59.1	70.2	47.9	25.3	81.8	85.7	39.3
VL-Rethinker-7B (Wang et al., 2025b)	56.7	87.9	62.7	74.9	54.2	32.3	80.8	84.7	39.8
Vision-G1 (Zha et al., 2025)	53.4	88.0	63.1	76.1	50.0	31.3	82.1	85.6	41.0
ARES-7B (Chen et al., 2025c)	67.9	-	65.3	74.6	56.5	51.9	-	-	47.0
OVR-7B (Wei et al., 2025)	-	86.6	62.7	72.1	54.6	51.8	-	-	44.5
VisionZero (Wang et al., 2025c)	58.8	-	65.8	73.1	52.1	28.5	84.8	86.3	-
OneThinker-8B (Feng et al., 2025b)	70.6	86.6	70.6	77.6	64.3	48.3	85.2	76.4	44.0
Qwen3-VL-Instruct-8B	60.2	85.1	68.5	74.2	58.1	45.4	82.3	82.8	44.5
Qwen3-VL GRPO	69.7	86.0	71.7	78.1	64.7	52.6	85.1	82.4	50.5
Qwen3-VL GDPO	69.8	86.1	71.5	78.0	64.9	49.7	85.1	82.4	51.6
OpenVLThinkerV2 (Ours)	**71.6**	**88.2**	**73.8**	**79.5**	**65.8**	**53.4**	**87.5**	**87.4**	**53.0**

Table 1: Performance comparison of various models across visual reasoning benchmarks. We report our reproduced results for Qwen3-VL-Instruct-8B under the same setting. Our model consistently outperforms strong open-source baselines.

entropy explosion is drastically exacerbated in complex or out-of-distribution (OOD) tasks, indicating a detrimental inclination to sample low-probability regions of the action space.

To mitigate these training instabilities, we propose a task-level entropy shaping mechanism. Analogous to our response length shaping, we impose a strict regularization envelope to bound the model's exploration within an optimal, task-specific zone. Let H_{task} denote the average entropy loss for a given task, computed over the generated negative log-probabilities. We define a target exploration interval bounded by a minimum entropy threshold, H_{min}, to prevent collapse, and a maximum entropy threshold, H_{max}, to prevent explosion. We formulate the entropy regularization loss, $\mathcal{L}_{\text{ent_reg}}$, as a margin-based penalty that applies a linear correction only when the entropy falls outside this optimal envelope:

$$\mathcal{L}_{\text{ent_reg}} = \max(0, H_{\text{task}} - H_{\text{max}}) + \max(0, H_{\text{min}} - H_{\text{task}}) \tag{10}$$

This regularization term is then added to the final optimization objective, scaled by a weighting hyperparameter λ_{ent}, ensuring the model maintains a stable balance between exploration and exploitation across highly diverse task topologies.

Discussion. While our response length and entropy shaping mechanisms introduce specific hyperparameters (e.g., L_{min}, H_{min}), we argue that these boundaries can be intuitively selected to guide the model's behavioral trajectory—such as encouraging longer reasoning chains for complex tasks—without the need for delicate tuning. Although our current implementation relies on simple empirical observations to establish a general optimization trend, it already yields substantial performance improvements. We leave the systematic exploration and automated search of these hyperparameters to future work. We emphasize that even coarse, trend-based shaping is sufficient to achieve robust and significant gains.

3 Experiments

3.1 Setup

Training Details. We train our model on AWS Trainium instances (specifically, Trn1.32xlarge). Our RL optimization is initialized from the Qwen3-VL-Instruct-8B (Bai et al., 2025) and trained using a filtered subset of the OneThinker-600k dataset (Feng et al., 2025b). Our sole filtering criterion is the removal of video and tracking data, which require substantially longer sequence lengths and impose a heavy computational burden; this yields a final training set of 162,146 samples. We introduce no additional data curation. We optimize for a single epoch using AdamW, with a batch size of 128 and a learning rate of 2×10^{-6}. The maximum generation length is capped at 4096 tokens. Following the practices outlined in Yu et al. (2025a), we disable KL regularization and apply dynamic data filtering, actively discarding rollouts that are uniformly correct or incorrect to maintain high-quality gradient signals. The entire training process requires approximately three days.

Benchmarks. To comprehensively evaluate our model, we adopt a variety of benchmarks across a wide-range of tasks. For visual reasoning, we evaluate on general multimodal VQA (Yue et al., 2024; Liu et al., 2024a; Chen et al., 2024) and math-focused VQA (Lu et al., 2024; Zhang et al., 2024; Wang et al., 2024a). For vision-centric perception, we assess document understanding (Mathew et al., 2021; 2022; Liu et al., 2024b) and visual

	(a) Document Understanding		
Model	**DocVQA**	**OCRBench**	**InfoVQA**
GPT-5 (Singh et al., 2026)	91.5	810	79.0
Gemini 2.5 Pro (Comanici et al., 2025)	92.6	866	84.2
DocThinker-7B (Yu et al., 2025b)	78.8	-	52.3
VisionThink (Yang et al., 2025a)	94.4	808	-
DeepEyesV2 (Hong et al., 2025)	-	882	-
RegionDoc-R1 (Wang et al., 2025a)	95.3	-	83.2
VisionZero (Wang et al., 2025c)	95.2	885	82.3
OneThinker-8B (Feng et al., 2025b)	95.0	833	84.8
Qwen3-VL-Instruct	95.3	819	83.1
Qwen3-VL GRPO	95.9	875	84.9
Qwen3-VL GDPO	95.6	897	83.5
OpenVLThinkerV2 (Ours)	**96.7**	**911**	**86.4**

