H-OPD: Confidence Aware Heterogeneous Multi-Teacher Multimodal On-policy Distillation

Qixiang Yin[1,5], Huanjin Yao[2], Cai Yuchen[3], Jianghao Chen[5], Ziyi Wang[1], Min Yang[1,*], Fei Su[1,4], Zhicheng Zhao[1,4,*]

[1] Beijing University of Posts and Telecommunications [2] HKUST [3] USTC
[4] Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism [5] Zhongguancun Academy, Beijing, China
[*] Corresponding Author {buptyqx,zhaozc}@bupt.edu.cn

Abstract

On-policy distillation (OPD) has recently emerged as an effective post-training paradigm by providing supervision on student-generated trajectories. However, existing OPD methods for multimodal reasoning usually rely on a static teacher routing, assigning each sample to a single teacher based on modality or task type. This ignores that visual grounding and abstract reasoning may dominate different decoding steps, making a single teacher insufficient for the full trajectory. To this end, H-OPD is proposed as a confidence-aware heterogeneous multi-teacher OPD framework for multimodal reasoning. By verifying the complementarity of heterogeneous teachers in the same reasoning process, H-OPD replaces task or sample level teacher routing with token-level teacher arbitration along the shared student trajectory. H-OPD employs vision-to-language description transfer to enable text-only teachers to access key visual semantics, and uses a confidence-aware arbitration mechanism to dynamically combine vision-language teacher and text-only teachers at each token. Extensive evaluations over 11 widely-used reasoning benchmarks showcase the superior performance of our method. Code is available at https://github.com/buptyqx/H-OPD.

1 Introduction

Effective post-training for multimodal reasoning remains an open challenge for MLLMs. Existing post-training frameworks typically follow a two-stage paradigm that combines supervised fine-tuning (SFT) (Ouyang et al., 2022a) with reinforcement learning (RL) (Guo et al., 2025). SFT relies on a fixed demonstration distribution and is therefore prone to exposure bias. RL can reduce distribution shift through online sampling, but it generally depends on sequence-level rewards and incurs much higher training costs. In contrast, on-policy distillation (OPD) provides efficient supervision on

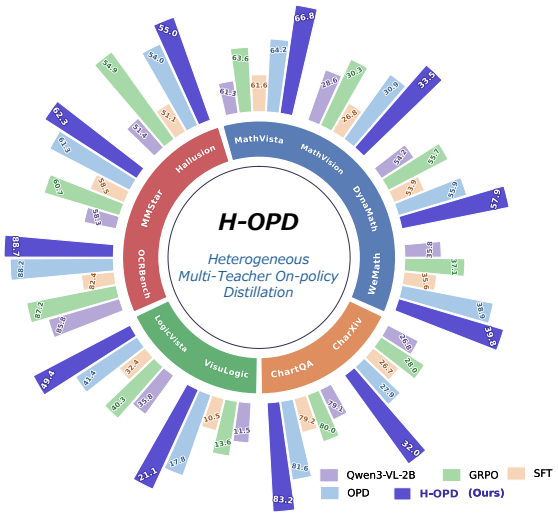


Figure 1: Performance comparison across 11 multimodal reasoning benchmarks. H-OPD consistently outperforms standard OPD and GRPO, leading to stronger overall performance.

student-generated trajectories and has emerged as a promising paradigm for reasoning post-training. However, an effective framework for high-quality, fine-grained, and token-state-aware supervision is still lacking in multimodal reasoning.

As illustrated in Fig. 2(a), existing approaches (Lu and Lab, 2025; Yang et al., 2026a) typically adopt relatively coarse-grained teacher routing strategies, they either train separate teacher-student systems for different modalities, or assign a single teacher to each input instance at the sample or task level, routing multimodal tasks to the vision-language(VL) teacher and text-only tasks to the text teacher. These methods commonly assume that the entire generation trajectory of a given input should be supervised by the same teacher. In fact, multimodal reasoning is not a static process governed by a single capability, but a dynamic autoregressive process with varying requirements for visual perception and abstract reasoning at different decoding steps. Prior studies show that multimodal

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 4780–4797
October 24-29, 2026 ©2026 Association for Computational Linguistics

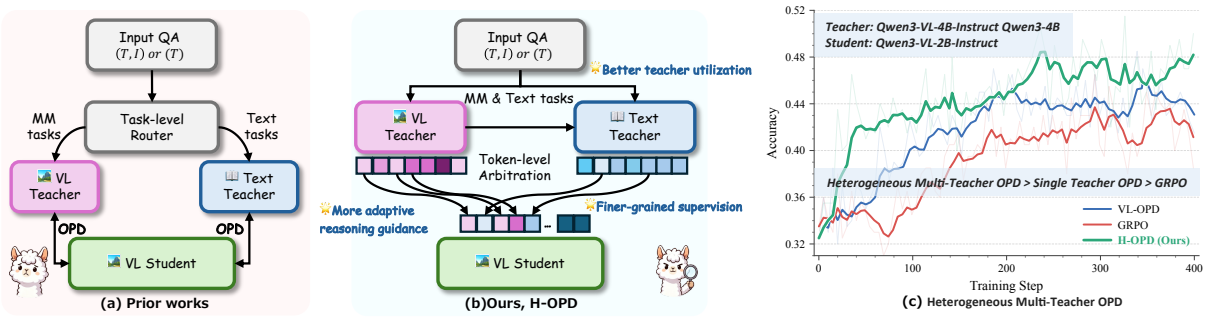


Figure 2: Comparison between conventional task-level routing and our token-level arbitration OPD framework. (a) Conventional routing assigns each input to a single heterogeneous teacher based on task type. (b) Our method exposes the same input to both a VL teacher and a text teacher, and performs token-level arbitration along the shared student trajectory. (c) H-OPD consistently outperforms OPD and GRPO, highlighting the advantage of heterogeneous multi-teacher OPD.

models can correctly perceive visual evidence yet still make reasoning errors, as perceptual information may be misinterpreted or overwritten during later reasoning process (Liu et al., 2025a; Wang et al., 2025; Xu et al., 2026). This suggests that teacher reliability fluctuates along the response trajectory depending on token states. Accordingly, teacher allocation should be refined from the task or sample level down to the token level.

Furthermore, we argue that VL and text-only teachers are not redundant in multimodal reasoning; instead, they provide fine-grained complementary supervision throughout the reasoning trajectory. The former is more effective at perception-heavy steps, while the latter can be more reliable at reasoning-intensive stages. Therefore, instead of statically assigning a single teacher to each sample, it is more appropriate to expose the same input to multiple heterogeneous teachers and dynamically arbitrate their supervisory weights at each decoding token. Fig. 2(b) illustrates this central idea: *rather than performing task-level routing, our method performs token-level arbitration along the shared student trajectory.*

Motivated by these observations, we propose **H-OPD**, a **h**eterogeneous multi-teacher OPD framework for multimodal reasoning. The core idea of H-OPD is to feed the same input into both VL and text-only teachers, and perform token-level teacher arbitration rather than task/sample-level teacher selection. Specifically, H-OPD first performs vision-to-language description transfer, converting key visual semantics of images into a textual proxy so that the text teacher lacking visual perception can also access task-relevant visual information in textual form. Building on this design, H-OPD further introduces a confidence-based token-level arbitration mechanism, which adaptively allocates supervisory weights according to the prediction entropy of each teacher at the current token, thereby enabling fine-grained collaborative distillation between visual grounding and abstract reasoning. As a result, heterogeneous teachers no longer compete as sample-level alternatives, but collaborate with each other at the token level. Empirically, as shown in Fig. 1 and Fig. 2(c), H-OPD substantially outperforms conventional OPD and GRPO, achieving both higher training efficiency and superior test reasoning accuracy. In summary, the main contributions of this work are as follows:

1. By exploring the transition from task/sample-level routing to token-level arbitration, we verify the complementarity of heterogeneous teachers in the same multimodal reasoning trajectory.
2. We propose H-OPD, to the best of our knowledge, the first heterogeneous multi-teacher on-policy distillation framework that enables token-level supervision by integrating complementary signals from VL and text-only teachers.
3. We conduct extensive experiments on 11 benchmarks, and the results demonstrate the superiority of H-OPD across diverse multimodal reasoning tasks.

