CONTRASTIVE ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student model on trajectories sampled from its own policy, providing dense token-level supervision by minimizing the divergence between the teacher's and student's output distributions at each token. Although existing OPD approaches effectively distill strong reasoning capabilities into student models, they inherently inherit uncurated chain-of-thought traces, exacerbating overthinking and reasoning redundancy. To address this limitation, we introduce COPD, a contrastive OPD framework. Specifically, for each token generated by the student, a frozen teacher evaluates the current student state under two contrasting prompts that induce low and high reasoning effort, respectively. The resulting difference in log-probabilities serves as a token-level advantage signal to guide the policy update. Rather than merely mimicking a single teacher distribution, COPD guides the student model toward acquiring concise and efficient reasoning strategies. We evaluate COPD across 9 multimodal benchmarks covering both reasoning and understanding tasks. Empirical results show that COPD substantially reduces reasoning length without hurting task performance, consistently improving efficiency across different tasks and model scales. In addition, this contrastive paradigm extends naturally to on-policy self-distillation (OPSD), establishing self-contrastive supervisory signals that enable a single model to compress its own reasoning without an external teacher.

1 INTRODUCTION

Large vision-language models (LVLMs) have made substantial progress in multimodal tasks (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; 2024; Chen et al., 2024b; Bai et al., 2023; Wang et al., 2024; Bai et al., 2025b;a; Caffagni et al., 2024; Jin et al., 2025a). By decomposing complex problems into explicit intermediate steps, these models steadily improve their performance in tasks such as image understanding and vision-language reasoning (Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2022; Zhang et al., 2023). Consequently, step-by-step reasoning has become the primary paradigm for complex problem solving (Wei et al., 2022; Yao et al., 2023; Huang & Chang, 2023). However, stronger reasoning capabilities often introduce severe redundancy into the reasoning chains. A central challenge is therefore to improve reasoning efficiency without sacrificing performance.

Existing reasoning-compression methods predominantly rely on supervised fine-tuning (SFT) or reinforcement learning with verifiable rewards (RLVR) (Munkhbat et al., 2025; Jin et al., 2025b; Han et al., 2025; Li et al., 2026c; Liu et al., 2025; Li et al., 2026a; Dumitru et al., 2025). However, they either demand scarce, curated concise traces or require ground-truth verifiers, limiting their applicability to unlabeled scenarios. On-policy distillation (OPD) provides a label-free alternative by distilling teacher feedback along student-generated sequences (Agarwal et al., 2024; Lu & Lab, 2025). Yet, standard OPD aligns the student with only a single teacher distribution, failing to capture how token compatibility shifts across different reasoning efforts.

To this end, we introduce COPD, a contrastive on-policy distillation framework. By pairing light- and heavy-thinking (LT/HT) contexts, this approach converts the implicit preferences of the teacher into explicit token-level supervisory signals. Specifically, the student model first generates reasoning trajectories on-policy. A frozen teacher then evaluates this trajectory at the token level under two contrasting prompts specifying low and high reasoning effort, respectively. The difference between the resulting log-likelihoods serves as a token-level advantage signal to guide student optimization. Under this design, the teacher functions not merely as an imitation target for a single distribution, but

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

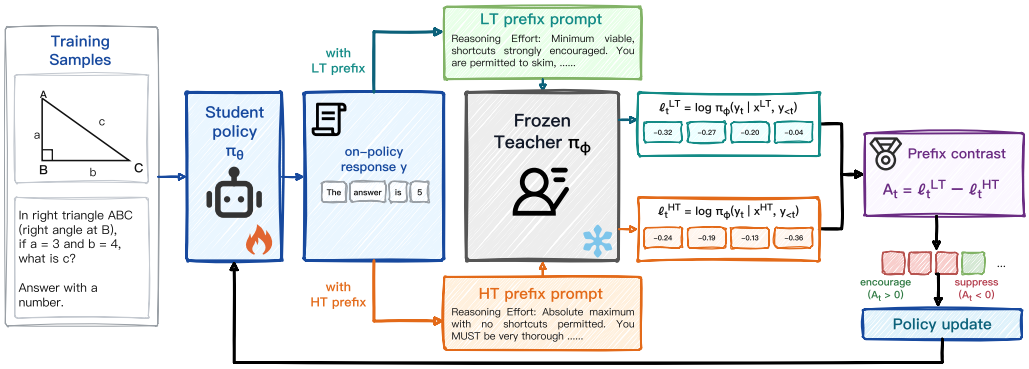


Figure 1: **Overview of the COPD training pipeline.** The student model first samples a response trajectory given the input. A frozen teacher then evaluates the generated sequence token by token under contrasting light- and heavy-thinking (LT and HT) prompts. The resulting difference in log-probabilities serves as a token-level advantage signal to guide the student policy update.

as a contrastive reference across different reasoning efforts. Consequently, this contrastive signal biases the student toward concise outputs aligned with the light-thinking condition, enabling adaptive reasoning compression without predefined length budgets.

We systematically evaluate COPD across 9 multimodal benchmarks. Empirical results show that across diverse task types and model scales, COPD substantially shortens reasoning length while maintaining or improving task accuracy, thereby optimizing the trade-off between performance and inference cost. For instance, with Qwen3-VL-8B as the teacher and Qwen3-VL-2B (Bai et al., 2025a) as the student, COPD improves accuracy from 55.2% to 57.0% compared to ExOPD (Yang et al., 2026), while cutting response length by 56.3%. In addition, this contrastive mechanism extends naturally to on-policy self-distillation (OPSD), constructing self-contrastive supervisory signals without an external teacher to compress reasoning lengths effectively.

The main contributions of this paper are summarized as follows: **1)** We introduce COPD, a contrastive on-policy distillation framework that derives dense token-level advantage signals from paired teacher evaluations under contrasting light- and heavy-thinking instructions, achieving adaptive reasoning compression with negligible performance loss. **2)** We extend this contrastive paradigm to on-policy self-distillation (OPSD), enabling models to self-induce concise and efficient reasoning strategies entirely free of external supervision. **3)** Extensive experiments across 9 multimodal benchmarks demonstrate the consistent effectiveness and robust generalization of COPD across diverse tasks and model scales.

2 COPD: Contrastive On-Policy Distillation

2.1 Problem Setup and Overview

Let x denote an input prompt, π_θ the student policy, and π_ϕ a frozen teacher model. On-policy distillation samples a response $y = (y_1, \ldots, y_{T_y})$ of length $T_y = |y|$ from the student policy $\pi_{\theta_{\text{old}}}$ and obtains teacher supervision at each student-visited prefix $(x, y_{<t})$ (Agarwal et al., 2024; Lu & Lab, 2025). While this mechanism provides dense supervision along student-generated trajectories, its training signal is derived from a single teacher conditional distribution at each visited prefix.

As shown in Figure 1, COPD casts teacher scoring as a paired comparison across different reasoning efforts. Specifically, given a student-generated trajectory, the frozen teacher scores each token conditioned on the shared multimodal input and the preceding prefix, differing solely in the light- versus heavy-thinking prompt. The difference between the resulting log-likelihoods constitutes a token-level preference signal. Consequently, COPD preserves the student-induced state distribution inherent to OPD, while replacing single-context imitation with a dense estimate of whether the current student action aligns more closely with concise or exhaustive reasoning.

2.2 PAIRED TEACHER SCORING

Let ι^{LT} and ι^{HT} denote light- and heavy-thinking instructions, respectively. Given input x, the student policy first samples a complete response trajectory:

$$y = (y_1, \ldots, y_{T_y}) \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x), \tag{1}$$

Subsequently, the frozen teacher receives the same input and evaluates the identical trajectory y, computing token-level conditional log-probabilities under both light- and heavy-thinking instructions:

$$\ell_t^m = \log \pi_\phi(y_t \mid x^m, y_{<t}), \qquad x^m = [\iota^m; x], \quad m \in \{\mathrm{LT}, \mathrm{HT}\}, \quad t = 1, \ldots, T_y. \tag{2}$$

Intuitively, the same token can exhibit markedly different conditional likelihoods depending on the intended reasoning effort (Han et al., 2025; Shi et al., 2024). Paired scoring builds directly on this observation: through two forward passes of the frozen teacher, it evaluates the identical generation step under contrasting instruction contexts, yielding the relative likelihood of the token across the two regimes.

2.3 CONTRASTIVE TOKEN-LEVEL ADVANTAGE

COPD defines the token-level advantage as the log-likelihood contrast between light- and heavy-thinking contexts:

$$A_t = \ell_t^{\mathrm{LT}} - \ell_t^{\mathrm{HT}} = \log \pi_\phi(y_t \mid x^{\mathrm{LT}}, y_{<t}) - \log \pi_\phi(y_t \mid x^{\mathrm{HT}}, y_{<t}). \tag{3}$$

A positive value ($A_t > 0$) indicates that, conditioned on the shared student's prefix, the teacher considers y_t more aligned with the light-thinking context, whereas a negative value indicates the converse. To stabilize policy updates, we apply two-sided clipping to bound the advantage scale. Formally, the effective advantage is:

$$\tilde{A}_t = \mathrm{stopgrad}(\mathrm{clip}(A_t, -A_{\max}, A_{\max})), \qquad A_{\max} > 0. \tag{4}$$

Importantly, this advantage signal incorporates no explicit length penalty. Even tokens appearing late in the sequence can receive high positive advantages, provided the teacher determines that they effectively advance concise reasoning. Consequently, reasoning compression emerges as a natural outcome of policy optimization: training reinforces tokens that make progress, summarize, or directly deliver the answer, while penalizing those that prolong redundant heavy-thinking.

2.4 POLICY UPDATE

Given the clipped contrastive advantage $\tilde{A}_t$, COPD optimizes the student policy via a clipped surrogate objective (Schulman et al., 2017). Let

$$r_t(\theta) = \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\theta_{\mathrm{old}}}(y_t \mid x, y_{<t})}$$

denote the probability ratio between the current student policy and the old student policy used to generate the sampled responses. The clipped objective is formulated as:

$$J_{\mathrm{COPD}}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D} \\ y \sim \pi_{\theta_{\mathrm{old}}}(\cdot|x)}} \left[\frac{1}{T_y} \sum_{t=1}^{T_y} \min \left(r_t(\theta)\tilde{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}})\, \tilde{A}_t \right) \right]. \tag{5}$$

This objective increases the likelihood of tokens aligned with the light-thinking context while penalizing those favored under the heavy-thinking context. In this formulation, the contrastive advantage $\tilde{A}_t$ functions as a scalar weight that modulates the policy gradient, directly steering the optimization of the student parameters θ according to the teacher's relative preference. The overall training procedure is summarized in Algorithm 1.

2.5 SELF-DISTILLATION VARIANT

This paired contrastive mechanism also applies without relying on an external teacher. We denote this variant as COPSD, where the frozen reference policy is instantiated as a static snapshot of the student model's initial weights. Beyond its practical utility, COPSD allows us to investigate whether the compression signal inherently hinges on teacher scale. If a single base model assigns distinct likelihoods to the identical action under light- versus heavy-thinking contexts, the supervisory signal is driven by context-induced relative preferences rather than external teacher capacity alone.