	(b) Spatial Reasoning		
Model	**EmbSpatial**	**RefSpatial**	**RoboSpatial**
GPT-5 (Singh et al., 2026)	82.9	23.8	53.5
Gemini 2.5 Pro (Comanici et al., 2025)	79.1	36.5	47.5
SpatialRGPT-8B (Cheng et al., 2024)	59.6	-	66.7
SpaceLLaVA-13B (Foutter et al., 2024)	49.4	3.25	61.0
RoboBrain2.0-7B (Team et al., 2025)	-	32.5	59.6
RoboRefer-8B-SFT (Zhou et al., 2025a)	-	48.4	58.3
G^2VLM (Hu et al., 2025)	62.4	43.5	62.7
OneThinker-8B (Feng et al., 2025b)	79.9	38.9	61.7
Qwen3-VL-Instruct	78.2	43.9	60.6
Qwen3-VL GRPO	79.9	43.3	61.7
Qwen3-VL GDPO	79.9	43.0	61.7
OpenVLThinkerV2 (Ours)	**83.1**	**44.6**	**63.2**

Table 2: Results on (a) document understanding and (b) spatial reasoning benchmarks. We compare against both general multimodal models and task-specific expert baselines.

grounding (Kazemzadeh et al., 2014; Yu et al., 2016). For tasks that demand both rigorous visual perception and multi-step reasoning, we test on chart understanding (Kembhavi et al., 2016; Masry et al., 2022; Wang et al., 2024b) and spatial reasoning (Du et al., 2024; Zhou et al., 2025a; Song et al., 2025). We report our reproduced results for Qwen3-VL-Instruct. All models are evaluated following prior Qwen3-VL's default generation hyperparameters.

3.2 Main Results

Visual Reasoning Tasks. As illustrated in Table 1, we evaluate OpenVLThinkerV2 across a diverse suite of tasks encompassing general scientific knowledge, mathematics, chart understanding, and complex multimodal reasoning. Compared to strong open-source baselines, including the recent generalist model OneThinker-8B, OpenVLThinkerV2 achieves superior performance across all evaluated benchmarks. Under controlled experimental conditions (utilizing identical training data and compute resources), our model consistently outperforms both standard GRPO and GDPO (Liu et al., 2026) variants, underscoring the efficacy of G^2RPO. Notably, OpenVLThinkerV2 reaches 71.6% on MMMU, 88.2% on MMBench, and 73.8% on MMStar, surpassing GPT-4o, as well as 87.4% on ChartQA, which exceeds the performance of Gemini 2.5 Pro.

Document Understanding Tasks. We evaluate the vision-centric capabilities of OpenVLThinkerV2, particularly document understanding, as detailed in Table 2(a). Our model achieves state-of-the-art performance, outperforming both strong open-source and leading proprietary models across all three benchmarks. Notably, OpenVLThinkerV2 attains a score of 911 on OCRBench, surpassing DeepEyesV2, which is a specialized model that employs dynamic zoom-in tools for enhanced document parsing, and significantly outperforms frontier proprietary models such as GPT-5 and Gemini 2.5 Pro.

Spatial Reasoning Tasks. Spatial reasoning rigorously tests a model's capacity for both granular visual perception and complex multimodal logic. We benchmark OpenVLThinkerV2 against specialized expert models fine-tuned explicitly on spatial reasoning datasets, such as RoboRefer. Despite not finetuned on this data, our model achieves the highest performance on EMbSpatial, and performs on par with the spatial-expert SpatialRGPT on the RoboSpatial, while underperformed the finetuned expert model RoboRefer-SFT marginally. Furthermore, our model significantly surpass the capabilities of GPT-5 and Gemini 2.5 Pro.

Grounding Tasks. We further evaluate our model on vision-centric tasks that require formatted, high-precision numerical outputs, such as visual grounding in Table 3. OpenVLThinkerV2 demonstrates SOTA localization capabilities across the widely used RefCOCO, RefCOCO+, and RefCOCOg benchmarks, achieving 93.4%/88.2%/90.4%, respectively. Our model consistently outperforms previous baselines by a substantial margin, including the specialized visual grounding expert, Grounding DINO.

Model	**RefCOCO**	**RefCOCO+**	**RefCOCOg**
Gemini 1.5 Pro (Gemini Team, 2024)	73.2	62.5	75.2
Grounding DINO (Liu et al., 2023)	90.6	88.2	86.1
VLM-R1 (Shen et al., 2025)	90.5	84.3	87.1
DeepEyes (Zheng et al., 2025b)	89.8	83.6	86.7
OneThinker-8B (Feng et al., 2025b)	92.0	87.0	89.2
Qwen3-VL-Instruct	89.9	84.5	86.8
Qwen3-VL GRPO	92.1	87.7	89.6
Qwen3-VL GDPO	92.2	87.8	88.9
OpenVLThinkerV2 (Ours)	**93.4**	**88.2**	**90.4**

Table 3: Evaluation results for **Grounding** task. We report scores on val splits. Our model consistently outperforms previous baselines.