2 Related Work

2.1 MLLM Reasoning

Multimodal reasoning enhances slow-thinking capabilities (Li et al., 2025), enabling models to tackle complex reasoning tasks (Lu et al., 2023). Early work mainly relied on instruction tuning (Liu et al., 2023) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022b; Yu

et al., 2024; Yao et al., 2024). However, scaling RLHF remains difficult due to its heavy dependence on high-quality human annotations (Lee et al., 2023). As a result, recent studies have explored more efficient RL-based strategies. Inspired by DeepSeek-R1 (Guo et al., 2025; Cai et al., 2025a), a line of work has introduced rule-based reward functions to provide reliable supervision without costly manual labeling (Liu et al., 2025b; Yao et al., 2025). These methods design fine-grained rewards from multiple perspectives, including answer correctness (Huang et al., 2025; Yao et al., 2026), Chain-of-Thought completeness (Zhang et al., 2025; Cai et al., 2025b), and visual consistency (Yang et al., 2025; Yin et al., 2026), improving the reasoning accuracy of MLLMs.

2.2 On Policy Distillation

Unlike traditional off-policy Knowledge Distillation (KD) that suffers from exposure bias resembling SFT (Ouyang et al., 2022a; Lu and Lab, 2025), OPD (Gu et al., 2024; Yang et al., 2026a; Hou et al., 2026) aligns students on self-generated rollouts. By treating the teacher's token-level log-probabilities as dense rewards for RL algorithms like GRPO (Yang et al., 2026b), OPD efficiently bridges KD and RL (Agarwal et al., 2024; Cai et al., 2026). Building on these insights, recent paradigms have expanded to self-distillation (Zhao et al., 2026; Hübotter et al., 2026; Ding, 2026) and multi-teacher settings (DeepSeek-AI, 2026; Xiao et al., 2026) to optimally fuse domain expert capabilities via unified framework (Yang et al., 2026a; Yuan et al., 2026). Inspired by these breakthroughs, our approach achieves collaborative supervision from both VL and text-only teachers, improving multimodal reasoning in student models.

3 Methodology

The proposed **H-OPD** adaptively arbitrates between heterogeneous teachers along the student's trajectory, allowing complementary supervision from visual grounding and abstract reasoning. We first introduce the preliminaries in Sec. 3.1, then motivate token-level arbitration in Sec. 3.2, present the modality-bridging design in Sec. 3.3, and finally describe the confidence-aware arbitration objective in Sec. 3.4.

3.1 Preliminaries

On-policy distillation. OPD (Gu et al., 2024) addresses the exposure bias inherent in traditional off-policy KD by training the student model on its own generated trajectories rather than fixed teacher demonstrations. Given an input prompt x, the student policy auto-regressively generates a sequence $y \sim \pi_\theta(\cdot|x)$. At each decoding step t, the teacher model evaluates this self-generated state, providing a target token-level distribution $\pi_T = \pi_T(\cdot|x, y_{<t})$ as feedback for the student's parameterized distribution $\pi_\theta = \pi_\theta(\cdot|x, y_{<t})$. The student is then updated by minimizing the reverse KL divergence over its on-policy data:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}\left[\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_T)\right] \quad (1)$$

By actively exploring the state space and receiving immediate corrective signals from the teacher on actual visited states, OPD ensures better train-inference consistency and robustly aligns the student's behavior.

3.2 Motivation: From Task/Sample-level Routing to Token-level Arbitration

A common strategy for heterogeneous teacher distillation is to perform *sample-level routing*: each input instance x is assigned to a single teacher deemed most appropriate for that example. For example, multimodal questions may be routed to a vision-language (VL) teacher, while text-only questions are routed to a text teacher. Let $m^\star(x) \in \{1, \ldots, M\}$ denote the selected teacher for input x. Then the entire student trajectory for x is supervised by the same teacher, $\pi^{(t)}_{\text{task}}(\cdot \mid x, y_{<t}) = \pi_{m^\star(x)}(\cdot \mid x, y_{<t})$. A slightly more general variant uses a fixed instance-level mixture over teachers:

$$\pi^{(t)}_{\text{task}}(\cdot \mid x, y_{<t}) = \sum_{m=1}^{M} \alpha_m(x)\, \pi_m(\cdot \mid x, y_{<t}), \quad (2)$$

where, $\alpha(x) \in \Delta^{M-1}$ denotes an instance-specific mixing weight over the M teachers. Hard routing is recovered as a special case when $\alpha(x)$ is one-hot form.

Despite their difference, both formulations assume that teacher authority is fixed for the entire input instance. However, OPD supervises the student on token states visited by its own on-policy rollout. Therefore, teacher reliability should depend on the current decoding state $s_t = (x, y_{<t})$, rather than only on the input instance x.

This motivates a new paradigm: *instead of assigning one teacher to the whole sample, we expose*

the same input to multiple heterogeneous teachers and arbitrate their supervision at the token level. Concretely, we define a state-dependent teacher mixture

$$\pi_{\text{tok}}^{(t)}(\cdot \mid s_t) = \sum_{m=1}^{M} \lambda_m(s_t)\, \pi_m(\cdot \mid s_t). \quad (3)$$

Here, $\lambda(s_t) \in \Delta^{M-1}$ denotes a state-dependent mixture weight over the M teachers. Thus, multiple teachers may contribute to the same reasoning trajectory, while their relative authority can vary across decoding steps. Task-level routing is recovered as a special case when $\lambda_m(s_t) = \alpha_m(x)$ for all t; in particular, hard routing corresponds to one-hot $\alpha(x)$. Hence, sample-level routing and task-level fixed mixtures are both contained in the more general family of token-level arbitration.

Let $\pi_t^\star$ denote the ideal teacher distribution at state s_t. Define $q_{\lambda,t} = \sum_{m=1}^{M} \lambda_m(s_t)\pi_{m,t}$ and $q_{\alpha,t} = \sum_{m=1}^{M} \alpha_m(x)\pi_{m,t}$, as the token-level and task-level teacher mixtures, respectively, where $\pi_{m,t} \equiv \pi_m(\cdot \mid s_t)$. Under any fixed distribution over visited token states, token-level arbitration admits an approximation error no worse than that of task-level routing:

$$\inf_{\lambda} \mathbb{E}_{s_t}[\mathrm{KL}(\pi_t^\star \,\|\, q_{\lambda,t})] \le \inf_{\alpha} \mathbb{E}_{s_t}[\mathrm{KL}(\pi_t^\star \,\|\, q_{\alpha,t})]\,. \quad (4)$$

The inequality holds strictly when the most reliable teacher changes across token states within the same reasoning trajectory. This is particularly plausible in multimodal reasoning, where different segments of the response rely on different capabilities, such as visual grounding, abstract reasoning, and linguistic fluency.

This raises a question: *given that the VL teacher has access to the image, what additional benefit can a text-only teacher provide?* Studies indicate that although multimodal models can accurately perceive visual evidence, they are prone to errors in reasoning, because perceptual information may be misinterpreted or overwritten during later inference (Liu et al., 2025a; Wang et al., 2025; Xu et al., 2026). Therefore, the VL teacher and the text-only teacher can complement each other both across instances and within the same reasoning process: the VL teacher offers stronger perceptual grounding, while the text-only teacher should achieve higher reliability at reasoning-heavy steps.

To verify this hypothesis, we conduct a pilot experiment on 100 questions sampled

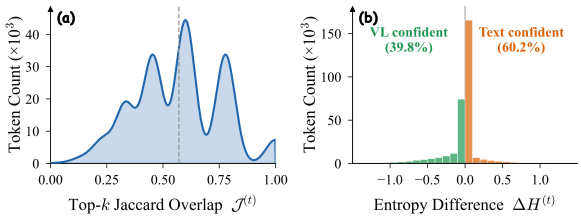


Figure 3: Pilot experiment on heterogeneous teacher distillation. (a) Distribution of top-k Jaccard overlap between the VL teacher and the text teacher. (b) Distribution of entropy difference, $\Delta H^{(t)} = H_{\text{VL}}^{(t)} - H_{\text{Text}}^{(t)}$.

from MMFineReason (Lin et al., 2026), using `Qwen3-VL-4B-Instruct` as the VL teacher and `Qwen3-4B-Instruct-2507` as the text teacher. We probe their token-level complementarity using two diagnostics. The first is the top-k Jaccard overlap, $\mathcal{J}^{(t)} = \frac{|\mathcal{S}_{\text{VL}}^{(t)} \cap \mathcal{S}_{\text{Text}}^{(t)}|}{|\mathcal{S}_{\text{VL}}^{(t)} \cup \mathcal{S}_{\text{Text}}^{(t)}|}$, which measures the overlap between their high-probability token sets at decoding step t. The second is the entropy difference, $\Delta H^{(t)} = H_{\text{VL}}^{(t)} - H_{\text{Text}}^{(t)}$, where positive values indicate higher uncertainty from the VL teacher and negative values indicate higher uncertainty from the text teacher.