Algorithm 1 COPD training procedure

Require: Initial student π_{θ_0}, frozen teacher π_ϕ, training distribution $\mathcal{D}$, batch size B, training steps K, advantage bound $A_{\max}$, PPO clipping bounds ϵ_{low} and ϵ_{high}

for $k = 0, \ldots, K-1$ **do**
  $\theta_{\text{old}} \leftarrow \theta_k$
  Sample $\mathcal{B}_k = (x_i)_{i=1}^{B}$, where $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{D}$
  for $i = 1, \ldots, B$ **do**
    Sample $y_i = (y_{i,1}, \ldots, y_{i,T_i}) \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i)$
    Construct $x_i^m = [\iota^m; x_i]$ for $m \in \{\text{LT}, \text{HT}\}$
    for $t = 1, \ldots, T_i$ **do**
      Compute $\ell_{i,t}^m = \log \pi_\phi(y_{i,t} \mid x_i^m, y_{i,<t})$ for $m \in \{\text{LT}, \text{HT}\}$
      $A_{i,t} \leftarrow \ell_{i,t}^{\text{LT}} - \ell_{i,t}^{\text{HT}}$
      $\tilde{A}_{i,t} \leftarrow \text{stopgrad}(\text{clip}(A_{i,t}, -A_{\max}, A_{\max}))$
    end for
  end for
  Update $\theta_{k+1} \approx \arg\max_\theta J_{\text{COPD}}(\theta)$ using $\mathcal{B}_k$ and the clipped ratios $[1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}]$
end for

3 Experiments

3.1 Setup

3.1.1 Models and Baseline

We conduct our primary experiments on Qwen3-VL-2B-Instruct and Qwen3.5-2B-Base (Bai et al., 2025a; Team, 2026). For training, we select approximately 16K instances from ViRL-39K (Wang et al., 2026), denoted as ViRL-16K[1]. We benchmark our method against the respective base models, Vanilla OPD (Lu & Lab, 2025), and two competitive OPD variants: ExOPD (Yang et al., 2026) and VA-OPD (Liu et al., 2026). Specifically, ExOPD formulates online distillation as reinforcement learning under dense KL constraints, applying reward extrapolation to push student performance beyond the teacher's ceiling. VA-OPD measures the teacher's token-level log-probability gap between inputs with versus without visual details, hierarchically reweighting rollouts and tokens during distillation to prevent visual information from being diluted by linguistic tokens.

3.1.2 Training Details

We implement our training pipeline within `verl`, using FSDP and vLLM for accelerated on-policy rollout generation and teacher scoring (Sheng et al., 2025; Zhao et al., 2023; Kwon et al., 2023). For each on-policy response, the frozen teacher evaluates generated tokens under contrasting light- and heavy-thinking prompts, with the resulting advantage $\ell_t^{\text{LT}} - \ell_t^{\text{HT}}$ clipped to $[-10, 10]$. Crucially, we incorporate no external task rewards, KL divergence penalties, or explicit length regularizers. Models are trained for 3 epochs using the AdamW optimizer. Complete optimization hyperparameters and sequence-length configurations are provided in Appendix A.1.

3.1.3 Evaluation Benchmarks

We evaluate on nine multimodal benchmarks, spanning four reasoning tasks (MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), WeMath (Qiao et al., 2025), LogicVista (Xiao et al., 2024)) and five understanding tasks (HalluBench (Guan et al., 2024), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024a), RWQA (xAI, 2024), MMGist (Yuan et al., 2026b))[2]. Table entries are formatted as Accuracy / Average Tokens / Acc@1K, where accuracy is reported as mean$_{\pm\text{std}}$ over three independent runs, and response length denotes the mean. To quantify efficiency, Acc@1K measures

[1] ViRL-39K reports empirical pass rates estimated via Qwen2.5-VL-32B (Bai et al., 2025b). To focus on non-trivial yet solvable instances, we retain only samples with pass rates strictly between 0 and 1. We then split the resulting 18K samples at a 9:1 ratio, yielding ViRL-16K for training and 2K for validation.

[2] Benchmark details and evaluation configurations are provided in the Appendix A.2 and A.3.

Table 1: **Main results of COPD.** Each benchmark cell reports Accuracy / Average Tokens; AVG reports Accuracy / Average Tokens / Acc@1K, with the standard deviation of accuracy shown as a subscript. The best accuracy, the shortest token length and the best Acc@1K are in bold.

Method	Reasoning				Understanding					AVG
	MathVista	MathVerse	WeMath	LogicVista	HalluBench	MMMU	MMStar	RWQA	MMGist	
Q3VL-2B	61.1/1.22K	47.9/2.13K	58.9/3.54K	38.7/4.45K	69.9/0.52K	45.6/3.08K	58.7/1.19K	67.5/0.17K	27.7/2.79K	$52.9_{\pm 0.4}$/2.12K/25.0
Qwen3-VL-8B→Qwen3-VL-2B (Avg Performance of Qwen3-VL-8B-Instruct with LT prompt: 62.9/0.55K/114.4)										
OPD	61.1/1.30K	49.7/2.55K	59.8/3.74K	40.0/4.49K	70.4/0.58K	46.2/3.14K	59.4/1.25K	**67.0**/0.21K	28.0/2.86K	$53.5_{\pm 0.1}$/2.23K/24.0
ExOPD	**66.8**/1.34K	52.2/2.92K	64.4/2.99K	41.6/3.70K	**71.2**/0.50K	47.0/2.95K	59.5/1.27K	66.3/0.31K	27.9/2.79K	$55.2_{\pm 0.1}$/2.08K/26.5
VA-OPD	61.5/1.33K	48.8/2.46K	60.5/3.59K	40.5/4.26K	69.8/0.56K	46.8/3.06K	58.1/1.24K	66.7/0.20K	27.5/2.82K	$53.4_{\pm 0.1}$/2.17K/24.6
COPD	62.9/**0.42K**	**53.0/0.73K**	**69.5/1.36K**	**48.4/1.89K**	68.1/**0.15K**	**52.6/1.67K**	**60.3/0.47K**	**67.0/0.03K**	**31.2/1.42K**	**57.0**$_{\pm 0.2}$**/0.91K/62.6**
Qwen3-VL-32B→Qwen3-VL-2B (Avg Performance of Qwen3-VL-32B-Instruct with LT prompt: 72.0/0.69K/104.3)										
OPD	63.4/1.34K	50.8/2.92K	62.1/3.77K	42.1/4.38K	**70.4**/0.63K	47.3/3.07K	60.5/1.25K	67.5/0.25K	28.9/2.92K	$54.8_{\pm 0.3}$/2.28K/24.0
ExOPD	**66.9**/1.20K	53.9/2.68K	65.7/2.82K	44.4/3.65K	67.9/0.35K	48.9/2.66K	**63.6**/0.99K	68.1/0.28K	30.3/2.56K	$56.6_{\pm 0.2}$/1.91K/29.6
VA-OPD	63.7/1.34K	50.2/2.87K	63.0/3.57K	42.9/4.19K	69.9/0.58K	47.7/2.97K	60.4/1.24K	**68.5**/0.27K	28.8/2.85K	$55.0_{\pm 0.2}$/2.21K/24.9
COPD	62.9/**0.39K**	**54.0/0.71K**	**71.6/1.16K**	**49.4/1.51K**	66.6/**0.15K**	**53.2/1.44K**	61.0/**0.41K**	64.9/**0.03K**	**31.1/1.23K**	**57.2**$_{\pm 0.1}$**/0.78K/73.3**
Q35-2B	68.8/2.25K	55.3/3.92K	60.9/4.18K	43.2/4.23K	69.9/0.99K	48.5/3.54K	62.4/1.90K	69.1/1.05K	30.3/3.47K	$56.5_{\pm 0.1}$/2.84K/19.9
Qwen3.5-4B→Qwen3.5-2B (Avg Performance of Qwen3.5-4B with LT prompt: 72.5/2.67K/27.2)										
OPD	70.2/2.95K	57.5/4.55K	62.3/5.09K	40.1/5.63K	70.4/1.73K	49.7/4.31K	62.2/2.48K	67.7/1.81K	29.4/4.47K	$56.6_{\pm 0.4}$/3.67K/15.4
ExOPD	69.8/3.08K	58.6/4.56K	62.6/5.14K	41.3/5.58K	71.7/1.73K	49.9/4.24K	63.6/2.53K	**68.7**/1.74K	29.0/4.48K	$57.3_{\pm 0.5}$/3.68K/15.6
VA-OPD	68.0/3.10K	56.9/4.60K	61.2/5.19K	41.0/5.45K	70.9/1.76K	49.2/4.28K	62.7/2.56K	68.3/1.74K	28.3/4.46K	$56.3_{\pm 0.3}$/3.68K/15.3
COPD	**71.2/1.49K**	**60.9/2.37K**	**68.9/2.64K**	**49.3/3.06K**	**71.9/0.63K**	**54.4/2.65K**	**64.9/1.29K**	68.3/**0.68K**	**31.7/2.57K**	**60.2**$_{\pm 0.1}$**/1.93K/31.2**
Qwen3.5-35A3B→Qwen3.5-2B (Avg Performance of Qwen3.5-35A3B with LT prompt: 78.5/2.58K/30.4)										
OPD	69.2/3.15K	58.6/4.62K	62.7/5.22K	39.5/5.69K	**71.5**/1.86K	50.1/4.44K	62.5/2.68K	65.7/2.01K	28.7/4.67K	$56.5_{\pm 0.1}$/3.82K/14.8
ExOPD	69.7/3.14K	58.8/4.60K	63.8/5.21K	39.9/5.56K	70.3/1.82K	50.2/4.35K	63.8/2.63K	67.1/1.94K	28.0/4.63K	$56.8_{\pm 0.1}$/3.76K/15.1
VA-OPD	69.2/3.21K	57.7/4.60K	62.0/5.24K	42.8/5.54K	71.2/1.81K	50.1/4.39K	62.6/2.62K	67.1/1.85K	27.7/4.60K	$56.7_{\pm 0.2}$/3.76K/15.1
COPD	**70.8/1.59K**	**60.1/2.54K**	**68.0/2.86K**	**48.4/3.24K**	71.1/**0.62K**	**52.9/2.85K**	**64.6/1.38K**	**69.5/0.70K**	**31.3/2.73K**	**59.6**$_{\pm 0.4}$**/2.06K/28.9**

the accuracy per thousand output tokens across the benchmark set $\mathcal{B}$:

$$\mathrm{Acc@1K} = 1000 \cdot \frac{\sum_{b \in \mathcal{B}} \mathrm{Acc}_b}{\sum_{b \in \mathcal{B}} \mathrm{Tokens}_b}, \tag{6}$$

where Acc_b and Tokens_b denote the accuracy and average response length on benchmark b.