Model	General VQA	Math VQA	Chart VQA	Grounding	Document Understanding	Spatial Reasoning
Qwen3-VL-Instruct-8B	71.3	59.2	69.9	87.1	86.8	60.9
Qwen3-VL + G^2RPO	76.9	64.8	74.5	90.2	90.6	62.3
+ task-level entropy loss	77.0	65.1	75.3	90.4	90.8	62.8
+ task-level length reward	77.4	65.7	75.4	90.5	91.1	63.2
OpenVLThinkerV2 (Ours)	**77.9**	**66.2**	**76.0**	**90.7**	**91.4**	**63.6**

Table 4: Ablation study evaluating the impact of different training components across the six main domains. Scores represent the average performance within each task category.

Overall Performance. In summary, OpenVLThinkerV2 exhibits balanced and robust performance across a broad spectrum of both vision-centric and reasoning-centric multimodal tasks. These extensive evaluations empirically validate the superiority of G^2RPO for multi-task RL optimization. Furthermore, they confirm that our task-level response length and entropy shaping mechanisms effectively stabilize training, accelerate convergence, and yield a highly capable multimodal generalist. We provide more details of how our model evolves its RL rewards during training, please refer to Appendix A.

3.3 Ablation Study

We conduct an ablation study to isolate the contributions of each component in OpenVLThinkerV2, as shown in Table 4. Compared to the Qwen3-VL-8B baseline, integrating the G^2RPO objective yields the most substantial initial performance improvement. Adding task-level entropy shaping further boosts performance, particularly in reasoning-centric tasks, with more modest gains observed in saturated or out-of-distribution domains (e.g., visual grounding and spatial reasoning). Alternatively, applying the task-level length reward provides even broader improvements, outperforming entropy shaping alone and highlighting its effectiveness as a regularization directive. Ultimately, combining both shaping mechanisms with G^2RPO achieves the highest overall gains. This synergistic effect demonstrates that response length and entropy shaping are highly complementary, empirically validating the collective efficacy of our proposed methodology.

4 Related Work

Group Relative Policy Optimization. Group Relative Policy Optimization (GRPO) (Guo et al., 2025), first introduced by DeepSeek-R1, has become the de-facto Reinforcement Learning (RL) objective for enhancing the reasoning capabilities of Large Language Models (LLMs) and Multimodal LLMs (MLLMs) (Zhang et al., 2025b; Dong et al., 2025; Yu et al., 2025a; Xie et al., 2025; Chen et al., 2025a; Feng et al., 2025c; Liu et al., 2025b; Zheng et al., 2025a; Gao et al., 2025). The success of GRPO has motivated its extension to MLLM post-training. However, the vast diversity of visual tasks exhibits significantly higher variance in reward topologies compared to standard LLM tasks, creating severe intra- and inter-task update imbalances. While recent works like EMA-GRPO (Feng et al., 2025b) attempt to address this using task-wise moving averages of reward variance, their reliance on linear scaling fails to guarantee inter-task gradient equity and leaves the optimization vulnerable to structural pathologies. In contrast, G^2RPO resolves this by enforcing a strict Gaussian topology. This non-linear mapping mathematically caps outliers and smooths bimodal rewards, promoting advantage-scale equity across tasks.

Multimodal Reasoning. A growing body of literature has integrated RL into MLLMs to enable complex reasoning across diverse visual tasks (Li et al., 2025b; Sun et al., 2025a; Feng et al., 2025a; Sun et al., 2025b; Zhou et al., 2025b; Duan et al., 2025; Chen et al., 2025b;c; Meng et al., 2025b). A primary challenge identified in these studies is preserving robust capabilities in both fine-grained perception and multi-step reasoning (Liu et al., 2025a; Tian et al., 2025; Tu et al., 2025; Yang et al., 2025b). To enhance perception, recent methods optimize auxiliary KL divergence objectives using corrupted visual inputs (Wang et al., 2025d; Zhang et al., 2025a), or insert explicit visual anchors or claims into the reasoning process via proprietary models and external captioners (Tian et al., 2025; Yang et al., 2025b; Tu et al., 2025). However, these approaches require costly data annotations or incur substantial computational overhead, severely limiting their scalability across diverse benchmarks. We bypass these burdens by framing this challenge directly as a multi-task optimization

problem. Through task-level response length and entropy shaping, our method accelerates stable convergence and effectively balances perception and reasoning capabilities.

Optimal Transport in LLM. Optimal Transport (OT) is not entirely new in LLM, they are adopted in specific areas like preference alignment (Melnyk et al., 2024; Li et al., 2025a; Na et al., 2026; Nanfack et al., 2026). Existing approaches typically leverage OT to compute semantic distances between token distributions, enforce stochastic dominance between in reward distributions, or dynamically map latent safety representations, which fundamentally treating it as a distance metric. In contrast, G^2RPO fundamentally repurposes 1D OT as a universal advantage normalization mechanism, directly addressing the extreme inter-task reward variance and heavy-tail topologies inherent in multi-domain multimodal RL.

5 Conclusion

We present OpenVLThinkerV2, a robust, general-purpose multimodal model with multi-task reinforcement learning post-training. To address the extreme variance in reward topologies across diverse visual tasks, we propose G^2RPO to map each task's advantage distribution to converge to $\mathcal{N}(0, 1)$, promoting advantage-scale equity across tasks. Furthermore, we introduce task-level response length and entropy shaping mechanisms, effectively balancing fine-grained perception and complex multi-step reasoning capabilities. Crucially, our method extends well beyond multimodal tasks. Since G^2RPO is inherently designed to harmonize highly divergent reward topologies, it is naturally suited for broader LLM applications that suffer from similar reward heterogeneity, such as SWE coding and GUI tasks. We leave the large-scale expansion of this framework to future work, and we encourage the community to build upon these principles to foster more stable, scalable, and equitable RL optimization paradigms across all language and multimodal foundation models.