Fig. 3 supports this view. The Jaccard overlap is far from concentrated near 1, indicating that the two teachers tend to favor different candidate tokens at the same position. Meanwhile, the entropy difference spans both positive and negative regions with substantial mass, showing that neither teacher is uniformly more confident throughout decoding. Overall, these observations suggest that the two teachers provide complementary supervision within the same multimodal reasoning trajectory, thereby motivating token-level arbitration over sample-level routing.

3.3 Bridging the Modality Gap with Vision-to-Language Description Transfer

As shown in Fig. 4, given a multimodal text-image input (x, I), the student policy π_θ autoregressively generates an output sequence y, with the next-token distribution at step t defined as $p_\theta^{(t)}(v) = \pi_\theta(v \mid x, I, y_{<t})$. Unlike standard online distillation relying on a single modality-aligned teacher, our framework employs two teachers with asymmetric modality access: a vision-language teacher τ_V conditioning on (x, I), and a text-only teacher τ_T restricted to x. Since τ_T intrinsically lacks visual grounding, its direct application yields unfaithful supervision. Thus, the core challenge lies in effec-

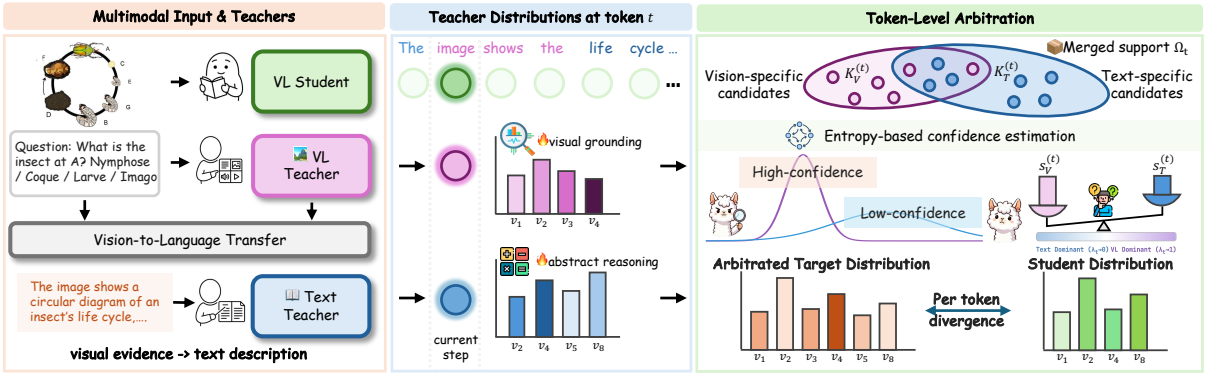


Figure 4: Overview of H-OPD, which consists of a vision-to-language description transfer module for enabling the text-only teacher to access visual semantics, and a confidence-aware arbitration mechanism for dynamically integrating the vision-language teacher and the text-only teacher at the token level.

tively injecting visual context into τ_T while synergizing the complementary capabilities of both teachers during online training.

Vision-to-Language Description Transfer. Rather than forcing the text teacher to process visual features directly, we translate visual semantics into explicit textual proxies seamlessly consumable by τ_T. Specifically, for a given sample (x, I), we leverage the visionlanguage teacher to extract a task-relevant image description d. We then synthesize a modality-adapted prompt:

$$\tilde{x} = x^{\text{txt}} \oplus [\texttt{IMAGE DESCRIPTION}] \oplus d, \quad (5)$$

where x^{txt} denotes the original question text with the <image> placeholder removed, and $\oplus$ denotes string concatenation. This transformation disentangles the multimodal reasoning task by offloading visual *perception* into the textual proxy d, thereby enabling the text-only teacher τ_T to focus exclusively on logical *reasoning*. Because this perceptual stage is distilled into d offline during data preprocessing, it introduces no additional serial overhead in the online rollout-and-distill loop.

Heterogeneous Teacher and Student Policies. Given the augmented prompt, the two teachers supervise the same student prefix through two parallel inference paths. At step t, the vision-language teacher produces $q_V^{(t)}(v) = \tau_V(v \mid x, I, y_{<t})$, while the text teacher produces $q_T^{(t)}(v) = \tau_T(v \mid \tilde{x}, y_{<t})$. To ensure tractability, we restrict the supervision to the top-k predictions from each teacher. Let $K_V^{(t)}$ and $K_T^{(t)}$ denote the corresponding top-k token sets, and let $\bar{q}_V^{(t)}$ and $\bar{q}_T^{(t)}$ denote the teacher distributions renormalized on their returned supports. The unified candidate vocabulary is defined as:

$$\Omega_t = K_V^{(t)} \cup K_T^{(t)}, \qquad |\Omega_t| \leq 2k. \quad (6)$$

When extending $\bar{q}_V^{(t)}$ and $\bar{q}_T^{(t)}$ to Ω_t, tokens absent from a teacher's top-k set are assigned zero probability. Crucially, conditioned on $y_{<t}$, the two teachers infer independently and asynchronously, bounding the per-step latency by the slower teacher alone.

3.4 Confidence-Aware Token-Level Teacher Arbitration

The two teachers provide token-level supervision over the unified candidate set Ω_t. Since different tokens rely on visual grounding and abstract reasoning to different extents, static ensembling is suboptimal, especially in the asymmetric modality setting. Confidence-Aware Token-Level Teacher Arbitration assigns token-wise supervisory emphasis according to predictive uncertainty. Teacher reliability is measured by predictive entropy, with higher entropy corresponding to greater uncertainty and lower entropy corresponding to higher confidence. For the renormalized teacher distributions over Ω_t, the confidence scores are defined as:

$$s_V^{(t)} = -H\left(\bar{q}_V^{(t)}\right) \qquad s_T^{(t)} = -H\left(\bar{q}_T^{(t)}\right) \quad (7)$$

where $H(\cdot)$ denotes Shannon entropy. These confidence scores are then transformed into a token-level arbitration weight:

$$\lambda_t = \frac{\exp\left(s_V^{(t)}/\tau_f\right)}{\exp\left(s_V^{(t)}/\tau_f\right) + \exp\left(s_T^{(t)}/\tau_f\right)} \quad (8)$$

where $\tau_f > 0$ is a temperature parameter controlling the sharpness of arbitration. As $\tau_f \to \infty$,

Table 1: Main results across multiple capability benchmarks. `Self` denotes the student model itself, Qwen3-VL-xB denotes `Qwen3-VL-`x`B-Instruct`. We report open-source checkpoints,and post-training results under different teacher configurations.* denotes evaluation on official weights using VLMEvalKit (Duan et al., 2024).