3.2 Main Results

Accuracy–length Pareto frontier. Table 1 demonstrates that COPD establishes a superior accuracy–length Pareto frontier across all four teacher–student configurations, achieving the highest average accuracy alongside the shortest response lengths. For Qwen3-VL-8B→2B, relative to the strong baseline ExOPD, COPD improves accuracy by 1.8 pp while cutting response length by 56.3%, thereby increasing Acc@1K by 136.2%. This advantage consistently holds across the remaining model and teacher pairings, where COPD outperforms ExOPD with a 91.4%–147.6% boost in Acc@1K. These results confirm that COPD steers the student policy toward concise reasoning behaviors while achieving better performance.

Broad compression with task-level heterogeneity. Under the Qwen3-VL-8B→2B configuration, relative to the un-distilled Qwen3-VL-2B base model, COPD reduces generation length across all nine benchmarks while improving accuracy on seven. The most pronounced gains appear on complex reasoning benchmarks, namely WeMath (+10.6 pp accuracy with 61.6% fewer tokens) and LogicVista (+9.7 pp accuracy with 57.5% fewer tokens), demonstrating that aggressive compression does not compromise essential reasoning steps on tasks with inherently verbose initial traces. In contrast, on HalluBench and RWQA where initial responses are already brief, accuracy decreases slightly by 1.8 pp and 0.5 pp alongside substantial token reductions of 71.2% and 82.4%, respectively. This pattern indicates that while the contrastive advantage effectively eliminates redundant rationales, it delineates an operational boundary: aggressive compression provides a narrower safety margin when tasks inherently elicit concise responses.

Gains do not follow monotonic teacher scaling. Scaling the Qwen3-VL teacher from 8B to 32B improves average accuracy by 0.2 pp, reduces response length by 14.3%, and boosts Acc@1K by 17.1%. Conversely, scaling the Qwen3.5 teacher from 4B to 35A3B decreases accuracy by 0.6 pp, lengthens responses by 6.7%, and lowers Acc@1K by 7.4%. Consequently, increasing teacher scale shifts the trade-off frontier without reliably optimizing it. This non-monotonicity is consistent with

Table 2: **Generalization across various model scales.**

Method	Reasoning	Understanding	AVG
Qwen3.5-4B	$72.2_{\pm 0.1}$/2.87K/25.2	$64.9_{\pm 0.1}$/1.85K/35.1	$68.1_{\pm 0.1}$/2.30K/29.6
Qwen3.5-4B $\xleftarrow{\text{COPD}}$ Qwen3.5-35A3B	$\mathbf{73.9}_{\pm 0.2}$**/1.52K/48.6**	$\mathbf{65.2}_{\pm 0.5}$**/1.22K/53.4**	$\mathbf{69.1}_{\pm 0.2}$**/1.35K/51.2**
Qwen3.5-9B	$75.8_{\pm 0.4}$/2.73K/27.8	$67.6_{\pm 0.4}$/1.85K/36.5	$71.2_{\pm 0.2}$/2.24K/31.8
Qwen3.5-9B $\xleftarrow{\text{COPD}}$ Qwen3.5-35A3B	$76.8_{\pm 0.3}$**/1.52K/50.5**	$68.1_{\pm 0.1}$**/1.29K/52.8**	$72.0_{\pm 0.1}$**/1.40K/51.4**
Qwen3.5-9B $\xleftarrow{\text{COPD}}$ Qwen3.5-122A10B	$\mathbf{78.6}_{\pm 0.4}$/1.99K/39.5	$\mathbf{68.8}_{\pm 0.5}$/1.56K/44.1	$\mathbf{73.2}_{\pm 0.4}$/1.75K/41.8

Table 3: **Generalization across training data under the Qwen3-VL-2B $\xleftarrow{\text{COPD}}$ Qwen3-VL-8B setting.** Geo3K (24 epochs) is iteration-matched to ViRL-16K (3 epochs).

Dataset	Epoch	Reasoning	Understanding	AVG
-	-	$51.6_{\pm 0.6}$/2.84K/18.2	$53.9_{\pm 0.2}$/1.55K/34.8	$52.9_{\pm 0.4}$/2.12K/25.0
Geo3K	3	$54.3_{\pm 0.1}$/2.11K/25.7	$54.4_{\pm 0.1}$/1.25K/43.5	$54.3_{\pm 0.1}$/1.63K/33.3
Geo3K	24	$58.1_{\pm 0.3}$/1.24K/46.9	$55.2_{\pm 0.1}$/0.83K/66.5	$56.5_{\pm 0.1}$/1.01K/55.9
ViRL-16K	3	$\mathbf{58.5}_{\pm 0.3}$**/1.10K/53.2**	$\mathbf{55.8}_{\pm 0.2}$**/0.75K/74.4**	$\mathbf{57.0}_{\pm 0.2}$**/0.91K/62.6**

recent findings in on-policy distillation: simply deploying a larger or higher-scoring teacher does not necessarily yield more effective token-level supervisory signals (Li et al., 2026b). Nevertheless, COPD achieves the highest Acc@1K and shortest average responses across all four teacher–student configurations, confirming that our consistent gains stem primarily from the contrastive preferences extracted from the teacher, rather than teacher capacity alone.

3.3 Ablation Studies

In this section, we systematically conduct ablation studies across multiple dimensions. Unless noted otherwise, all experiments adhere to the same training and evaluation configurations as the main experiments (see Appendices A.4–A.11 for more experimental results).

3.3.1 Generalization on Scales

As shown in Table 2, the contrastive signal generalizes effectively across different student scales while maintaining its compression efficacy. Holding the 35A3B teacher fixed, COPD shortens average response length by 41.3% and 37.5%, improves average accuracy by 1.0 pp and 0.8 pp, and boosts Acc@1K by 73.0% and 61.6% on the 4B and 9B students, respectively. Conversely, fixing the 9B student and scaling the teacher from 35A3B to 122A10B yields an additional 1.2 pp in accuracy, but at the cost of a 25.0% increase in response length, which reduces Acc@1K by 18.7%. These findings indicate that teacher scaling primarily shifts the accuracy–efficiency trade-off, whereas the efficiency benefits of COPD persist consistently across all evaluated student scales.

3.3.2 Generalization Across Training Data

Table 3 compares COPD trained on Geo3K (Lu et al., 2021) against ViRL-16K. Relative to the baseline, Geo3K (3 epochs) increases average accuracy from 52.9% to 54.3% while reducing average response length by 23.1%. Extending Geo3K training to 24 epochs yields 56.5% accuracy at 1.01K tokens, elevating Acc@1K to 55.9. In comparison, ViRL-16K (3 epochs) achieves 57.0% accuracy at 0.91K tokens, attaining the highest Acc@1K of 62.6. These results demonstrate that COPD does not depend on a specialized data distribution, consistently improving the accuracy–length trade-off across both domain-specific and general multimodal corpora.

3.3.3 Generalization on Various Prompts

As shown in Table 4, COPD is robust to prompt phrasing: the Simple, Complex, and default templates achieve comparable accuracy (56.2%–57.0%), response length (0.82–0.92K tokens), and Acc@1K (61.5–68.5), all consistently outperforming the base model. This consistency confirms that the gains stem from the contrastive formulation itself rather than delicate prompt engineering.

Table 4: **Robustness on various LT/HT prompts using Qwen3-VL-8B as the teacher.**

Method	Reasoning	Understanding	AVG
Qwen3-VL-2B	$51.6_{\pm 0.6}$/2.84K/18.2	$53.9_{\pm 0.2}$/1.55K/34.8	$52.9_{\pm 0.4}$/2.12K/25.0
Simple	$58.2_{\pm 0.5}$/1.12K/52.0	$55.3_{\pm 0.3}$/0.76K/72.8	$56.6_{\pm 0.3}$/0.92K/61.5
Complex	$57.6_{\pm 0.5}$**/0.97K/59.4**	$55.0_{\pm 0.2}$**/0.70K/78.6**	$56.2_{\pm 0.3}$**/0.82K/68.5**
COPD (Ours)	$\mathbf{58.5}_{\pm 0.3}$/1.10K/53.2	$\mathbf{55.8}_{\pm 0.2}$/0.75K/74.4	$\mathbf{57.0}_{\pm 0.2}$/0.91K/62.6

Table 5: **Contrastive on-policy self-distillation.** COPSD uses a frozen copy of the initial student model as the paired LT–HT scorer, eliminating the need for an external teacher.

Method	Reasoning	Understanding	AVG
Qwen3-VL-2B	$51.6_{\pm 0.6}$/2.84K/18.2	$53.9_{\pm 0.2}$/1.55K/34.8	$52.9_{\pm 0.4}$/2.12K/25.0
+COPSD	$\mathbf{58.3}_{\pm 0.5}$**/0.92K/63.4**	$\mathbf{55.4}_{\pm 0.2}$**/0.64K/86.6**	$\mathbf{56.7}_{\pm 0.3}$**/0.76K/74.6**
Qwen3-VL-4B	$\mathbf{69.3}_{\pm 0.4}$/1.56K/44.4	$63.0_{\pm 0.2}$/1.03K/61.2	$65.8_{\pm 0.2}$/1.26K/52.2
+COPSD	$69.0_{\pm 0.3}$**/0.65K/106.2**	$\mathbf{63.4}_{\pm 0.1}$**/0.53K/119.6**	$\mathbf{65.9}_{\pm 0.1}$**/0.58K/113.6**
Qwen3-VL-8B	$\mathbf{71.0}_{\pm 0.6}$/1.41K/50.4	$64.4_{\pm 0.2}$/0.92K/70.0	$\mathbf{67.3}_{\pm 0.4}$/1.14K/59.0
+COPSD	$70.8_{\pm 0.0}$**/0.57K/124.2**	$\mathbf{64.5}_{\pm 0.0}$**/0.48K/134.4**	$\mathbf{67.3}_{\pm 0.0}$**/0.52K/129.4**
Qwen3.5-122A10B	$84.4_{\pm 0.1}$/3.81K/22.2	$75.1_{\pm 0.3}$/3.10K/24.2	$79.2_{\pm 0.3}$/3.42K/23.2
+COPSD	$\mathbf{86.1}_{\pm 0.1}$**/3.14K/27.4**	$\mathbf{76.3}_{\pm 0.1}$**/2.55K/29.9**	$\mathbf{80.7}_{\pm 0.0}$**/2.81K/28.7**

3.3.4 Contrastive On-Policy Self-Distillation

While larger external teachers provide benefits, they are not strictly necessary for acquiring concise reasoning behaviors. Results across four model scales in Table 5 demonstrate that a frozen snapshot of the student itself suffices to construct effective contrastive signals. For the 2B model, COPSD improves average accuracy by 3.8 pp while reducing response length by 64.2%, resulting in a 198.4% boost in Acc@1K. The same trend consistently emerges across the 4B, 8B, and 122A10B scales: average accuracy is maintained or improved, response length shrinks by 17.8%–54.4%, and Acc@1K increases by 23.7%–119.3%. These findings indicate that effective compression does not strictly hinge on teacher scale. Instead, the model's own relative preference between light- and heavy-thinking contexts provides sufficient supervision to steer generation toward concise outputs.