Acknowledgments

We thank anonymous reviewers for their helpful comments. This work was partially supported by U.S. ONR grant N00014-23-1-2780, DARPA ANSR program FA8750-23-2-0004, Amazon, and Apple. We thank Amazon for their Trainium compute resources. Chang and Peng were supported in part by a grant from DARPA to the Simons Institute for the Theory of Computing.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *ArXiv preprint*, abs/2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. *ArXiv preprint*, abs/2508.11800, 2025. URL https://arxiv.org/abs/2508.11800.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, Decem-*

ber 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *ArXiv preprint*, abs/2503.09567, 2025a. URL https://arxiv.org/abs/2503.09567.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *ArXiv preprint*, abs/2506.04207, 2025b. URL https://arxiv.org/abs/2506.04207.

Shuang Chen, Yue Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Haoxi Li, Manyuan Zhang, Jiayu Chen, Song Guo, and Nanyun Peng. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. *ArXiv preprint*, abs/2510.08457, 2025c. URL https://arxiv.org/abs/2510.08457.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *ArXiv preprint*, abs/2504.02546, 2025. URL https://arxiv.org/abs/2504.02546.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261, 2025. URL https://arxiv.org/abs/2507.06261.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *ArXiv preprint*, abs/2503.17352, 2025. URL https://arxiv.org/abs/2503.17352.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *ArXiv preprint*, abs/2507.19849, 2025. URL https://arxiv.org/abs/2507.19849.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.33. URL https://aclanthology.org/2024.acl-short.33/.

Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *ArXiv preprint*, abs/2510.11718, 2025. URL https://arxiv.org/abs/2510.11718.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *ArXiv preprint*, abs/2503.21776, 2025a. URL https://arxiv.org/abs/2503.21776.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. Onethinker: All-in-one reasoning model for image and video, 2025b. URL https://arxiv.org/abs/2512.03043.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *ArXiv preprint*, abs/2505.10978, 2025c. URL https://arxiv.org/abs/2505.10978.

Matthew Foutter, Daniele Gammelli, Justin Kruger, Ethan Foss, Praneet Bhoj, Tommaso Guffanti, Simone D'Amico, and Marco Pavone. Space-llava: a vision-language model adapted to extraterrestrial applications, 2024. URL https://arxiv.org/abs/2408.05924.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL https://arxiv.org/abs/2511.20347.

Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530, 2024. URL https://arxiv.org/abs/2403.05530.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv preprint*, abs/2501.12948, 2025. URL https://arxiv.org/abs/2501.12948.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *ArXiv preprint*, abs/2511.05271, 2025. URL https://arxiv.org/abs/2511.05271.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *ArXiv preprint*, abs/2511.21688, 2025. URL https://arxiv.org/abs/2511.21688.

Wenke Huang, Quan Zhang, Yiyang Fang, Jian Liang, Xuankun Rong, Huanjin Yao, Guancheng Wan, Ke Liang, Wenwen He, Mingjun Li, et al. Mapo: Mixed advantage policy optimization. *ArXiv preprint*, abs/2509.18849, 2025. URL https://arxiv.org/abs/2509.18849.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *ArXiv preprint*, abs/2410.21276, 2024. URL https://arxiv.org/abs/2410.21276.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, and Anxiang Zeng. Optimal transport-based token weighting scheme for enhanced preference optimization, 2025a. URL https://arxiv.org/abs/2505.18720.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *ArXiv preprint*, abs/2505.15804, 2025b. URL https://arxiv.org/abs/2505.15804.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv preprint*, abs/2303.05499, 2023. URL https://arxiv.org/abs/2303.05499.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms, 2025a. URL https://arxiv.org/abs/2510.17771.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *ArXiv preprint*, abs/2503.20783, 2025b. URL https://arxiv.org/abs/2503.20783.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *ArXiv preprint*, abs/2503.01785, 2025c. URL https://arxiv.org/abs/2503.01785.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2199–2208, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2821–2831, 2022.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. Distributional preference alignment of llms via optimal transport, 2024. URL https://arxiv.org/abs/2406.05882.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *ArXiv preprint*, abs/2503.07365, 2025a. URL https://arxiv.org/abs/2503.07365.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *ArXiv preprint*, abs/2510.20579, 2025b. URL https://arxiv.org/abs/2510.20579.

Byeonghu Na, Hyungho Na, Yeongmin Kim, Suhyeon Jo, HeeSun Bae, Mina Kang, and Il-Chul Moon. Semantic-aware wasserstein policy regularization for large language model alignment, 2026. URL https://arxiv.org/abs/2602.01685.

Geraldin Nanfack, Eugene Belilovsky, and Elvis Dohmatob. Efficient refusal ablation in llm through optimal transport, 2026. URL https://arxiv.org/abs/2603.04355.

Bytedance Seed. Seed1.8 model card: Towards generalized real-world agency, 2026. URL https://arxiv.org/abs/2603.20633.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *ArXiv preprint*, abs/2504.07615, 2025. URL https://arxiv.org/abs/2504.07615.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Oral Presentation.

Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *ArXiv preprint*, abs/2505.18536, 2025a. URL https://arxiv.org/abs/2505.18536.