Setup (Teacher/Method)	Mathematics				Chart		Logic		General			Avg.
	Math Vista	Math Vision	Dyna Math	We Math	CharXiv (RQ)	Chart QA	Visu Logic	Logic Vista	OCR Bench(/10)	MMStar	Hallusion Bench	
Baseline Models												
Qwen3-VL-2B	61.3	28.6*	54.2	35.8*	26.8	79.1	11.5	35.8	85.8	58.3	51.4	48.1
Qwen3-VL-4B	73.7	44.7*	65.3	54.3*	39.7	84.6	19.0	53.2	88.1	69.8	57.6	59.1
Qwen3-VL-8B	77.2	45.4*	67.7	57.5*	46.4	89.6	22.5	55.3	89.6	70.4	61.1	62.1
Student: Qwen3-VL-2B-Instruct												
Self / GRPO (Guo et al., 2025)	63.6	30.3	55.7	37.1	28.0	80.0	13.6	40.3	87.2	60.7	54.9	50.1
Qwen3-VL-4B / SFT	61.6	26.8	53.9	35.6	26.7	79.2	10.5	32.4	82.4	58.5	51.1	47.2
Qwen3-VL-4B / OPD (Lu and Lab, 2025)	64.2	30.9	55.9	38.9	27.9	81.6	17.8	41.4	88.2	61.3	54.0	51.1
Qwen3-VL-4B / ExOPD (Yang et al., 2026b)	64.4	31.6	56.3	39.1	28.3	82.4	18.1	39.6	88.4	61.2	54.5	51.3
Qwen3-VL-4B + Qwen3-4B / H-OPD	**66.8**	**33.5**	**57.9**	**39.8**	**32.0**	**83.2**	**21.1**	**49.4**	**88.7**	**62.3**	**55.0**	**53.5**
Qwen3-VL-8B / SFT	62.2	28.2	55.1	36.4	26.8	80.4	10.9	33.6	83.3	60.7	51.3	48.1
Qwen3-VL-8B / OPD (Lu and Lab, 2025)	65.7	30.3	58.1	38.3	32.0	81.2	16.2	42.5	88.0	62.0	54.9	51.7
Qwen3-VL-8B / ExOPD (Yang et al., 2026b)	66.1	30.6	58.3	38.6	31.9	82.0	16.8	43.6	87.2	62.1	55.1	52.0
Qwen3-VL-8B + Qwen3-8B / H-OPD	**67.2**	**36.8**	**59.5**	**39.7**	**33.6**	**84.8**	**20.8**	**50.3**	**89.5**	**62.7**	**55.4**	**54.6**
Student: Qwen3-VL-4B-Instruct												
Self / GRPO (Guo et al., 2025)	76.1	46.7	67.9	56.2	40.9	88.0	21.0	54.1	89.1	70.0	60.2	60.9
Qwen3-VL-8B / SFT	72.1	42.1	65.5	54.3	39.6	84.8	18.1	51.5	87.7	68.0	55.4	58.1
Qwen3-VL-8B / OPD (Lu and Lab, 2025)	76.6	46.1	68.3	55.7	40.5	88.8	22.4	53.5	88.2	68.8	**60.9**	60.9
Qwen3-VL-8B / ExOPD (Yang et al., 2026b)	77.2	46.4	68.9	55.8	40.6	88.8	22.9	55.5	88.2	68.9	**60.9**	61.3
Qwen3-VL-8B + Qwen3-8B / H-OPD	**77.9**	**46.7**	**69.9**	**56.3**	**41.0**	**88.9**	**23.1**	**55.9**	**89.3**	**70.0**	59.3	**61.7**

the weighting approaches uniform averaging; as $\tau_f \to 0$, it approaches hard selection of the more confident teacher. The resulting arbitrated target distribution is

$$q_A^{(t)}(v) = \lambda_t \bar{q}_V^{(t)}(v) + (1 - \lambda_t)\bar{q}_T^{(t)}(v) \quad (9)$$

where $v \in \Omega_t$.This construction enables supervision to adapt to the requirements of the current token, visually grounded tokens are influenced more strongly by τ_V, whereas reasoning-dominant tokens place greater weight on τ_T.

Given the arbitrated target distribution, the student policy is optimized by minimizing the reverse KL divergence over the restricted support:

$$\mathcal{L}(\theta) = \mathbb{E}_{y\sim\pi_\theta}\left[\sum_{t=1}^{|y|} m_t\, D_{\mathrm{KL}}^{\Omega_t}\left(p_\theta^{(t)} \,\|\, q_A^{(t)}\right)\right] \quad (10)$$

where $m_t \in \{0, 1\}$ denotes the mask over valid response positions, and $D_{\mathrm{KL}}^{\Omega_t}$ denotes the KL divergence evaluated only on Ω_t.

Restricting distillation to the merged top-k support improves both efficiency and stability. Since $|\Omega_t| \leq 2k$, the per-step supervision cost is bounded by $\mathcal{O}(k)$, while the learning signal remains concentrated on the most relevant candidates.

4 Experiments

4.1 Experiments Setup

Training Dataset. We train our models on MMFineReason-123K (Lin et al., 2026). Under the same experimental setting, we use a multimodal teacher model to generate image descriptions and employ GPT-4.1-mini (Achiam et al., 2023) to assess the correctness of these descriptions. After filtering, we obtain 55K training samples, which are used consistently across all training settings.

Models. We conduct experiments on the Qwen3-VL (Bai et al., 2025) model family. Specifically, we use Qwen3-VL-4B-Instruct and Qwen3-VL-2B-Instruct as student models. For heterogeneous teacher ensembles, we employ a combination of Qwen3-VL-4B-Instruct, Qwen3-4B-Instruct-2507, Qwen3-VL-8B-Instruct, and Qwen3-8B.

Implementation Details. We implement H-OPD based on the verl (Sheng et al., 2025) framework, with vLLM (Kwon et al., 2023) used as the backend for both rollout generation and teacher inference. During training, the maximum prompt length and maximum response length are both set to 12,384 tokens. For teacher arbitration, we adopt confidence-aware fusion as the fusion strategy, with

the fusion temperature τ_f set to 1.0. The student models are optimized using Adam with a learning rate of 1×10^{-6} and a training batch size of 128. All experiments are conducted on a single compute node equipped with 8 NVIDIA B200 180GB GPUs.

Evaluation Benchmarks. We assess H-OPD using a comprehensive benchmark suite spanning four major axes of multimodal capability. Specifically, for multimodal mathematical reasoning, we include MathVista (Lu et al., 2023), MathVision (Wang et al., 2024a), DynaMath (Zou et al., 2025), and WeMath (Qiao et al., 2025); for multimodal document understanding, we consider CharXiv (Wang et al., 2024b) and ChartQA (Masry et al., 2022); for multimodal logical reasoning, we evaluate on VisuLogic (Xu et al., 2025) and LogicVista (Xiao et al., 2024); and for general multimodal visual question answering, we use OCRBench (Liu et al., 2024), MMStar (Chen et al., 2024), and HallusionBench (Guan et al., 2024). Together, these benchmarks provide a broad and systematic assessment of H-OPD across a wide range of multimodal reasoning tasks.

4.2 Main Results

To provide a comprehensive evaluation of H-OPD, we conduct experiments on models of different scales, including 2B, 4B, and 8B variants. As shown in Table 1, we organize the analysis into two parts. We first compare H-OPD against representative post-training baselines, including GRPO (Guo et al., 2025), OPD (Lu and Lab, 2025), and ExOPD (Yang et al., 2026b). We then examine its scaling properties by analyzing the effects of increasing both student and teacher model sizes.

Comparison with post-training strategies. We first compare H-OPD against representative post-training strategies on the Qwen3-VL-2B-Instruct student model. As shown in Table 1, H-OPD achieves a 5.5% improvement in average score over the baseline instruct model. In addition, compared with OPD and ExOPD, which rely on distillation from single VL teacher, H-OPD improves the average score from 51.1 and 51.3 to 53.6 under the 4B teacher setting, yielding relative gains of 4.9% and 4.5%, respectively. The gains are particularly pronounced on challenging reasoning benchmarks such as MathVista, LogicVista, and CharXiv. This demonstrates that combining supervision from heterogeneous teacher allows H-OPD to transfer complementary multimodal perception and textual reasoning abilities more effectively, leading to superior multimodal reasoning performance over existing post-training baselines.

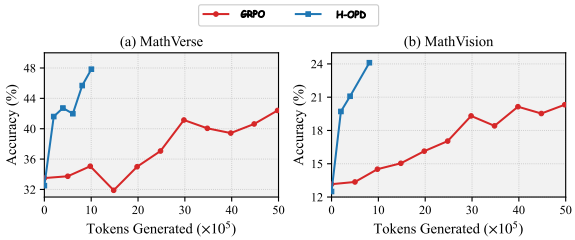


Figure 5: Comparison of training token efficiency between GRPO and H-OPD on (a) MathVerse and (b) MathVision.

Scaling up. As model scale increases on both the teacher and student sides, H-OPD continues to show strong performance. For the Qwen3-VL-2B-Instruct student, replacing the 4B teacher pair with the 8B teacher pair improves the average score from 53.6 to 54.6, while maintaining a clear advantage over OPD and ExOPD under the same teacher configuration. This suggests that H-OPD can effectively benefit from stronger teachers. In addition, when applied to a larger Qwen3-VL-4B-Instruct student, H-OPD improves the average score from the baseline model's 59.1 to 61.7, approaching the performance of its teacher Qwen3-VL-8B-Instruct.The improvements are particularly stable on multimodal mathematical reasoning and logical reasoning benchmarks, showing that H-OPD scales effectively on challenging cross-modal reasoning tasks. Overall, these results suggest that H-OPD benefits from both stronger teachers and larger students, highlighting its robustness as a general post-training framework for MLLM reasoning.