The gains of COPSD reflect a more effective use of the model's existing reasoning capabilities through policy adaptation. Prior work demonstrates that language models possess a latent capacity for concise reasoning, which can be elicited through self-training on their own generated trajectories while preserving accuracy (Munkhbat et al., 2025). Moreover, reasoning accuracy does not increase monotonically with chain-of-thought length: extending reasoning beyond an appropriate length can reduce performance (Wu et al., 2026). Analyses of reasoning traces further identify redundant repetition, excessive self-verification, and over-exploration of alternatives as common sources of unnecessary computation (Yi et al., 2026). In COPSD, the LT and HT scoring contexts expose different conditional preferences within the same frozen model. Optimizing their likelihood contrast reinforces student-generated continuations favored by the light-thinking context, shifting the generation policy toward more concise responses.

3.3.5 Can a single light-thinking teacher foster light-thinking in students?

The critical factor driving compression is the contrast between LT and HT prompts, rather than the LT score alone. Results in Table 6 illustrate this distinction: optimizing solely for the LT score lengthens outputs, whereas the contrastive objective improves both accuracy and efficiency. Specifically, adopting a unilateral objective $A_t = \ell_t^{\mathrm{LT}}$ yields only a marginal gain in average accuracy (52.9% $\rightarrow$ 53.2%) while expanding average response length from 2.12K to 2.25K tokens. These results indicate that conditional likelihood under LT alone fails to establish a definitive criterion for determining which reasoning steps to streamline. In contrast, the formulation $A_t = \ell_t^{\mathrm{LT}} - \ell_t^{\mathrm{HT}}$ evaluates each token under both contexts, converting teacher evaluations into a directional signal actively favoring concise reasoning. This contrastive objective elevates average accuracy to 57.0% while slashing average response length by 57.1%. These findings confirm that the HT score pro-

Table 6: **Ablation of the contrastive teacher signal with Qwen3-VL-8B as teacher.** Using only the light-thinking log-probability does not yield output-efficient reasoning, whereas the LT–HT contrast improves accuracy while reducing response length.

Method	Reasoning	Understanding	AVG
Qwen3-VL-2B	$51.6_{\pm 0.6}$/2.84K/18.2	$53.9_{\pm 0.2}$/1.55K/34.8	$52.9_{\pm 0.4}$/2.12K/25.0
$A_t = \ell_t^{\mathrm{LT}}$	$52.4_{\pm 0.6}$/3.05K/17.2	$53.9_{\pm 0.4}$/1.62K/33.3	$53.2_{\pm 0.1}$/2.25K/23.6
$A_t = \ell_t^{\mathrm{LT}} - \ell_t^{\mathrm{HT}}$	$\mathbf{58.5_{\pm 0.3}}$**/1.10K/53.2**	$\mathbf{55.8_{\pm 0.2}}$**/0.75K/74.4**	$\mathbf{57.0_{\pm 0.2}}$**/0.91K/62.6**

Table 7: **Comparison with other training paradigms.** We use Qwen3-VL-8B as the teacher and Qwen3-VL-2B as the student. **Note that when evaluated under LT and HT prompts, Qwen3-VL-8B achieves an average performance of 62.9/0.55K/114.4 and 66.3/1.94K/34.2, respectively.**

Method	Reasoning	Understanding	AVG
Qwen3-VL-2B	$51.6_{\pm 0.6}$/2.84K/18.2	$53.9_{\pm 0.2}$/1.55K/34.8	$52.9_{\pm 0.4}$/2.12K/25.0
+ PE with LT	$53.1_{\pm 0.1}$/2.00K/26.6	$53.3_{\pm 0.4}$/1.04K/51.3	$53.2_{\pm 0.3}$/1.47K/36.2
+ SFT on LT data	$52.6_{\pm 0.4}$/2.44K/21.6	$52.9_{\pm 0.4}$/1.37K/38.6	$52.8_{\pm 0.2}$/1.85K/28.5
+ DPO (LT data vs. HT data)	$53.8_{\pm 0.7}$/1.99K/27.0	$53.0_{\pm 0.1}$/1.15K/46.1	$53.4_{\pm 0.3}$/1.53K/34.9
+ COPD (Ours)	$\mathbf{58.5_{\pm 0.3}}$**/1.10K/53.2**	$\mathbf{55.8_{\pm 0.2}}$**/0.75K/74.4**	$\mathbf{57.0_{\pm 0.2}}$**/0.91K/62.6**

Table 8: **Training efficiency under the Qwen3-VL-8B→2B.** GPU hours denote the aggregate training time across all GPUs.

Method	GPU Hours	AVG
Base	–	$52.9_{\pm 0.4}$/2.12K/25.0
OPD	132	$53.5_{\pm 0.1}$/2.23K/24.0
ExOPD	150	$55.2_{\pm 0.1}$/2.08K/26.5
COPD	**60**	$\mathbf{57.0_{\pm 0.2}}$**/0.91K/62.6**

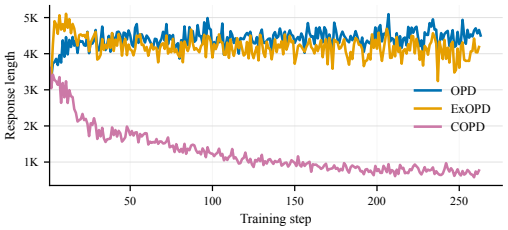


Figure 2: **Response length over training.** COPD progressively shortens rollout responses, whereas others remain at high response lengths.

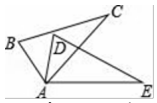
vides an indispensable baseline, enabling the contrastive signal to selectively penalize redundant continuations rather than merely imitate the teacher distribution under a LT prompt.

3.3.6 Comparison with Other Training Paradigms

Table 7 demonstrates that alternative training paradigms fail to fully reproduce the benefits of our LT–HT contrastive formulation. Specifically, we use Qwen3-VL-8B as the teacher to generate offline responses under the light- and heavy-thinking instructions for SFT and DPO (Rafailov et al., 2023) training. Among the evaluated alternatives, DPO achieves the highest average accuracy at 53.4%, whereas prompt engineering (PE)—which directly applies the light-thinking instruction to the student model during inference—yields the shortest average response (1.47K tokens) and highest Acc@1K (36.2). In comparison, COPD advances average accuracy to 57.0% (+3.6 pp over DPO), while simultaneously cutting response length by 38.1% and boosting Acc@1K by 72.9% relative to PE. These findings demonstrate that an on-policy, token-level contrastive advantage provides far more granular supervision and credit assignment than sequence-level imitation or offline preference optimization, suppressing redundant reasoning without compromising problem-solving accuracy.

3.4 Training Efficiency

Table 8 and Figure 2 demonstrate that COPD excels in both training and inference efficiency. It requires merely 60 GPU hours, reducing compute costs by 54.5% over Vanilla OPD and 60.0% over ExOPD. Concurrently, COPD delivers the highest average accuracy with the shortest outputs. Crucially, its training dynamics (Figure 2) show a progressive reduction in response length below 1K tokens, while competing methods remain largely static.



Question: As shown in the figure, triangle ABC congruent triangle ADE, Point A is their common vertex. if angle B = 70.0, angle C = 30.0, angle DAC = 35.0, then the degree of angle EAC is ()
Choices:
A:40° B:45°
C:35° D:25°

Qwen3.5-2B [3,680 tokens]

The user wants me to find the measure of angle EAC based on the provided image and text. **1. Analyze the given information:** $\triangle ABC \cong \triangle ADE$ ... Specifically, \angle B = \angle DAE$ ... Given $\angle B = 70^\circ$, $\angle C = 30^\circ$, $\angle DAC = 35^\circ$. ... **Step 1:** $\angle DAE = \angle B = 70^\circ$. **Step 2:** Looking at the diagram, $\angle DAE = \angle DAC + \angle EAC$. **Step 3:** $70^\circ = 35^\circ + \angle EAC$ ... $\angle EAC = 35^\circ$. Therefore, the degree of angle EAC is 35°. **Answer:** The correct option is C.

Qwen3.5-2B + COPD [1,626 tokens]

The user wants me to find the measure of angle EAC based on the given information about two congruent triangles. **1. Analyze the given information:** $\triangle ABC \cong \triangle ADE$ ... $\angle B = 70^\circ$, $\angle C = 30^\circ$, $\angle DAC = 35^\circ$. **2. Determine missing angles in $\triangle ABC$:** $\angle BAC = 180^\circ - 70^\circ - 30^\circ = 80^\circ$. ... **4. Use the congruence property:** Since $\triangle ABC \cong \triangle ADE$, then $\angle DAE = \angle BAC = 80^\circ$. **5. Calculate $\angle EAC$:** $\angle EAC = \angle DAE - \angle DAC$ ... $\angle EAC = 80^\circ - 35^\circ = 45^\circ$. **Answer:** B

Figure 3: **Case study.** COPD preserves decisive steps while removing redundancy.

3.5 Case Study

Figure 3 illustrates a representative case demonstrating this behavioral shift. The base model (Qwen3.5-2B) produces an overly verbose trajectory of 3,680 tokens, yet falters during repetitive self-checking and derives an incorrect answer. In contrast, COPD reaches the correct solution directly with only 1,626 tokens (44.2% of the base model's length). This confirms that verbose deliberation does not prevent reasoning errors; instead, COPD effectively prunes unproductive verification loops while preserving decisive reasoning steps. More case studies can be found in Appendix D.

4 Related Work

On-policy distillation and multimodal reasoning. While large vision-language models have made substantial progress across multimodal tasks through step-by-step reasoning, their rationales frequently accumulate redundant visual restatements and excessive deliberation (Jiang et al., 2025). To enhance student capabilities, on-policy distillation (OPD) mitigates the exposure bias of offline distillation by sampling rollouts from the student policy and having a teacher score the visited prefixes (Lu & Lab, 2025). Recent variants refine reward scaling (ExOPD (Yang et al., 2026)) or visual grounding (VA-OPD (Liu et al., 2026)) to improve the student performance. In contrast, COPD shifts the axis of supervision: the frozen teacher scores each token at the identical student state under contrasting light- and heavy-thinking prompts, providing a state-matched preference between concise and exhaustive reasoning.

Reasoning compression and preference learning. Prior efforts to compress reasoning chains primarily rely on supervised fine-tuning over curated concise rationales (Munkhbat et al., 2025) or reinforcement learning with explicit token budgets and heuristic length penalties (Han et al., 2025; Jin et al., 2025b; Liu et al., 2025). However, these approaches largely depend on scarce curated offline datasets, fragile length hyperparameters, or coarse outcome-level feedback. In contrast, COPD requires no external task rewards or explicit length regularizers; instead, it directly derives intrinsic token-level compression preferences from paired teacher likelihoods evaluated at student-visited states. A broader discussion of related works is provided in Appendix B.