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, et al. Spacevista: All-scale visual spatial reasoning from mm to km. *ArXiv preprint*, abs/2510.09606, 2025b. URL https://arxiv.org/abs/2510.09606.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, Yingbo Tang, Xiangqi Xu, Wei Guo, Yaoxu Lyu, Yijie Xu, Jiayu Shi, Mengfei Du, Cheng Chi, Mengdi Zhao, Xiaoshuai Hao, Junkai Zhao, Xiaojie Zhang, Shanyu Rong, Huaihai Lyu, Zhengliang Cai, Yankai Fu, Ning Chen, Bolun Zhang, Lingfeng Zhang, Shuyi Zhang, Dong Liu, Xi Feng, Songjing Wang, Xiaodan Liu, Yance Jiao, Mengsi Lyu, Zhuo Chen, Chenrui He, Yulong Ao, Xue Sun, Zheqi He, Jingshu Zheng, Xi Yang, Donghai Shi, Kunchang Xie, Bochao Zhang, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. Robobrain 2.0 technical report, 2025. URL https://arxiv.org/abs/2507.02029.

Kimi Team. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *ArXiv preprint*, abs/2509.25848, 2025. URL https://arxiv.org/abs/2509.25848.

Songjun Tu, Qichao Zhang, Jingbo Sun, Yuqian Fu, Linjing Li, Xiangyuan Lan, Dongmei Jiang, Yaowei Wang, and Dongbin Zhao. Perception-consistency multimodal large language models reasoning via caption-regularized policy optimization, 2025. URL https://arxiv.org/abs/2509.21854.

Chao Wang, Hehe Fan, Huichen Yang, Zhengdong Hu, Sarvnaz Karimi, Lina Yao, and Yi Yang. Regiondoc-r1: Reinforcing semantic layout-aware learning for document understanding, 2025a. URL https://openreview.net/forum?id=pfHm4YJTzC.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv preprint*, abs/2504.08837, 2025b. URL https://arxiv.org/abs/2504.08837.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play, 2025c. URL https://arxiv.org/abs/2509.25541.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning, 2025d. URL https://arxiv.org/abs/2507.06448.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/cdf6f8e9fd9aeaf79b6024caec24f15b-Abstract-Datasets_and_Benchmarks_Track.html.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, Chunrui Han, Yuang Peng, Qi Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, and Vishal M. Patel. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning, 2025. URL https://arxiv.org/abs/2507.05255.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *ArXiv preprint*, abs/2502.14768, 2025. URL https://arxiv.org/abs/2502.14768.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Visionthink: Smart and efficient vision language model via reinforcement learning. *ArXiv preprint*, abs/2507.13348, 2025a. URL https://arxiv.org/abs/2507.13348.

Siqi Yang, Zilve Gao, Haibo Qiu, Fanfan Liu, Peng Shi, Zhixiong Zeng, Qingmin Liao, and Lin Ma. Learning when to look: A disentangled curriculum for strategic perception in multimodal reasoning, 2025b. URL https://arxiv.org/abs/2512.17227.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv preprint*, abs/2503.14476, 2025a. URL https://arxiv.org/abs/2503.14476.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding, 2025b. URL https://arxiv.org/abs/2508.08589.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Vision-g1: Towards general vision language reasoning with multi-domain data curation, 2025. URL https://arxiv.org/abs/2508.12680.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Shuoshuo Zhang, Yizhen Zhang, Jingjing Fu, Lei Song, Jiang Bian, Yujiu Yang, and Rui Wang. See less, see right: Bi-directional perceptual shaping for multimodal reasoning, 2025a. URL https://arxiv.org/abs/2512.22120.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *ArXiv preprint*, abs/2506.03106, 2025b. URL https://arxiv.org/abs/2506.03106.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *ArXiv preprint*, abs/2507.18071, 2025a. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *ArXiv preprint*, abs/2505.14362, 2025b. URL https://arxiv.org/abs/2505.14362.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *ArXiv preprint*, abs/2506.04308, 2025a. URL https://arxiv.org/abs/2506.04308.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *ArXiv preprint*, abs/2504.21277, 2025b. URL https://arxiv.org/abs/2504.21277.

A More Results

We demonstrate G^2RPO's training stability and efficiency in accuracy reward, length reward, format reward and structure reward.

Accuracy Reward. As showcased in Figure 5, comparing to GRPO and GDPO baselines, G^2RPO demonstrates an early convergence at the first of the 100 training steps. While GRPO is oscillating between 0.685 and 0.695 and GDPO oscillate to lower accuracy reward around step 240, our method (G^2RPO) is consistently learning and improving its accuracy reward.

Length Reward. As illustrated in Figure 6, G^2RPO consistently demonstrate significantly higher length reward than the two other baselines.