4.3 Discussion

Token Efficiency. Figure 5 shows that H-OPD is substantially more token-efficient than GRPO. Across both MathVerse and MathVision, H-OPD attains higher accuracy within the first 100 training steps while using far fewer generated tokens, indicating faster convergence and more effective learning from each rollout.

Stronger RL Teacher. We further evaluate H-OPD with a stronger RL-trained teacher. Specifically, the teacher model is a version of Qwen3-VL-4B-Instruct trained with GRPO on the same training data. As shown in Table 2, compared with OPD and ExOPD, H-OPD improves the average score by 3.9 and 4.0 points, respectively. These

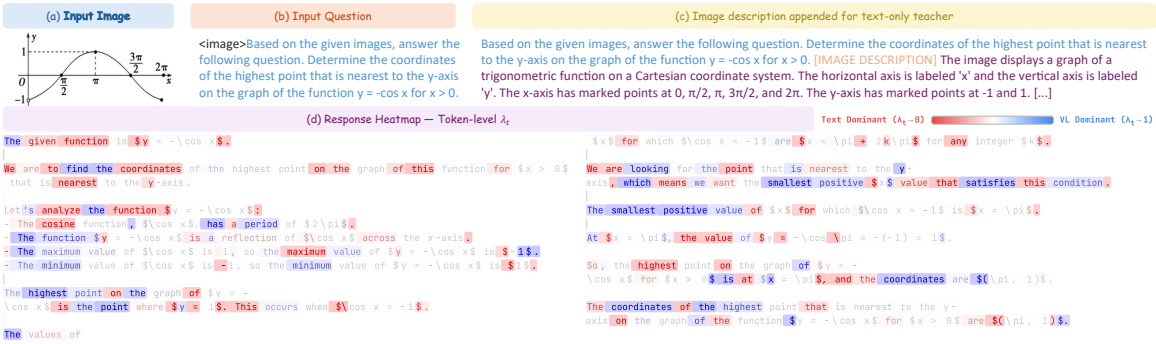


Figure 6: Visualization of the token-level teacher arbitration in H-OPD. (a) Input image. (b) Input question. (c) Textual description of the original image constructed for the text teacher. (d) Token-level heatmap of λ_t.

Table 2: Results with a stronger RL-trained teacher on selected benchmarks. Qwen3-VL-2B-Instruct is used as the student, and Qwen3-VL-4B-RL is a stronger teacher further trained with GRPO on the same data.

Method	MathVista	CharXiv	LogicVista	MMStar
Qwen3-VL-2B(S)	61.3	26.8	35.8	58.3
Qwen3-VL-4B-RL(T)	76.1	40.9	54.1	70.0
OPD	68.8	32.4	42.5	63.3
ExOPD	69.2	30.2	42.9	64.0
H-OPD(Ours)	**69.9**	**34.5**	**51.4**	**66.6**

Table 3: Ablation studies on key hyperparameters in H-OPD.

(a) The impact of τ_f

τ_f	MathVista	MathVision
0.1	66.1	31.0
0.5	66.8	33.5
1.0	66.2	32.9
1.5	66.2	32.3

(b) The impact of top-k

k	MathVista	Time/step
4	66.2	62 s
8	66.8	92 s
32	67.1	718 s

results suggest that H-OPD can more effectively exploit the capabilities of a stronger RL-trained teacher, enabling better transfer of advanced reasoning skills to the student model.

4.4 Ablation Study

We conduct ablation studies with Qwen3-VL-2B-Instruct as the student and Qwen3-VL-4B-Instruct together with Qwen3-4B as the teachers. As shown in Table 3, $\tau_f = 0.5$ achieves the best performance on both MathVista and MathVision, indicating that a moderate fusion temperature provides the best balance between the two teacher signals. For top-k, setting $k = 8$ offers a good trade-off between performance and efficiency, while a larger value brings only marginal gains at higher cost.

4.5 Qualitative Evaluation

To intuitively demonstrate the effectiveness of the token-level teacher arbitration coefficient λ_t in H-OPD, we visualize its behavior with a token-level heatmap. As shown in Figure 6(c), we first present the textual description of the original image generated via Vision-to-Language Description Transfer, which is used to construct the rollout trajectory for the text teacher. Figure 6(d) further shows the token-wise values of λ_t, where color indicates whether each token is predominantly dominant by the text teacher or the vision-language teacher. From the visualization, we observe a clear functional specialization between the two teachers: the vision-language teacher mainly supervises tokens associated with visual perception, while the text teacher primarily guides tokens related to logical reasoning. This observation verifies that H-OPD delivers complementary rather than redundant supervision signals, allowing distinct teacher models to guide different aspects of the student's capabilities.

5 Conclusion

We propose H-OPD, a confidence-aware heterogeneous multi-teacher on-policy distillation framework for multimodal reasoning. By shifting from task and sample-level teacher routing to token-level teacher arbitration, H-OPD enables a vision-language teacher and a text-only teacher to provide complementary supervision within the same reasoning trajectory. Through vision-to-language description transfer and confidence-aware arbitration, H-OPD effectively combines visual grounding and abstract reasoning signals during student generation. Experiments on 11 multimodal benchmarks show that H-OPD consistently outperforms strong post-training baselines and achieves higher token efficiency, retaining efficacy with advanced teachers and larger student models.

6 Limitations

In this section, we discuss the limitations of our method. First, the effectiveness of the text-only teacher depends on the quality of the transferred image description used to bridge the modality gap, when this description is incomplete, noisy, or weakly aligned with the task, the resulting supervision may be suboptimal. Second, our arbitration uses predictive entropy over truncated top-k distributions as a proxy for teacher reliability, but low entropy does not always imply token-level correctness, especially when teachers are overconfident or poorly calibrated. While merged top-k supervision improves efficiency, multi-teacher querying still introduces additional training overhead, and truncation may discard informative low-probability candidates. In addition, due to limited computational resources, we do not evaluate our method on larger-scale settings such as MoE models or 32B-level dense models. Although we discuss scalability in the main text, further validation at larger scales remains important future work.

7 Acknowledgments

This work was supported by the Key Research and Development Program of Xizang (XZ202601ZY0042), Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing (GJJ-26) and the BUPT Innovation and Entrepreneurship Support Program (2025-YC-T033).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*.

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, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Yuchen Cai, Ding Cao, Liang Lin, Chunxi Luo, Xin Xu, Kai Yang, Weijie Liu, Saiyong Yang, Tianxiang Zhao, Guangzhong Sun, and 1 others. 2026. Learning to foresee: Unveiling the unlocking efficiency of on-policy distillation. *arXiv preprint arXiv:2605.11739*.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guiquan Liu, and Junfeng Fang. 2025a. On predictability of reinforcement learning dynamics for large language models. *arXiv preprint arXiv:2510.00553*.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guiquan Liu, and Junfeng Fang. 2025b. On predictability of reinforcement learning dynamics for large language models. *arXiv preprint arXiv:2510.00553*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Ken Ding. 2026. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, and 1 others. 2026. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and 1 others. 2026. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. 2026. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. *arXiv preprint arXiv:2601.21821*.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025a. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

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

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/on-policy-distillation.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022b. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and

1 others. 2025. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, and 1 others. 2026. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Haolei Xu, Haiwen Hong, Hongxing Li, Rui Zhou, Yang Zhang, Longtao Huang, Hui Xue, Yongliang Shen, Weiming Lu, and Yueting Zhuang. 2026. Seeing but not thinking: Routing distraction in multimodal mixture-of-experts. *arXiv preprint arXiv:2604.08541*.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. 2026a. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2025. Look-back: Implicit visual re-focusing in mllm reasoning. *arXiv preprint arXiv:2507.03019*.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026b. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. 2026. Mm-deepresearch: A simple and effective multimodal agentic search baseline. *arXiv preprint arXiv:2603.01050*.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and 1 others. 2025. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*.

Qixiang Yin, Huanjin Yao, Jianghao Chen, Jiaxing Huang, Zhicheng Zhao, and Fei Su. 2026. Towards efficient multimodal unified reasoning model via model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9378–9388.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. 2026. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

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

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pages 48337–48383.