5 Conclusion

We propose COPD, a contrastive on-policy distillation framework for adaptive reasoning compression. By evaluating each student-generated token under paired light- and heavy-thinking prompts, a frozen teacher model converts absolute token likelihoods into relative preference signals that encourage concise reasoning. This design preserves the benefits of online supervision in OPD without relying on predefined length budgets or auxiliary penalties. Extensive experiments, including evaluations on nine multimodal benchmarks across four teacher–student configurations and self-distillation variants, show that COPD consistently improves the accuracy–efficiency Pareto frontier. These results suggest that concise reasoning can emerge from relative preferences without sacrificing problem-solving accuracy. More broadly, future work could extend this pairwise comparison mechanism to other behavioral alignment tasks, such as enforcing safety constraints, controlling writing style, and improving factuality.

AI USE STATEMENT

Generative AI tools were utilized in this work for the following purposes: language polishing, refining manuscript organization, checking LaTeX formatting consistency, assisting in literature search for related work, and optimizing code for figure plotting. These tools were not used to generate or modify experimental data. All authors have reviewed the AI-assisted content and take full responsibility for the final content, claims, and artifacts of this work.

ETHICS STATEMENT

The experiments in this work are conducted exclusively on publicly released academic benchmarks and research datasets, involving no human participants, private user information, or sensitive personal data. Regarding broader societal and environmental impacts, reducing reasoning length directly curtails the per-query computational overhead and aggregate carbon footprint of deploying large multimodal models at scale. However, aggressive reasoning compression presents critical trade-offs: eliding intermediate deductive chains may diminish model transparency and interpretability in safety-critical domains such as medical diagnosis, legal analysis, and autonomous control. Furthermore, overly concise responses may project unwarranted certainty even when predictions are erroneous. We recommend that deployments in high-stakes environments integrate compressed reasoning models with robust confidence calibration, explicit visual grounding checks, and adaptive fallback mechanisms to full chains of thought.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility of all algorithmic and empirical findings, Section 2 provides complete mathematical formulations and procedural descriptions for on-policy rollouts, paired teacher scoring, contrastive token advantage, and policy optimization. Full experimental configurations—including distributed training parameters, optimizer schedules, hardware environments, and standardized evaluation protocols across all nine multimodal benchmarks—are documented in Section 3 and Appendices A.1–A.3. The exact instruction templates for light- and heavy-thinking prompts are fully reported in Appendix C and Tables 14–19.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Razvan-Gabriel Dumitru, Darius Peteleaza, Vikas Yadav, and Liangming Pan. Conciserl: Conciseness-guided reinforcement learning for efficient reasoning models. In *EMNLP (Findings)*, pp. 17099–17123, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, pp. 1049–1065, 2023.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025a.

Zhensheng Jin, Xinze Li, Yifan Ji, Chunyi Peng, Zhenghao Liu, Qi Shi, Yukun Yan, Shuo Wang, Furong Peng, and Ge Yu. Recut: Balancing reasoning length and accuracy in llms via stepwise trails and preference optimization. In *EMNLP (Findings)*, pp. 14269–14282, 2025b.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yanhao Li, Lu Ma, Jiaran Zhang, Lexiang Tang, Wentao Zhang, and Guibo Luo. Leash: Adaptive length penalty and reward shaping for efficient large reasoning model. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2846–2856, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, Kai Jia, and Zhifang Sui. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21135–21156, 2026c.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, et al. Dler: Doing length penalty right-incentivizing more intelligence per token via reinforcement learning. *arXiv preprint arXiv:2510.15110*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

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 *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25127–25152, 2025.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise thoughts: Impact of output length on llm reasoning and cost. *Information Sciences*, pp. 123939, 2026.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. 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)*, pp. 20023–20070, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Jiacheng Ruan, Dan Jiang, Xian Gao, Ting Liu, Yuzhuo Fu, and Yangyang Kang. Mme-sci: A comprehensive and challenging science benchmark for multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8760–8768, 2026.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. Crisp: Compressed reasoning via iterative self-policy distillation. *arXiv preprint arXiv:2603.05433*, 2026.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4323–4332, 2019.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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. *Advances in Neural Information Processing Systems*, 38:30865–30891, 2026.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. In *International Conference on Learning Representations*, volume 2026, pp. 127108–127138, 2026.

xAI. Realworldqa: A benchmark for real-world spatial understanding. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024. Accessed: 2025-04-26.

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

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Jingyang Yi, Jiazheng Wang, and Sida Li. Shorterbetter: Guiding reasoning models to find optimal inference length for efficient reasoning. *Advances in Neural Information Processing Systems*, 38: 39011–39043, 2026.

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

Wenzhen Yuan, Jiacheng Ruan, Wutao Xiong, Chengping Zhao, Ting Liu, and Yuzhuo Fu. Mmgist: A comprehensive multimodal benchmark for 2027. *arXiv preprint arXiv:2606.22437*, 2026b.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Ke Zhang, Yunjie Tian, Dongdi Zhao, Yijiang Li, Yuanye Liu, Vishal M Patel, and Di Fu. On-policy distillation with best-of-n teacher rollout selection. *arXiv preprint arXiv:2605.09725*, 2026a.

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.

Yaocheng Zhang, Jiajun Chai, Yuqian Fu, Songjun Tu, Xiaohan Wang, Wei Lin, Guojun Yin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Are full rollouts necessary for on-policy distillation? *arXiv preprint arXiv:2605.31490*, 2026b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

A Extended Experimental Setup and Analysis

A.1 Training Infrastructure and Hyperparameters

We implement our on-policy distillation pipeline within `verl` (Sheng et al., 2025), combining Fully Sharded Data Parallel (FSDP (Zhao et al., 2023)) for distributed policy updates with vLLM (Kwon et al., 2023) for high-throughput rollout generation and teacher scoring. At each iteration, the student policy samples response trajectories for a batch of inputs from ViRL-16K. Conditioned on the shared multimodal input and generated response prefix, the frozen teacher scores each sampled trajectory twice under contrasting light- and heavy-thinking prompts.

Following Section 2, the token-level contrastive advantage and its clipped form are defined as:

$$A_t = \ell_t^{\mathrm{LT}} - \ell_t^{\mathrm{HT}}, \qquad \tilde{A}_t = \mathrm{stopgrad}(\mathrm{clip}(A_t, -10, 10)) \,. \tag{7}$$

This clipped advantage functions as a scalar weight in the policy-gradient objective defined in the main paper, with gradients updating the student parameters exclusively. Bounding the advantage magnitude prevents destabilizing updates and ensures numerical stability.

Regarding optimization hyperparameters, we employ the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-6}, weight decay of 1×10^{-2}, and a maximum gradient clipping norm of 1.0. The global batch size is 192 with a PPO mini-batch size of 48. Training runs for 3 epochs, with maximum sequence lengths for input prompts and model responses capped at 4,096 and 8,192 tokens, respectively. The training objective incorporates no auxiliary task, format, accuracy, or explicit length rewards.

A.2 Evaluation Benchmarks

We evaluate on nine multimodal benchmarks, comprising a total of $18{,}656$ test instances. For structured presentation and analysis, we partition these benchmarks into four reasoning suites and five understanding suites:

Multimodal Reasoning. MathVista (Lu et al., 2024) (*testmini*, $1{,}000$ questions) benchmarks mathematical reasoning across diverse visual contexts, including natural photographs, synthetic charts, geometric diagrams, and academic figures, using both free-form generation and multiple-choice formats. MathVerse (Zhang et al., 2024) (*testmini*, $3{,}940$ evaluation instances) evaluates plane geometry, solid geometry, and functions through five visual variants of 788 core problems that systematically redistribute information between text and diagrams; we exclude the text-only variant to maintain strict multimodal evaluation. WeMath (Qiao et al., 2025) (*testmini*, $1{,}740$ questions) probes hierarchical visual mathematical reasoning by pairing complex multi-step problems with their prerequisite sub-problems, featuring an explicit "No correct answer" option to penalize superficial guesswork. LogicVista (Xiao et al., 2024) (448 multiple-choice questions) assesses deductive, inductive, numerical, spatial, and mechanical reasoning, sourcing visual questions directly from licensed intelligence and aptitude tests.

Multimodal Understanding. HallusionBench (Guan et al., 2024) (HalluBench; 951 image-grounded questions) diagnoses language hallucinations and visual illusions via paired yes/no queries targeting original versus human-edited images. We evaluate on the 951 image-bearing instances from the full $1{,}129$-question benchmark, omitting the 178 text-only controls. MMMU (Yue et al., 2024) (*dev+val*, $1{,}050$ questions) tests college-level domain expertise across six disciplines spanning STEM, business, and the humanities. MMStar (Chen et al., 2024a) ($1{,}500$ questions across six core capabilities) enforces strict visual dependency through rigorous expert human verification, filtering out instances solvable via unimodal language priors. RealWorldQA (xAI, 2024) (RWQA; 765 questions) evaluates physical spatial understanding in real-world scenes. MMGist (Yuan et al., 2026b) ($7{,}262$ questions spanning seven capability dimensions) is distilled from eighteen public benchmarks via text-ablation filtering, cross-model saturation screening, and human-verified anomaly removal.

Prompt and Evaluation Protocol. All primary evaluations follow a strict zero-shot protocol, where model inputs are restricted to the visual inputs, question text, and candidate options when available. For HallusionBench, we append a standard yes/no response formatting directive. Input format-

Table 9: **Scientific reasoning on MMESci-VisionOnly.** Performance across high-school Physics, Math, Chemistry, Biology, and overall evaluation (All). Each cell reports accuracy (mean$_{\pm\text{std}}$), average response length (in thousands of tokens), and the efficiency index Acc@1K.

Method	Physics	Math	Chemistry	Biology	All
Qwen3-VL-2B	$6.9_{\pm 0.2}$/5.00K/1.4	$13.2_{\pm 1.7}$/6.27K/2.1	$14.1_{\pm 1.6}$/3.90K/3.6	$15.9_{\pm 1.9}$/3.47K/4.6	$11.5_{\pm 0.2}$/4.90K/2.3
Qwen3-VL-8B→Qwen3-VL-2B					
OPD	$6.8_{\pm 0.4}$/6.38K/1.1	$10.7_{\pm 1.7}$/7.32K/1.5	$14.4_{\pm 0.0}$/4.93K/2.9	$16.3_{\pm 1.5}$/4.34K/3.8	$10.9_{\pm 0.4}$/6.03K/1.8
ExOPD	$6.4_{\pm 0.6}$/7.03K/0.9	$13.6_{\pm 0.7}$/7.34K/1.9	$13.0_{\pm 1.3}$/5.77K/2.3	$14.2_{\pm 1.4}$/5.02K/2.8	$10.9_{\pm 0.9}$/6.55K/1.7
VA-OPD	$8.0_{\pm 0.6}$/6.31K/1.3	$13.9_{\pm 0.7}$/7.21K/1.9	$16.2_{\pm 0.7}$/4.78K/3.4	$\mathbf{16.6}_{\pm 0.9}$/4.38K/3.8	$12.7_{\pm 0.3}$/5.95K/2.1
COPD	$\mathbf{9.8}_{\pm 0.4}$**/2.25K/4.4**	$\mathbf{14.1}_{\pm 0.0}$**/3.36K/4.2**	$\mathbf{17.4}_{\pm 0.3}$**/1.92K/9.1**	$15.6_{\pm 0.5}$**/1.63K/9.6**	$\mathbf{13.4}_{\pm 0.1}$**/2.40K/5.6**

ting and response constraints remain identical across all evaluated models, without custom system prompts or in-context demonstrations.