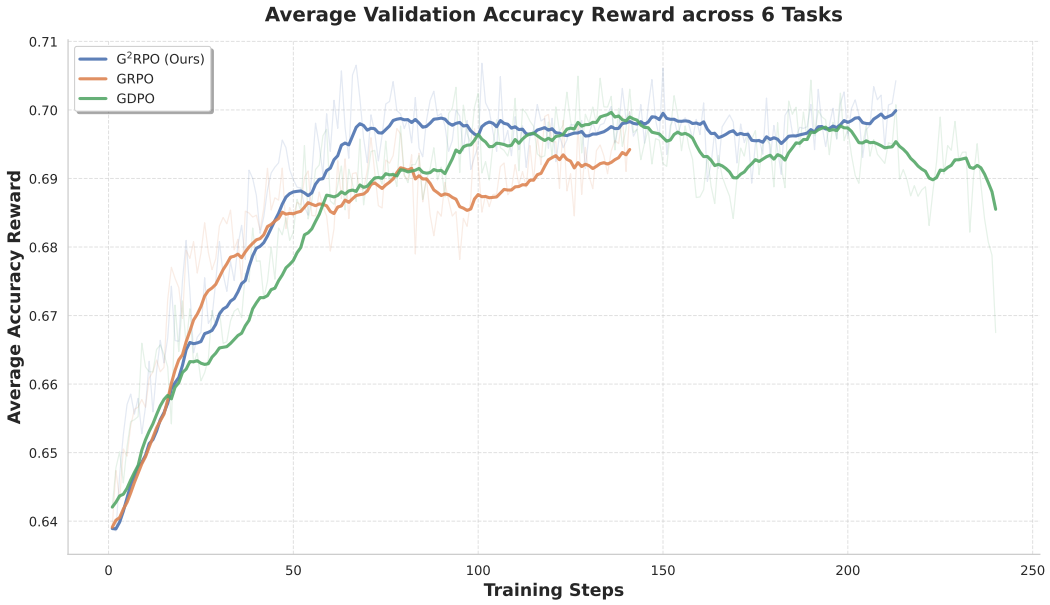

Figure 5: Average Accuracy Reward Comparison across all tasks on the Validation set during Training. G^2RPO demonstrates stable and superior performance overall.

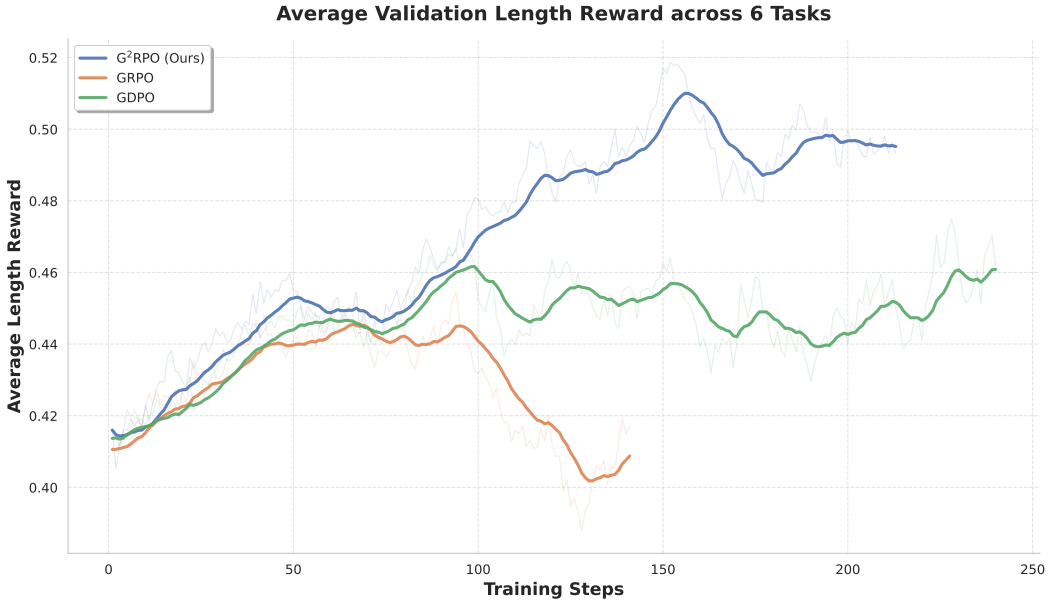

Figure 6: Average Length Reward Comparison across all tasks on the Validation set during Training. G^2RPO demonstrates stable and superior performance overall.

Format Reward. Following Feng et al. (2025b), we adopt format reward which enclose the thinking process with <thinking> and </thinking> tokens and enclose the final answer with the <answer> and </answer> tokens. While GDPO has a theoretical advantage maintaining a better format reward, it demonstrates highest format reward at the beginning but then converges to lower result, as illustrated in Figure 7. G^2RPO maintains the best format reward at the end of the training.

Structure Reward. Following Feng et al. (2025b), we also adopt structure reward, specifically for tasks that require a strict output structure. For example the grounding task output adopt the format of <answer>"boxes": [a, b, c, d]</answer>. While GDPO has a theoret-