Appendix

A Benchmark

A.1 Multimodal Mathematical Reasoning

- **MathVista** (Lu et al., 2023): MathVista evaluates the comprehensive mathematical reasoning capabilities of models within complex visual contexts, such as geometric figures, function plots, scientific diagrams, and natural images. In this paper, we utilize the mathvista MINI subset, which contains 1,000 questions.

- **MathVision** (Wang et al., 2024a): MathVision is a visual mathematical reasoning benchmark specifically designed for real-world mathematics competition-level problems. It covers 16 distinct mathematical disciplines categorized into 5 difficulty levels. In our evaluation, we use the mathvision MINI subset, comprising 304 questions.

- **DynaMath** (Zou et al., 2025): DynaMath evaluates the robustness of vision-language models in mathematical reasoning by dynamically modifying visual numerical values or function-graph features within the same problem. The benchmark includes 501 high-quality seed questions across multiple topics, each expanded into 10 variants, for a total of 5,010 questions.

- **WeMath** (Qiao et al., 2025): WeMath aims to explore and evaluate the human-like problem-solving processes of large language models. WeMath deconstructs complex visual mathematical problems into step-wise sub-problems to deeply investigate the knowledge application and underlying logic during the model's reasoning process. In our experiments, a total of 1,740 questions from this benchmark are utilized.

A.2 Document/Chart Understanding

- **CharXiv (RQ)** (Wang et al., 2024b): CharXiv (RQ) is a benchmark designed to evaluate the understanding of authentic and complex charts sourced from scientific literature. This dataset comprises 2,323 highly challenging chart question-answering pairs, manually curated and verified by domain experts.

- **ChartQA** (Masry et al., 2022): ChartQA is specifically tailored for chart question answering, demanding models to integrate visual feature extraction, data reading, and logical reasoning to answer complex questions concerning real-world charts. For our evaluation, we utilize the chartqa test split, which consists of 1,250 questions.

A.3 Multimodal Logical Reasoning

- **VisuLogic** (Xu et al., 2025): VisuLogic is a purely vision-centric logical reasoning benchmark. It is dedicated to evaluating the native visual logical reasoning capabilities in spatial, attribute, and quantitative pattern recognition. The dataset contains 1,000 human-verified visual logic multiple-choice questions.

- **LogicVista** (Xiao et al., 2024): LogicVista evaluates the comprehensive logical cognitive abilities of models within everyday visual contexts. The benchmark consists of 448 meticulously designed multiple-choice questions, each accompanied by detailed, human-annotated reasoning processes.

A.4 General Multimodal Visual Question Answering

- **OCRBench** (Liu et al., 2024): OCRBench comprehensively assesses the OCR and text reasoning capabilities of multimodal models. The benchmark comprises 1,000 manually filtered questions.

- **MMStar** (Chen et al., 2024): MMStar is an elite-level vision-dependent benchmark, primarily introduced to mitigate the data leakage issue prevalent in traditional benchmarks. It contains 1,500 highly challenging questions subjected to rigorous human review, ensuring that each problem strictly necessitates visual content for its resolution.

- **HallusionBench** (Guan et al., 2024): HallusionBench is specifically developed to diagnose and evaluate language hallucination and visual illusion in multimodal large models. The benchmark comprises 346 images and 1,129 meticulously designed adversarial question-answering pairs.

B H-OPD Details

This section provides supplementary implementation details omitted from the main method description for clarity, with a focus on practical aspects

Algorithm 1 H-OPD Training with Confidence-Aware Token-Level Arbitration

Require: Training set $\mathcal{D} = \{(x, I)\}$; student π_θ; VL teacher π_V; text teacher π_T; top-k size k; temperature τ_f

Stage I: Offline vision-to-language transfer

for all $(x, I) \in \mathcal{D}$ **do**
  Generate a task-relevant description d from image I using π_V
  Construct $\tilde{x} \leftarrow x^{\text{txt}} \oplus [\texttt{IMAGE DESCRIPTION}] \oplus d$, and store $(x, I, \tilde{x})$
end for

Stage II: Online on-policy distillation

repeat
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  for all $(x, I, \tilde{x}) \in \mathcal{B}$ **do**
    Roll out $y \sim \pi_\theta(\cdot \mid x, I)$ and initialize $\mathcal{L} \leftarrow 0$
    for $t = 1$ to $|y|$ **do**
      Obtain student prediction $\pi_\theta^{(t)} \leftarrow \pi_\theta(\cdot \mid x, I, y_{<t})$
      Query teachers $\pi_V^{(t)} \leftarrow \pi_V(\cdot \mid x, I, y_{<t})$ and $\pi_T^{(t)} \leftarrow \pi_T(\cdot \mid \tilde{x}, y_{<t})$
      Extract teacher candidates $K_j^{(t)} \leftarrow \mathrm{TopK}(\pi_j^{(t)}, k)$ for $j \in \{V, T\}$, set $\Omega_t \leftarrow K_V^{(t)} \cup K_T^{(t)}$
      Truncate and normalize each teacher distribution on its candidates:

$$\bar{\pi}_j^{(t)}(v) \leftarrow \frac{\pi_j^{(t)}(v)\mathbf{1}[v \in K_j^{(t)}]}{\sum_{u \in K_j^{(t)}} \pi_j^{(t)}(u)}, \qquad j \in \{V, T\},\ v \in \Omega_t$$

      Measure teacher confidence by entropy: $s_j^{(t)} \leftarrow -H(\bar{\pi}_j^{(t)})$, for $j \in \{V, T\}$
      Compute the confidence-aware arbitration weight: $\lambda_t \leftarrow \dfrac{\exp(s_V^{(t)}/\tau_f)}{\exp(s_V^{(t)}/\tau_f) + \exp(s_T^{(t)}/\tau_f)}$
      Fuse teacher predictions: $\pi_A^{(t)} \leftarrow \lambda_t \bar{\pi}_V^{(t)} + (1 - \lambda_t)\bar{\pi}_T^{(t)}$
      Accumulate token-level distillation loss: $\mathcal{L} \leftarrow \mathcal{L} + m_t D_{\mathrm{KL}}^{\Omega_t}\big(\pi_\theta^{(t)} \,\|\, \pi_A^{(t)}\big)$
    end for
    Update θ using $\nabla_\theta \mathcal{L}$
  end for
until convergence

of data construction, teacher querying, restricted-support distillation, and optimization. As shown in Algorithm 1, we present the full training pipeline and further elaborate on several necessary design details.

B.1 Vision-to-Language Description Construction

For each training sample, we first construct a text-only surrogate input for the text teacher by generating an image description with the VL teacher. In practice, this description is prompted to capture the visual information that is most relevant to downstream reasoning, rather than producing a generic caption. We use a concise instruction that encourages the model to summarize objects, quantities, spatial relations, text in the image, and other visually grounded cues that may affect the final answer. Prompt B.1 presents the full prompt used in the transfer process. The generated description is then concatenated with the original textual question to form the text teacher input.

To reduce noise propagation, we keep the description-generation prompt fixed across all training samples and use deterministic decoding. We additionally truncate overly long descriptions to a predefined maximum length. This design stabilizes the transferred visual context throughout training and avoids introducing unnecessary variance into the text-teacher branch. Furthermore, after constructing the training data, we employ GPT-4.1-mini as a filter to remove samples whose generated

descriptions are incorrect, inaccurate, or incomplete.

Vision-to-Language Description Transfer Prompt

```
You are an expert vision-to-text
translator. Please observe the image
carefully and follow these rules to
provide an objective and exhaustive
transcription:

1. For general natural scenes: Describe
the content in extreme detail (subjects,
actions, environment, colors, spatial
relationships).
2. For mathematical or geometric problems:
 Rigorously describe the geometric
structures, lines, coordinate axes, ALL
numerical values, alphabetical labels,
and math symbols.
3. For tables or structured data:
Transcribe directly into Markdown table
or LaTeX code.
4. For text/OCR: Transcribe verbatim and
accurately.

Original Question:
{clean_question}

CRITICAL: DO NOT output any greetings,
conversation, or "Okay". Output the pure
visual description directly below.
```

B.2 On-Policy Rollout

Given a rollout $y = (y_1, \ldots, y_T)$ from the student, at each step t we query the teachers using the same prefix $y_{<t}$ so that all distributions are conditioned on the identical student state. This ensures that teacher feedback is aligned with the trajectory actually visited by the student. For the VL teacher, the input consists of the original multimodal sample; for the text teacher, the input consists of the question concatenated with the transferred image description. We compute teacher distributions independently at every decoding step, and no hidden states are shared across the two teachers. In our implementation, teacher outputs are obtained with standard next-token logits, followed by softmax normalization and top-k truncation.