A.3 Evaluation Implementation Details

Serving and Decoding. We serve standalone model checkpoints via vLLM (v0.18.0) in bfloat16 precision. Unless specified otherwise, decoding hyperparameters are set to `max_tokens` $=$ 8192, `temperature` $= 0.7$, `top_p` $= 0.8$, `top_k` $= 20$, `presence_penalty` $= 1.5$, `repetition_penalty` $= 1.0$, and `seed` $= 3407$.

Repeated Evaluations. Primary results summarize three evaluation passes, with accuracy reported as mean$_{\pm\text{std}}$ and response length reported as the mean.

Automated Grading. We employ Qwen3.6-Flash at temperature 0 as the automated judge, prompted with the question, ground-truth answer, and candidate response to output either `correct` or `incorrect`. When a response contains a `\boxed{}` expression, only the boxed content is extracted for grading. Accuracy is computed as the proportion of responses graded correct.

Evaluation Metrics. Sequence lengths are computed using each model's native tokenizer, measuring the complete assistant response inclusive of reasoning segments. Reasoning, Understanding, and overall AVG denote the averages across four, five, and all nine benchmarks, respectively. To measure performance density, Acc@1K calculates the accuracy per thousand output tokens across the benchmark set $\mathcal{B}$:

$$\text{Acc@1K} = 1000 \cdot \frac{\sum_{b \in \mathcal{B}} \text{Acc}_b}{\sum_{b \in \mathcal{B}} \text{Tokens}_b}, \tag{8}$$

where Acc_b and Tokens_b denote the accuracy and average response length on benchmark b, respectively.

Evaluated Variations. To examine behavior across diverse inference settings, Figure 4 varies the generation budget over $\{2048, 4096, 8192, 16384\}$ tokens. In addition, the text-only evaluation in Table 10 expands the generation budget to 32K tokens, Table 12 examines inference with thinking disabled (`enable_thinking=false`), and Table 13 evaluates greedy decoding (temperature 0).

A.4 Generalization to High-School STEM Reasoning

Table 9 evaluates out-of-distribution (OOD) transfer to MMESci-VisionOnly (Ruan et al., 2026), a comprehensive and challenging high-school level multimodal scientific reasoning benchmark spanning Physics, Math, Chemistry, and Biology. The un-distilled Base model achieves an overall accuracy of 11.5% with an average response length of 4.90K tokens and an Acc@1K of 2.3. Conventional on-policy distillation baselines fail to improve efficiency: both OPD and ExOPD degrade accuracy to 10.9% while substantially inflating response lengths to 6.03K and 6.55K tokens, respectively. While VA-OPD marginally improves accuracy to 12.7%, it likewise incurs severe generation overhead, averaging 5.95K tokens.

In contrast, COPD achieves the highest overall accuracy of 13.4% at only 2.40K tokens, outperforming Base by 1.9 pp while cutting response length by 51.0%, which translates to an Acc@1K of 5.6 (more than doubling VA-OPD's 2.1). Across disciplines, COPD consistently slashes response lengths by more than half across all four subjects while matching or surpassing competing base-

Table 10: **Generalization to text-only reasoning under an avg@8 protocol.** Evaluated on Qwen3-VL-8B→2B with temperature $T = 0.7$ and a 32K-token generation budget. Accuracy is reported as mean$_{\pm\text{std}}$ across eight independent samples per problem. The AVG column summarizes average accuracy, average response length (in thousands of tokens), and the efficiency index Acc@1K.

Method	MATH500	GPQA	AIME24	AIME25	AMC23	AVG
Qwen3-VL-2B	$80.6_{\pm 1.1}$	$11.2_{\pm 1.8}$	$19.2_{\pm 4.3}$	$17.5_{\pm 5.0}$	$53.4_{\pm 3.3}$	$36.4_{\pm 1.4}$/19.42K/1.9
+ OPD	$81.6_{\pm 0.8}$	$11.2_{\pm 1.5}$	$17.5_{\pm 2.4}$	$\mathbf{21.7}_{\pm 5.9}$	$54.4_{\pm 2.2}$	$37.3_{\pm 1.7}$/19.09K/2.0
+ ExOPD	$\mathbf{83.4}_{\pm 0.7}$	$12.7_{\pm 1.8}$	$20.4_{\pm 5.5}$	$19.6_{\pm 6.3}$	$54.4_{\pm 3.7}$	$38.1_{\pm 2.2}$/17.62K/2.2
+ VA-OPD	$82.2_{\pm 0.6}$	$12.1_{\pm 1.1}$	$18.3_{\pm 5.9}$	$\mathbf{21.7}_{\pm 7.6}$	$54.4_{\pm 3.7}$	$37.7_{\pm 2.0}$/19.10K/2.0
+ COPD	$81.4_{\pm 0.8}$	$\mathbf{13.1}_{\pm 1.1}$	$\mathbf{21.7}_{\pm 4.0}$	$20.0_{\pm 4.0}$	$\mathbf{56.6}_{\pm 3.5}$	$\mathbf{38.5}_{\pm 1.2}$**/8.54K/4.5**

lines. Specifically, it establishes leading accuracies in Physics (9.8% vs. Base 6.9%), Math (14.1% vs. Base 13.2%), and Chemistry (17.4% vs. Base 14.1%), with its Acc@1K surging to 9.1 in Chemistry. Even in Biology, COPD maintains competitive accuracy (15.6%) with dramatically curtailed verbosity (1.63K vs. 4.38K tokens for VA-OPD), yielding an efficiency index of 9.6. These findings demonstrate that COPD generalizes effectively to demanding high-school scientific reasoning, consistently fostering concise yet accurate derivation.

A.5 Generalization to Text-Only Reasoning

Table 10 evaluates cross-modal transfer to five challenging text-only reasoning benchmarks: MATH500 (Lightman et al., 2024), GPQA (Rein et al., 2023), AIME24[3], AIME25[4], and AMC23[5]. Under the Qwen3-VL-8B→2B distillation setting, decoding follows standard evaluation protocols for long-form reasoning using avg@8 sampling with temperature 0.7 and a 32K-token generation budget. Relative to the Base model, COPD improves average accuracy from 36.4% to 38.5% while reducing average response length from 19.42K to 8.54K tokens (a 56.0% reduction), more than doubling the efficiency index Acc@1K from 1.9 to 4.5.

Across all evaluated methods, COPD attains the highest average accuracy alongside the most compact response length. It achieves peak performance on GPQA, AIME24, and AMC23, while ExOPD leads on MATH500 and OPD ties with VA-OPD on AIME25. Crucially, conventional on-policy distillation baselines continue to generate exceedingly verbose reasoning chains averaging 17.62K–19.10K tokens. Relative to the strongest baseline ExOPD, COPD yields an accuracy gain of 0.4 pp while slashing response length by 51.5%. These results demonstrate that the conciseness incentives learned by COPD are modality-agnostic, transferring seamlessly to pure-text complex mathematical and scientific reasoning without sacrificing performance.

A.6 Accuracy per Token Across Generation Budgets

Experimental Setup. Figure 4 illustrates the Pareto trade-offs on Qwen3.5-2B ← Qwen3.5-35A3B across four approaches: the Base model, OPD, ExOPD, and COPD. Across four generation budgets ($\{2048, 4096, 8192, 16384\}$ tokens on a $\log_2$ scale), the vertical axis reports average accuracy across nine multimodal benchmarks, while marker size (bubble area) represents average response length.

Comparison at Matched Budgets. Across all four budget tiers, COPD strictly Pareto-dominates the Base model, OPD, and ExOPD, simultaneously achieving superior accuracy and substantially more concise responses (consistently smaller bubble areas). Under the tightest budget of 2,048 tokens, standard OPD and ExOPD inherit verbose teacher reasoning and suffer from frequent generation truncation, dropping below the Base model to 52.1% and 51.6%, respectively. In contrast, COPD reliably completes derivation within the restricted budget, attaining 55.1% accuracy (+3.5 pp over ExOPD). Scaling the budget to 4K, 8K, and 16K tokens steadily elevates the accuracy of COPD to 57.9%, 59.6%, and 60.2%, continuously establishing the uppermost Pareto frontier.

[3] https://huggingface.co/datasets/AI-MO/aimo-validation-aime

[4] https://huggingface.co/datasets/opencompass/AIME2025

[5] https://huggingface.co/datasets/AI-MO/aimo-validation-amc

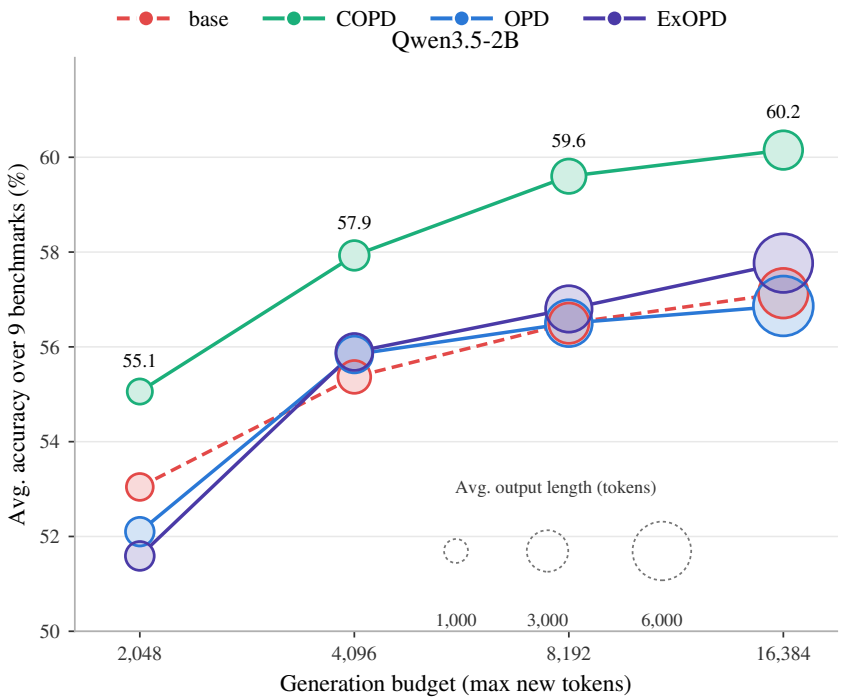


Figure 4: **Accuracy and output length across four generation budgets on Qwen3.5-2B ← Qwen3.5-35A3B.** The horizontal axis denotes the maximum token generation budget on a $\log_2$ scale ($\{2048, 4096, 8192, 16384\}$); the vertical axis denotes accuracy averaged across nine multimodal benchmarks. Bubble area represents average output length. Curves compare the Base model (red dashed), standard OPD (blue solid), ExOPD (purple solid), and COPD (green solid). Numerical labels indicate the accuracy of COPD.