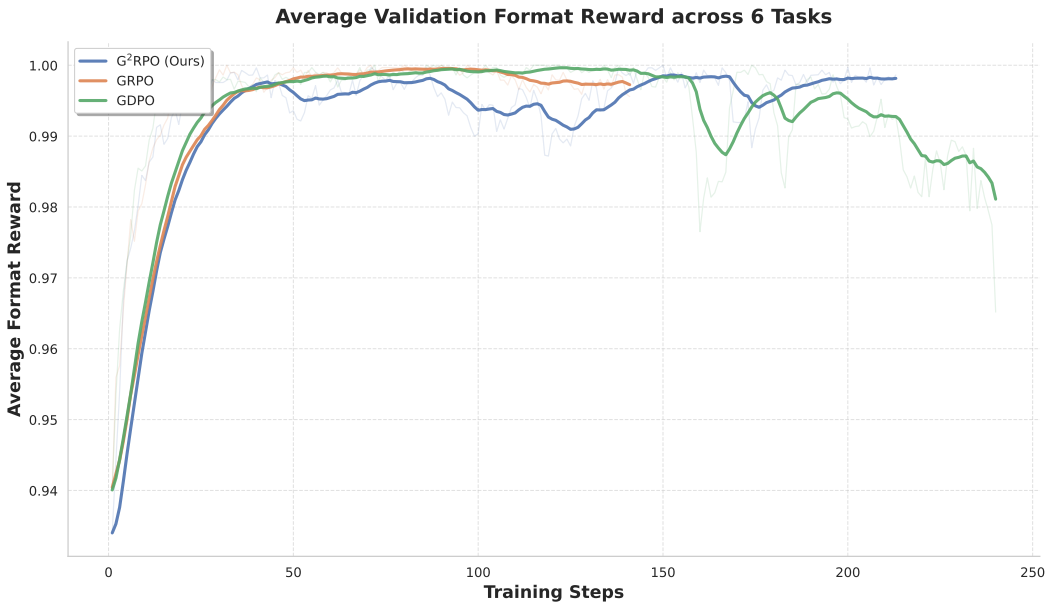

Figure 7: Average Format Reward Comparison across all tasks on the Validation set during Training. G^2RPO demonstrates stable and superior performance overall.

ical advantage maintaining a better structure reward, similar to format reward result, it demonstrates highest structure reward at the beginning but then converges to lower result. G^2RPO maintains the best structure reward at the end of the training.

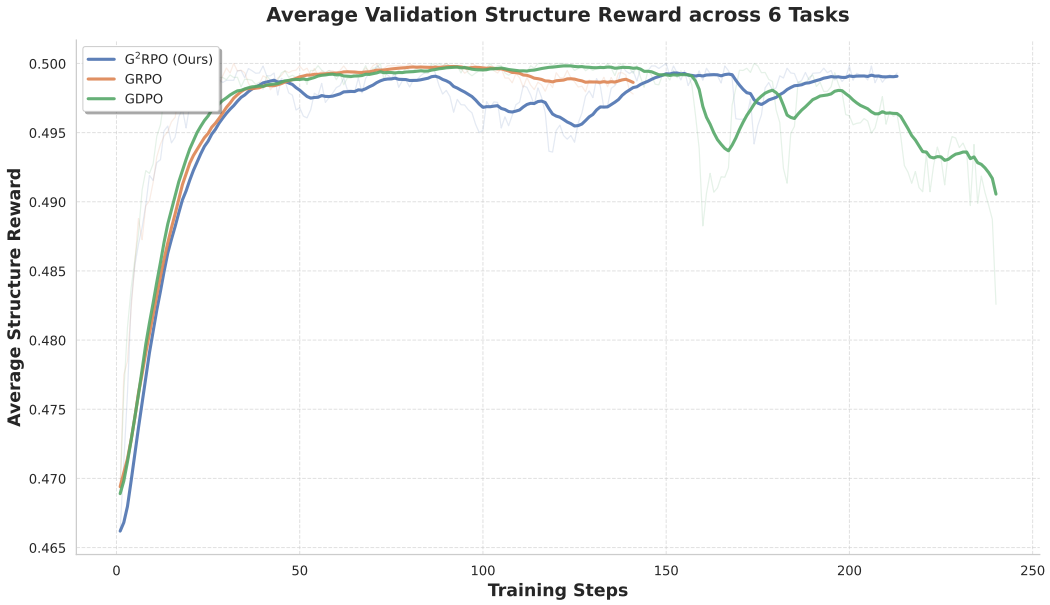

Figure 8: Average Structure Reward Comparison across all tasks on the Validation set during Training. G^2RPO demonstrates stable and superior performance overall.

B Task-Level Shaping Hyperparameters

Table 5 reports the complete set of task-level response length and entropy shaping thresholds used in all our experiments. Consistent with our claim in the main text, we do not perform

any hyperparameter search over these values; we only shape the coarse behavioral trend of each task family. Concretely, General VQA and Math Reasoning share identical response-length thresholds because we treat both as reasoning-heavy tasks, while Grounding and Document Understanding share thresholds as the most vision-centric tasks. For entropy shaping, all tasks are shaped toward the same target exploration zone $[H_{\text{min}}, H_{\text{max}}] = [0.2, 0.7]$ with a fixed regularization weight $\lambda_{\text{ent}} = 0.02$, so that no task suffers from entropy explosion or collapse.

Task	L_{min}	L_{low}	L_{high}	L_{max}	H_{min}	H_{max}	λ_{ent}
General VQA	200	2500	10000	16000	0.2	0.7	0.02
Math Reasoning	200	2500	10000	16000	0.2	0.7	0.02
Grounding	0	30	200	300	0.2	0.7	0.02
Document Understanding	0	30	200	300	0.2	0.7	0.02
Chart VQA	100	600	1500	3000	0.2	0.7	0.02
Spatial Reasoning	100	500	1000	1400	0.2	0.7	0.02

Table 5: Complete task-level response length (L) and entropy (H) shaping thresholds used in our experiments. We shape only the coarse per-family trend and perform no hyperparameter search.

C Controlled Comparison with EMA-GRPO

To further isolate the contribution of G^{2}RPO, we add a Qwen3-VL + EMA-GRPO baseline trained under the exact same single-node data and compute budget as our model. Table 6 reports the average performance across the three benchmarks within each task category, following the protocol of Table 4. Note that OneThinker-8B is itself trained with EMA-GRPO, but under a substantially stronger regime (an additional SFT stage, 64 GPUs across 8 nodes, and more RL data); we include it here only for reference. Under identical controlled conditions, Qwen3-VL + G^{2}RPO consistently outperforms both OneThinker-8B and Qwen3-VL + EMA-GRPO, and our full model OpenVLThinkerV2 achieves the best performance overall.