Since the two teachers may assign probability mass to different candidate sets, H-OPD performs distillation on the merged support $\Omega_t = K_V^{(t)} \cup K_T^{(t)}$. Let $\pi_\theta^{\Omega_t}(\cdot \mid s_t)$ denote the student distribution restricted to Ω_t and renormalized over this set:

$$\pi_\theta^{\Omega_t}(v \mid s_t) = \frac{\pi_\theta(v \mid s_t)}{\sum_{u \in \Omega_t} \pi_\theta(u \mid s_t)} \tag{11}$$

The arbitrated teacher target $\pi_A^{(t)}$ is already defined on Ω_t. We then compute the stepwise restricted-support KL divergence as

$$D_{\mathrm{KL}}^{\Omega_t}\left(\pi_\theta^{(t)} \,\middle\|\, \pi_A^{(t)}\right) = \sum_{v \in \Omega_t} \pi_\theta^{\Omega_t}(v \mid s_t) \log \frac{\pi_\theta^{\Omega_t}(v \mid s_t)}{\pi_A^{(t)}(v \mid s_t)}. \tag{12}$$

This formulation avoids forcing the student to match teacher probabilities over the entire vocabulary, while still preserving competition among the candidate tokens that either teacher considers important.

B.3 Support Merging and Zero-Filling

After obtaining the top-k token sets $K_V^{(t)}$ and $K_T^{(t)}$, we construct the union support Ω_t. Each teacher distribution is represented on Ω_t by assigning zero probability to tokens outside its own top-k set. Formally, for a teacher-specific truncated distribution $\bar{\pi}_i^{(t)}$ with support $K_i^{(t)}$, we use

$$\tilde{\pi}_i^{(t)}(v) = \begin{cases} \bar{\pi}_i^{(t)}(v), & v \in K_i^{(t)}, \\ 0, & v \in \Omega_t \setminus K_i^{(t)}, \end{cases} \tag{13}$$

where $i \in \{V, T\}$. The final arbitrated distribution is then computed on the shared support using these zero-filled vectors. In practice, this implementation is simple and avoids introducing additional approximation beyond the top-k truncation already used in the method.

The token-level mask m_t is used to exclude positions that are not suitable for distillation. In particular, we set $m_t = 0$ for padded positions and for decoding steps after the end-of-sequence token has already been produced. For valid response tokens, we set $m_t = 1$ and apply the distillation loss normally.

Depending on the data format, certain special tokens can also be excluded from distillation if they do not carry semantic content for reasoning, such as template separators or formatting markers introduced by the chat prompt. This masking rule prevents such tokens from dominating the loss and makes the optimization more focused on content-bearing decisions.

B.4 Optimization and Reproducibility Details

We train the student with standard minibatch optimization and compute the H-OPD loss over all valid response positions in each sampled rollout.

Table 4: Hyperparameter settings for the training and distillation processes.

Parameter Category	Hyperparameter	Value
Model Architecture	Student Model	Qwen3-VL-2B-Instruct
	VL Teacher Model	Qwen3-VL-4B-Instruct
	Text Teacher Model	Qwen3-4B-Instruct-2507
Data Configuration	Max Prompt Length	12,384
	Max Response Length	12,384
	Max Number of Tokens	24,769
	Train Batch Size	128
Distillation Settings	Distillation Loss Mode	`kl_topk`
	Top-K (for KL divergence)	8
	Loss Max Clamp	10.0
	Log Prob Min Clamp	-10.0
	Fusion Mode	`token-level`
	Fusion Temperature (τ)	0.5
Training & Optimization	Training Framework	verl (Sheng et al., 2025)
	Rollout Framework	vLLM (Kwon et al., 2023)
	Learning Rate (Actor)	1e-6
	PPO Mini Batch Size	128
	Student Micro Batch Size per GPU	1
	Total Epochs	15
	Sequence Parallel Size	1
Hardware & Execution	GPU Type	NVIDIA B200
	Total Number of GPUs	8
	GPUs for Student Model	4
	GPUs for VL Teacher Model	2
	GPUs for Text Teacher Model	2
	GPU Memory Utilization	0.8
	FSDP Parameter Offload	True
	FSDP Optimizer Offload	True

Teacher models remain frozen throughout training, and only the student parameters are updated. For reproducibility, Table 4 summarizes the experimental settings for distilling Qwen3-VL-2B-Instruct using Qwen3-VL-4B-Instruct and Qwen3-4B-Instruct-2507 as teacher models.

From a systems perspective, H-OPD is more computationally demanding than single-teacher OPD because each student rollout requires querying two teachers at every decoding step. To improve efficiency, we cache the transferred image descriptions offline before training and reuse them throughout all epochs. This reduces repeated multimodal preprocessing and makes the text teacher branch substantially cheaper during optimization.

C Additional Analyses and Discussion

This appendix provides additional analyses of Heterogeneous On-Policy Distillation (H-OPD), focusing on compute-normalized efficiency, the contribution of token-level teacher arbitration, robustness to imperfect visual descriptions, teacher-scale sensitivity, and the interpretation of teacher specialization. Unless otherwise noted, the student is Qwen3-VL-2B-Instruct, and the heterogeneous teacher pair consists of a vision–language (VL) teacher and a text-only teacher.

Table 5: Best MathVerse-200 accuracy within the first 400 training steps. GPU-hours are computed using eight GPUs.

Method	Best Acc.	Step	Time (h)	GPU-h
GRPO	46.5	285	23.7	190.0
OPD	48.5	339	18.1	145.2
H-OPD	**51.5**	311	21.7	173.6

C.1 Compute-Normalized Efficiency

H-OPD evaluates two teacher distributions on each student-generated trajectory and therefore introduces additional teacher-side forward computation relative to single-teacher OPD. Nevertheless, this overhead should be considered jointly with the rollout cost and convergence rate. In particular, GRPO samples multiple responses per input ($N = 8$ in our experiments) to estimate group-relative advantages, whereas OPD-style methods obtain a dense token-level learning signal from each student trajectory. We therefore compare GRPO, OPD, and H-OPD using wall-clock time and GPU-hours, in addition to training steps and generated tokens.

Best performance within 400 steps. Table 5 reports the best MathVerse-200 validation accuracy reached during the first 400 optimization steps and the compute required to reach that checkpoint. H-OPD attains the highest accuracy, improving over GRPO and OPD by 5.0 and 3.0 percentage points, respectively.

Cost-matched comparison. To avoid favoring methods with different per-step costs, Table 6 compares the best-so-far accuracy under identical wall-clock and GPU-hour budgets. H-OPD is consistently competitive and yields the best result at every reported budget. The advantage is especially pronounced early in training, indicating that heterogeneous token-level supervision improves the utility of each rollout.

Time to target accuracy. We additionally measure the compute required to reach fixed accuracy thresholds (Table 7). H-OPD reaches all shared thresholds with less wall-clock time and fewer GPU-hours than both baselines. GRPO does not reach 48.0%, while only H-OPD reaches 50.0% within the observed training window.

These results complement the token-efficiency analysis in the main paper. They do not imply that

Table 6: Cost-matched MathVerse-200 results. Each entry is the best validation accuracy attained within the corresponding wall-clock and GPU-hour budget.

Compute budget		Best accuracy			H-OPD improvement	
Time (h)	GPU-h	GRPO	OPD	H-OPD	vs. GRPO	vs. OPD
5	40	38.0	42.0	**46.5**	+8.5	+4.5
10	80	38.5	45.0	**47.0**	+8.5	+2.0
15	120	44.0	47.0	**47.0**	+3.0	+0.0
20	160	44.0	48.5	**49.5**	+5.5	+1.0
22	176	44.5	48.5	**51.5**	+7.0	+3.0

Table 7: Wall-clock hours and GPU-hours required to reach a target MathVerse-200 accuracy. A dash denotes that the target was not reached.

Target	GRPO	OPD	H-OPD
40.0	10.83 / 86.67	3.58 / 28.60	**1.40 / 11.20**
45.0	23.75 / 190.00	9.62 / 77.00	**2.45 / 19.60**
48.0	–	18.15 / 145.20	**15.05 / 120.40**
50.0	–	–	**21.70 / 173.60**

Table 8: Controlled arbitration ablations on LogicVista.