Table 11: **Comparison with length-penalty GRPO on Qwen3.5-2B.** Qwen3.5-4B is used as the teacher for COPD. The number of rollouts of GRPO is set to 8. Each cell reports accuracy, average response length, and Acc@1K. Training cost reflects total GPU hours.

Method	Reasoning	Understanding	AVG	GPU Hours
Qwen3.5-2B	57.1/3.65K/15.6	56.0/2.19K/25.6	56.5/2.84K/19.9	-
+ length-penalty GRPO	59.9/2.47K/24.3	57.7/1.58K/36.5	58.6/1.98K/29.6	371
+ COPD	**62.6/2.39K/26.2**	**58.2/1.56K/37.3**	**60.2/1.93K/31.2**	62

Comparison Across Budgets. Cross-budget comparisons highlight an even more pronounced efficiency advantage: COPD achieves 57.9% accuracy under a moderate 4,096-token budget, already surpassing the peak performance of Base (57.1%), OPD (56.8%), and ExOPD (57.7%) operating under a fourfold larger budget of 16,384 tokens, while requiring less than half the response length. Furthermore, extending the generation budget from 8,192 to 16,384 tokens yields negligible accuracy gains for the baselines despite massive bubble enlargement (with ExOPD swelling to nearly 6,000 tokens), indicating that conventional on-policy distillation leads to vacuous token proliferation under higher budgets. In contrast, COPD effectively checks redundant reasoning, ensuring that expanded computational capacity directly translates into higher task accuracy.

A.7 Comparison with Length-Penalty Reinforcement Learning

Table 11 compares length-penalty GRPO against COPD on Qwen3.5-2B, using Qwen3.5-4B as the teacher for COPD. GRPO increases average accuracy from 56.5% to 58.6% while shortening responses from 2.84K to 1.98K tokens. COPD outperforms GRPO across all metrics, achieving 60.2% accuracy at 1.93K tokens (a 1.6 pp accuracy gain and 2.5% shorter outputs), yielding an

Table 12: **Evaluation with the no-thinking mode in Qwen3.5.** All models are evaluated with `enable_thinking=false` at inference time under an 8192-token generation budget (`max_tokens` = 8192). Distilled models use Qwen3.5-35B-A3B as the teacher. Cells report Accuracy / Average Tokens (in thousands) / Acc@1K across four Reasoning benchmarks, five Understanding benchmarks, and their overall average (AVG). COPD achieves the highest overall accuracy and most compact responses across both student model scales.

Method	Reasoning	Understanding	AVG
Qwen3.5-2B	55.6/2.13K/26.1	55.8/0.98K/56.9	55.7/1.50K/37.1
+ OPD	56.1/2.84K/19.8	56.0/1.39K/40.3	56.0/2.03K/27.6
+ COPD	**57.9/0.54K/107.2**	**57.2/0.38K/150.5**	**57.5/0.45K/127.8**
Qwen3.5-9B	75.0/1.56K/48.1	65.7/1.05K/62.6	69.8/1.28K/54.5
+ OPD	75.6/3.12K/24.2	66.5/1.97K/33.8	70.5/2.48K/28.4
+ COPD	**76.2/0.84K/90.7**	**67.2/0.57K/117.9**	**71.2/0.69K/103.2**

Acc@1K of 31.2 versus 29.6. Crucially, COPD requires only 62 GPU hours for training, compared to 371 GPU hours for GRPO.

The performance and efficiency advantages of COPD over length-penalty reinforcement learning stem fundamentally from the granularity of supervision. Sequence-level length penalties in RL (e.g., GRPO) impose a uniform scalar penalty across the entire trajectory, failing to distinguish indispensable deductive tokens from superficial verbosity. Consequently, length-constrained RL risks truncating valid reasoning chains and requires extensive exploratory rollouts (e.g., N=8), incurring substantial computational overhead (371 GPU hours). In contrast, COPD provides dense, token-level contrastive guidance at every decoding step, selectively curbing vacuous elaboration while safeguarding core deductive steps. This token-level precision allows COPD to establish a superior accuracy–length frontier while slashing training compute by 83.3%.

A.8 Accuracy and Output Length with the No Thinking Mode

Table 12 evaluates Qwen3.5-2B and Qwen3.5-9B when the reasoning block is explicitly disabled via `enable_thinking=false` at inference time under an 8192-token budget. At the 2B scale, COPD achieves 57.5% average accuracy with an average response length of only 0.45K tokens. Relative to the Base model (55.7% at 1.50K tokens), this represents a 1.8 pp accuracy gain alongside a 70.0% reduction in output length. OPD fails to produce concise answers, inflating output length to 2.03K tokens. Consequently, COPD elevates Acc@1K to 127.8, achieving a $3.44\times$ efficiency gain over the Base baseline (37.1).

At the 9B scale, COPD demonstrates a similarly dominant efficiency advantage, reaching 71.2% average accuracy with only 0.69K tokens. Compared to Base (69.8% at 1.28K tokens), COPD delivers a 1.4 pp accuracy improvement while cutting response length by 46.1%, boosting Acc@1K to 103.2 ($1.89\times$ that of Base at 54.5). In contrast, OPD generates 2.48K tokens—over $3.5\times$ the length of COPD—while stalling at 70.5% accuracy.

Crucially, these results demonstrate that the conciseness induced by COPD is not an artifact of the explicit thinking block. When `enable_thinking=false` is enforced at inference time, un-distilled and standardly distilled models continue to produce verbose responses (1.28K–2.48K tokens). Merely turning off the thinking flag during decoding fails to instill reasoning efficiency; in contrast, COPD intrinsically steers the policy toward dense, non-redundant derivation.

A.9 Lexical Composition of Compressed Reasoning Chains

Figure 5 dissects the lexical composition of reasoning chains produced by Base, conventional on-policy distillation baselines (OPD, ExOPD, VA-OPD), COPD, and the Qwen3-VL-8B teacher. We aggregate task restatement, step enumeration, and hedging markers under the category of *boilerplate*. Panels (a)–(c) report absolute counts and length-normalized densities across functional markers, while panel (d) details instance-level paired comparisons across all $18{,}656$ evaluation samples.

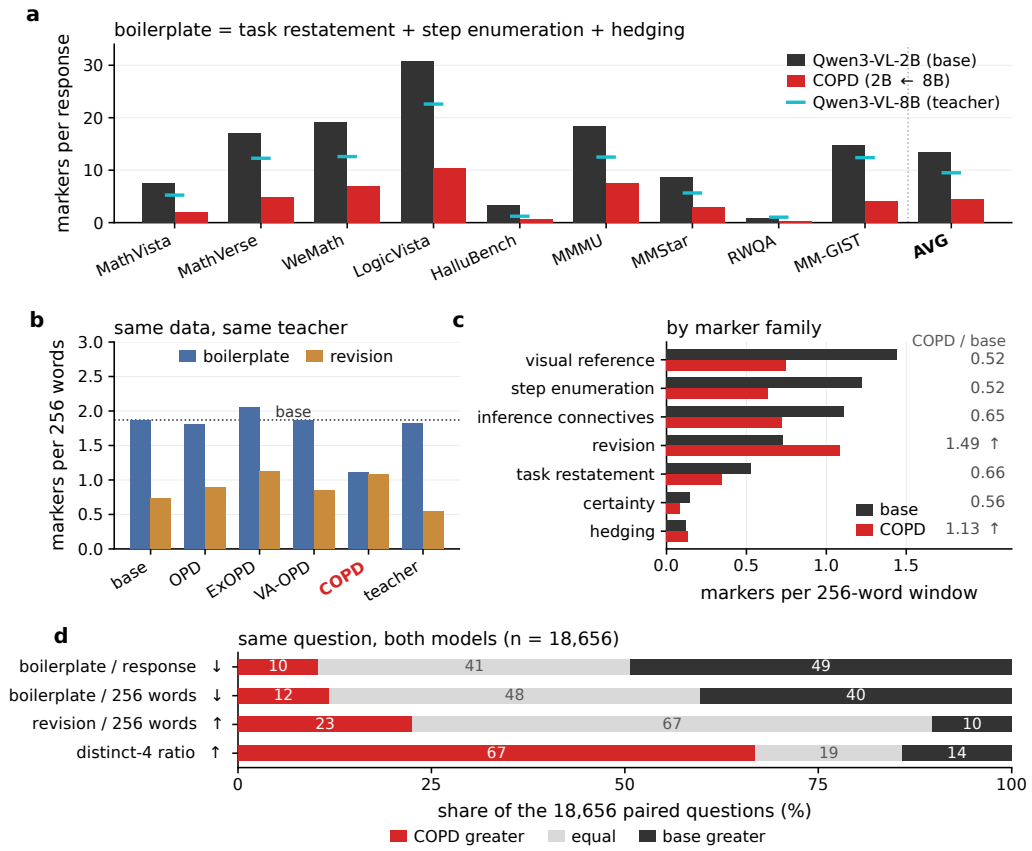


Figure 5: **Lexical marker statistics for Base, COPD, and comparison models.** (a) Frequency of boilerplate markers (task restatement, step enumeration, and hedging) per response across nine multimodal benchmarks; AVG denotes the macro-average. (b) Normalized densities of boilerplate and revision markers per 256-word window across distillation baselines. (c) Density comparison across seven marker families within a 256-word window, annotated with COPD/Base ratios; revision and hedging marker densities increase, while densities for the other five marker families decrease. (d) Distribution of paired differences over 18,656 questions, where red denotes higher values for COPD, gray denotes ties, and dark gray denotes higher values for Base; arrows indicate the preferred direction for each metric.

Boilerplate Suppression Across Absolute and Normalized Metrics. As shown in panel (a), COPD consistently generates fewer boilerplate markers than Base across all nine benchmarks, curtailing the macro-average from 13.4 to 4.4 markers per response, well below the teacher's 9.5. This suppression holds equally under length normalization: panel (b) shows that boilerplate density falls from 1.87 markers per 256 words in Base to 1.12 in COPD, whereas standard OPD baselines remain elevated at 1.80–2.05. Consequently, COPD curtails the concentration of rhetorical scaffolding within the reasoning text itself.