Model	**General VQA**	**Math VQA**	**Chart VQA**	**Grounding**	**Document Understanding**	**Spatial Reasoning**
Qwen3-VL-Instruct-8B	71.3	59.2	69.9	87.1	86.8	60.9
OneThinker-8B	75.9	63.4	68.5	89.4	87.7	60.2
Qwen3-VL + EMA-GRPO	75.9	64.0	72.3	89.6	89.2	61.4
Qwen3-VL + G^{2}RPO	76.9	64.8	74.5	90.2	90.6	62.3
OpenVLThinkerV2 (Ours)	**77.9**	**66.2**	**76.0**	**90.7**	**91.4**	**63.6**

Table 6: Controlled comparison including a Qwen3-VL + EMA-GRPO baseline trained under identical data and compute. Scores are averaged over the three benchmarks in each task category.

D Policy Gradient Derivation of G^{2}RPO

To derive the policy gradient for our proposed G^{2}RPO, we begin with the fundamental expected return objective and systematically replace the linear standard advantage with our non-linear Optimal Transport advantage mapping.

The Base Policy Gradient. For an autoregressive language model parameterised by θ, given a query $q \sim \mathcal{D}$ and a response y, the standard policy gradient theorem dictates that the gradient of the expected return $\mathcal{J}(\theta)$ is driven by the advantage $A(q, y)$:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{q\sim\mathcal{D}, y\sim\pi_\theta}\left[\sum_{t=1}^{|y|} \nabla_\theta \log \pi_\theta(y_t|q, y_{<t}) A(q, y)\right] \tag{11}$$

Group Sampling and Importance Sampling. To stabilize training without a separate value network, we adopt the group-sampling strategy of GRPO. For a query q, we sample a group

of G responses $\mathcal{G} = \{y_1, \ldots, y_G\}$ from the behavior policy $\pi_{\theta_{\text{old}}}$. To optimize the new policy π_θ using these trajectories, we introduce the token-level importance sampling ratio:

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|q, y_{i,<t})} \tag{12}$$

Injecting the Gaussian Advantage $\widehat{A}^{\mathbf{G^2RPO}}$. In standard GRPO, the advantage $\widehat{A}_i$ is computed via scalar standardization, which is vulnerable to heavy-tailed distributions and multi-task scale discrepancies. We replace this with our Gaussian Optimal Transport advantage, denoted as $\widehat{A}_i^{\text{G}^2\text{RPO}}$

For the sampled group's rewards $\mathcal{R}_\tau = \{R_1, \ldots, R_G\}$, we calculate the rank-based empirical probability $p_i = \frac{\text{rank}(R_i) - 0.5}{G}$. The advantage is strictly mapped to the standard normal distribution $\mathcal{N}(0,1)$ using the inverse CDF Φ^{-1}. Incorporating our tie-breaking strategy over the set of indices $\mathcal{K}_{R_i}$ sharing identical rewards, the advantage becomes:

$$\widehat{A}_i^{\text{G}^2\text{RPO}} = \frac{1}{|\mathcal{K}_{R_i}|} \sum_{j \in \mathcal{K}_{R_i}} \sqrt{2}\,\text{erfinv}(2p_j - 1) \tag{13}$$

Because $\widehat{A}_i^{\text{G}^2\text{RPO}}$ perfectly matches the quantiles of $\mathcal{N}(0,1)$, it guarantees $\sum_i \widehat{A}_i^{\text{G}^2\text{RPO}} \approx 0$ and $\sum_i (\widehat{A}_i^{\text{G}^2\text{RPO}})^2 \approx 1$, inherently satisfying the zero-mean and unit-variance requirements for stable policy updates, while systematically mitigating the influence of reward outliers.

The Clipped Surrogate Objective. To prevent destructively large policy updates, we bind the policy update using a PPO-style clipping mechanism. By integrating our Gaussian advantage $\widehat{A}_i^{\text{G}^2\text{RPO}}$ into the surrogate objective, we define the formal loss function for G^{2}RPO:

$$\begin{aligned} \mathcal{J}_{\text{G}^2\text{RPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \\ \min \left(r_{i,t}(\theta) \widehat{A}_i^{\text{G}^2\text{RPO}}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right) \widehat{A}_i^{\text{G}^2\text{RPO}} \right) \Bigg] \end{aligned} \tag{14}$$

The Final Gradient Update. During backpropagation, the gradient of our surrogate objective with respect to θ determines the parameter updates. Let $\mathcal{L}_{\text{clip}}^{i,t}(\theta)$ denote the inner clipped term. The empirical gradient is evaluated as:

$$\nabla_\theta \mathcal{J}_{\text{G}^2\text{RPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \nabla_\theta \mathcal{L}_{\text{clip}}^{i,t}(\theta) \right] \tag{15}$$

where the gradient of the token-level loss evaluates to:

$$\nabla_\theta \mathcal{L}_{\text{clip}}^{i,t}(\theta) = \begin{cases} \widehat{A}_i^{\text{G}^2\text{RPO}}\, r_{i,t}(\theta) \nabla_\theta \log \pi_\theta(y_{i,t}|q, y_{i,<t}) & \text{if } r_{i,t}(\theta) \text{ is not clipped} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$