Method	Accuracy
Student baseline	35.8
Uniform two-teacher averaging	31.3
Sample-level teacher routing	42.5
Text teacher only + description	26.6
H-OPD (token-level arbitration)	**49.4**

the second teacher is computationally free; rather, they show that its additional forward cost is offset by faster improvement under matched end-to-end compute budgets.

C.2 Mechanism Analysis: Arbitration and Routing

We isolate the source of H-OPD improvements by comparing token-level confidence-aware arbitration with three alternatives: (i) uniform averaging of both teacher distributions at every token, (ii) sample-level routing that assigns one teacher to an entire response, and (iii) supervision from the text teacher alone after conditioning it on a VL-generated image description. All variants use the same Qwen3-VL-2B-Instruct student and the same Qwen3-VL-4B-Instruct and Qwen3-4B-Instruct-2507 teacher pair.

Uniform averaging underperforms the student baseline, demonstrating that the benefit does not follow merely from exposing the student to two teachers. Heterogeneous teachers can assign high probability to incompatible continuations, and an unconditional average may blur both signals. Sample-level routing performs substantially better, confirming that selecting among heterogeneous teachers is useful, but remains 6.9 points below token-level arbitration. This gap supports the central hypothesis that teacher reliability varies within a single multimodal trajectory as decoding alternates between perceptual grounding and symbolic reasoning.

The text-teacher-only variant performs poorly because textualized image evidence cannot replace direct supervision from a VL teacher. The vision-to-language transfer module should therefore be interpreted as an interface that enables a text teacher to participate in multimodal distillation, rather than as a substitute for visual supervision.

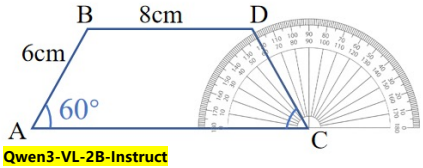
D The Use of Large Language Models (LLMs)

In this paper, large language model (LLM) (Team et al., 2023) is only utilized to assist with the refinement of the English language. No LLMs are employed to generate new innovative ideas, and the research process is conducted by the researchers (humans).

E Case studies

Fig. 7 presents a representative case study comparing the original model, the OPD baseline, and H-OPD on a geometry reasoning example. The original model and OPD both fail on this instance, but their errors stem from the same source: they first misinterpret the visual structure of the trapezoid and then build subsequent reasoning on top of that incorrect perception. Specifically, both models confuse the roles of the parallel sides and non-parallel sides, leading to an incorrect geometric decomposition and ultimately an erroneous final answer. This example shows that once the visual grounding is wrong, downstream reasoning can remain internally fluent yet still be fundamentally invalid.

In contrast, H-OPD correctly identifies the geometric configuration from the diagram and then follows a consistent reasoning process to derive the correct answer. This behavior highlights the benefit of heterogeneous token-level arbitration: the VL teacher contributes more reliable visually grounded signals when precise perception is required, while the text-side reasoning remains coherent after the correct structure has been established. The example therefore illustrates that the gain of H-OPD is not merely better final-answer accuracy, but a more faithful coupling between perception and reasoning. In multimodal reasoning tasks, such improvement is particularly important because reasoning quality is heavily constrained by the correctness of the underlying visual interpretation.



Question: As shown in the diagram, using a protractor to measure the angle, ∠C=60°, then the length of CD in the trapezoid is ( ) cm.

Answer: 6

knowledge concept: Properties and Understanding of Trapezoids

Qwen3-VL-2B-Instruct

We are given a trapezoid $ABCD$ with:\n\n- **$AB = 8$ cm (top base),\n- $AD = 6$ cm (left leg),**n- $BC = 6$ cm (right leg),\n- $\\angle A = 60^\\circ$,\n- A protractor is used to measure $\\angle C = 60^\\circ$,\n- We are to find the length of CD.\n\nWait — let's carefully analyze the diagram and the given information.\n\n---\n\n### Step 1: Understand the trapezoid\n\nFrom the description:\n\n- Trapezoid $ABCD$ with AB and CD as the two parallel sides (top and bottom).\n- Since AB and CD are the bases**, and $AB = 8$ cm, $AD = 6$ cm, $BC = 6$ cm,** and $\\angle A = 60^\\circ$.\n\nBut the key point is: **the angle at A is 60°**, and **the angle at C is also 60°** (as per the protractor measurement).\n\nSo, we have a trapezoid where:\n\n- $AB \\parallel CD$ (assuming standard labeling),\n- $\\angle A = 60^\\circ$,\n- $\\angle C = 60^\\circ$,\n- We are to find the length of CD.\n\n---\n\n### Step 2:

… (Omitted for brevity.)

But since the trapezoid is symmetric, the horizontal projection of BC is 3 cm to the left, and the height is $3\\sqrt{3}$ cm.\n\nSo, from point C, the horizontal distance to the right is 3 cm (since the projection is to the left), and the vertical height is $3\\sqrt{3}$ cm.\n\nSo, the angle between BC and the base CD is the angle between the leg and the base — which is $ \\tan^{-1} \\left(\\frac{h}{\\text{horizontal projection}} \\right) = \\tan^{-1} \\left(\\frac{3\\sqrt{3}}{3} \\right) = \\tan^{-1}(\\sqrt{3}) = 60^\\circ $ — correct.\n\nSo, all checks out.\n\n---\n\n### ✅ **Final Answer:\n\n$$\n\\boxed{14}\n$$\n\nSo, the length of CD is $\\boxed{14}$ cm.**

incorrect perception, incorrect reasoning

Qwen3-VL-2B-Instruct + OPD

We are given a trapezoid $ABCD$ with the following information:\n\n- **$AB = 8$ cm (top base)**n- AD is the bottom base\n- AB is parallel to CD (since it's a trapezoid)\n- AC is a diagonal\n- $∠A = 60^\\circ$\n- $∠C = 60^\\circ$\n- A protractor is used to measure $∠C = 60^\\circ$\n- We are to find the length of CD\n\nLet's analyze the diagram and use geometric reasoning.\n\n---\n\n### Step 1: Understand the Trapezoid\nSince $AB \\parallel CD$, and AD is a leg, we can assume the trapezoid is **trapezoid $ABCD$** with AB and CD as the two parallel sides (bases).

… (Omitted for brevity.)

n\nSo $CD = 8$ cm\n\nTherefore, the length of CD is 8 cm.\n\nBut wait — is AC the same as AB?\n\nIn triangle $ABC**$, $AB = 8$ cm, and $AC = 8$ cm, so $AB = AC$, and** $∠A = 60^\\circ$, so triangle ABC is equilateral.\n\nSo $BC = 8$ cm, $AC = 8$ cm, $AB = 8$ cm.\n\nNow, in triangle ACD, $AC = 8$ cm, $∠A = 60^\\circ$, $∠C = 60^\\circ$, so $CD = 8$ cm.\n\nSo the length of CD is 8 cm.\n\n---\n\n**### Final Answer:\n$$\n\\boxed{8}\n$$.**

incorrect perception, incorrect reasoning

Qwen3-VL-2B-Instruct +H-OPD(Ours)

"We are given a trapezoid $ABCD$ with:\n\n- $AB = 6\\text{ cm}$ (the non-parallel side),\n- $BD = 8\\text{ cm}$ (the top base, which is parallel to AC),\n- $\\angle A = 60^\\circ$,\n- $\\angle C = 60^\\circ$ (measured with a protractor),\n- We are to find the length of CD.\n\n---\n\n### Step 1: Understand the figure\n\nFrom the diagram, it appears that AB and CD are the **non-parallel sides** (legs), and AD and BC are the **bases** — but wait, the diagram shows:\n\n- $BD = 8\\text{ cm}$

… (Omitted for brevity.)

\n\nNow, is this consistent? Yes — because the top base $BD = 8\\text{ cm}$, and the legs project 3 cm to each side, so the bottom base is 8 + 3 + 3 = 14 cm.\n\nSo the trapezoid is symmetric, with legs $AB = CD = 6\\text{ cm}$, and angles at A and C both 60°.\n\nTherefore, the length of CD is **6 cm**.\n\n---\n\n### ✅ Final Answer: $\\boxed{6}$ cm"

correct perception, correct reasoning

Figure 7: Case studies.