Functional Divergence Across Lexical Families. Panel (c) traces the origin of this density shift across individual marker categories. Within the boilerplate group, mechanical expressions drop precipitously: task restatements and step enumerations contract to $0.66\times$ and $0.52\times$ of Base density, respectively, driving the overall decline. Conversely, hedging markers—which often signal cautious verification—rise modestly to $1.13\times$. Across the remaining families, visual references fall to $0.52\times$, whereas revision markers markedly increase to $1.49\times$. This functional divergence demonstrates that COPD achieves selective compression: the policy purges perfunctory restatements and enumerations while concentrating critical self-monitoring and error-correction operations.

Table 13: **Evaluation under deterministic greedy decoding ($T = 0$).** All distilled methods employ Qwen3-VL-8B as the teacher and Qwen3-VL-2B as the student. Cells report Accuracy / Average Tokens (in thousands) / Acc@1K across Reasoning benchmarks, Understanding benchmarks, and their overall average (AVG).

Method	Reasoning	Understanding	AVG
Qwen3-VL-2B	47.2/3.24K/14.6	51.4/1.75K/29.4	49.6/2.41K/20.6
+ OPD	48.7/3.33K/14.6	53.1/1.78K/29.8	51.2/2.46K/20.8
+ ExOPD	53.0/3.07K/17.3	52.7/1.76K/29.9	52.8/2.35K/22.5
+ VA-OPD	49.1/3.26K/15.1	52.4/1.74K/30.1	50.9/2.42K/21.0
+ COPD	**55.3/1.52K/36.4**	**54.0/0.96K/56.3**	**54.6/1.21K/45.1**

Instance-Level Distributions and Lexical Diversity. Panel (d) evaluates these distributional shifts at sample granularity across all $18{,}656$ test instances. On absolute boilerplate counts, COPD uses fewer markers on 49% of questions and more on only 10% (41% tied). Controlling for sequence length, boilerplate density remains lower on 40% of questions while increasing on just 12%. In contrast, revision density rises on 23% of instances versus 10% declining. Concurrently, 4-gram diversity (distinct-4) is higher on 67% of instances (lower on 14%), demonstrating that the compressed chains generated by COPD consistently favor denser reasoning steps and richer lexical expression over formulaic templates.

A.10 Evaluation Under Deterministic Greedy Decoding

Table 13 reports evaluation under deterministic greedy decoding (`temperature` $= 0$) for the Qwen3-VL-8B→2B distillation setup. Greedy decoding eliminates decoding stochasticity, directly probing the intrinsic generation policy. COPD achieves an average accuracy of 54.6%, outperforming the un-distilled Base baseline by 5.0 pp and surpassing competitive baselines OPD, ExOPD, and VA-OPD by 3.4 pp, 1.8 pp, and 3.7 pp, respectively. Concurrently, COPD slashes average response length to 1.21K tokens—a 49.8% reduction relative to Base (2.41K) and a 48.5% reduction relative to ExOPD (2.35K)—more than doubling the efficiency index Acc@1K to 45.1 (compared to 20.6 for Base and 22.5 for ExOPD).

This superiority holds consistently across both benchmark partitions. On the Reasoning suite, COPD achieves 55.3% accuracy using only 1.52K tokens, notably outperforming ExOPD (53.0% at 3.07K tokens) while cutting generation length by more than half. On the Understanding suite, COPD similarly delivers higher accuracy (54.0% vs. 52.7%) at nearly half the token budget (0.96K vs. 1.76K). These results confirm that the accuracy–length Pareto advantage of COPD stems directly from the learned policy's deterministic argmax mode, establishing concise reasoning as the model's dominant generation behavior.

A.11 Positional and Lexical Localization of Contrastive Advantage

Figure 6 investigates the fine-grained mechanistic localization of the contrastive advantage by offline scoring Base and COPD responses across nine benchmarks (200 sampled questions per benchmark, up to a 768-token scoring window). We analyze how the advantage signal $A_t = \ell_t^{\mathrm{LT}} - \ell_t^{\mathrm{HT}}$ varies across generation positions and functional lexical categories.

Position Dependence. As depicted in panel (a), Base model responses exhibit a pronounced negative advantage band (blue) at initial token positions, whereas COPD exhibits a consistently positive opening region (deep red) across all benchmarks. The relative-position trajectory in panel (b) highlights this operational disparity: Base responses initiate with a sharp negative trough near -0.54 before flattening toward zero, reflecting an initial bias toward verbose, heavy-thinking preambles. In contrast, COPD eliminates this opening deficit, initiating near zero and sustaining positive advantage values ($+0.05$ to $+0.25$) throughout the remainder of the response. The absolute-position profiles in panel (c) corroborate this behavior: when isolating multi-step reasoning (≥ 20 tokens), Base responses incur negative opening advantages, whereas COPD consistently maintains superior

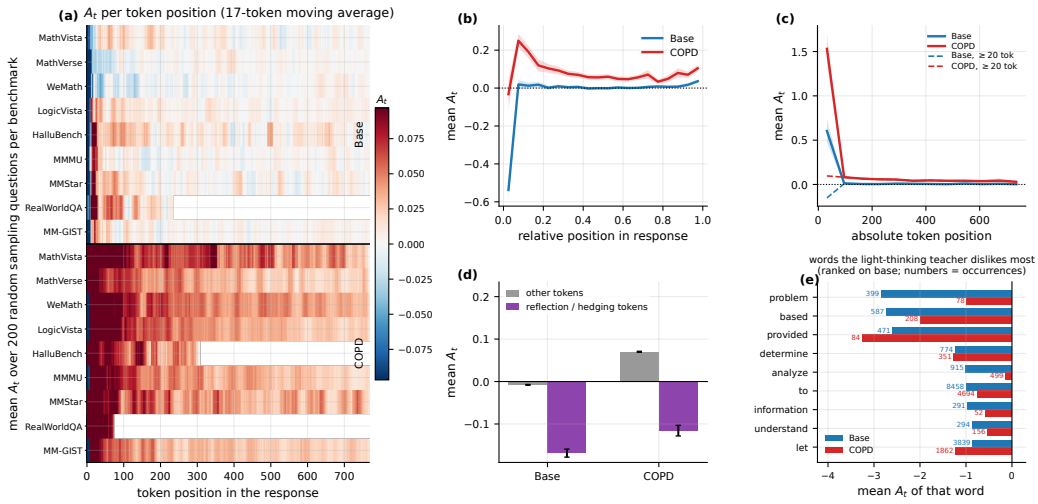


Figure 6: **Positional and lexical localization of the LT–HT contrast.** Offline scoring of sampled Base and COPD responses by the Qwen3-VL-8B teacher using $A_t = \ell_t^{\mathrm{LT}} - \ell_t^{\mathrm{HT}}$. (a) Position-wise mean advantage across nine multimodal benchmarks (17-token moving average). (b) Mean advantage across normalized relative response positions. (c) Mean advantage across absolute token positions; dashed curves restrict evaluation to responses with at least 20 tokens. (d) Mean advantage for reflection/hedging markers versus other sampled tokens. (e) Top vocabulary items most disfavored by the light-thinking context (ranked by mean A_t in Base), along with their occurrence counts across both models.

positive advantage values. This demonstrates that the contrastive objective exerts its strongest corrective pressure at the onset of generation, suppressing introductory bloat.

Lexical Categories. Panel (d) breaks down token advantage across semantic roles. While reflection and deliberative hedging markers naturally receive negative advantage under both models (as they are disproportionately favored by heavy-thinking contexts), other problem-solving tokens shift from near-zero (slightly negative) in Base to $+0.07$ in COPD. Furthermore, panel (e) isolates the specific vocabulary items most penalized by the light-thinking context—predominantly procedural framing verbs and filler nouns such as "let", "understand", "information", and "determine". In COPD responses, the occurrences of these formulaic markers drop substantially (e.g., "let" decreases from 3,839 to 1,862, and "information" from 291 to 52).

Together, these positional and lexical signatures elucidate the mechanics of reasoning compression under COPD. The contrastive objective acts as a targeted structural filter: it applies immediate negative feedback to suppress introductory boilerplate and formulaic hedging at the sequence opening, while steering the generative trajectory toward compact, informative transitions throughout downstream deduction.

B Expanded Related Work

Multimodal Large Language Models and Reasoning. Recent advances in multimodal large language models (MLLMs) integrate visual representations with autoregressive language generation to tackle complex deductive tasks across mathematical problem solving, chart comprehension, and visual question answering (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; 2024; Chen et al., 2024b; Bai et al., 2023; Wang et al., 2024; Bai et al., 2025b;a; Caffagni et al., 2024; Jin et al., 2025a). A substantial body of work on chain-of-thought (CoT) prompting and test-time scaling establishes that extending intermediate reasoning steps can improve task accuracy on challenging problems (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022; Zhou et al., 2022; Yao et al., 2023; Huang & Chang, 2023). However, such extended reasoning can contain redundant steps that increase token use without reliably improving accuracy (Munkhbat et al., 2025; Jin et al., 2025b).

Developing methods that condense reasoning chains while rigorously safeguarding multimodal reasoning capabilities remains a critical objective for deploying efficient MLLMs.

On-Policy Distillation. Knowledge distillation transfers predictive behavior and domain representations from a capable teacher to a compact student (Hinton et al., 2015; Gou et al., 2021; Sanh et al., 2019; Jiao et al., 2020; Sun et al., 2019; Kim & Rush, 2016). Whereas traditional offline distillation suffers from exposure bias on student-visited prefixes, on-policy distillation (OPD) evaluates teacher feedback directly along trajectories sampled from the student policy, closing the distribution shift between static targets and exploration prefixes (Agarwal et al., 2024). Subsequent research investigates reward scaling, exploration strategies, entropy regularization, and multimodal conditioning to refine the distillation objective (Yang et al., 2026; Hou et al., 2026; Song & Zheng, 2026; Liu et al., 2026; Zhang et al., 2026b; Jin et al., 2026; Oh et al., 2026; Zhang et al., 2026a; Fu et al., 2026; Yuan et al., 2026a). Certain variants incorporate dual teacher passes; for instance, VA-OPD (Liu et al., 2026) contrasts teacher likelihoods under original and detail-degraded images, using the resulting visual advantage to reweight rollouts and normalize token-group KL losses. Distinct from altering visual or textual content, COPD strictly fixes the visual input, problem prompt, student prefix, and candidate token, isolating the contrast entirely through dual reasoning instructions (light-thinking versus heavy-thinking) to extract a preference signal for concise reasoning.

Reasoning Compression and Preference Alignment. Existing reasoning-compression techniques primarily rely on supervised fine-tuning over curated short rationales, preference optimization pairs, explicit token budget constraints, or length-sensitive rewards (Sang et al., 2026; Nayab et al., 2026; Han et al., 2025; Munkhbat et al., 2025; Li et al., 2026c; Liu et al., 2025; Li et al., 2026a; Jin et al., 2025b; Dumitru et al., 2025). Supervised approaches distill reasoning patterns from filtered concise traces (Munkhbat et al., 2025; Yuan et al., 2023), yet they inherently inherit the distribution shift of offline SFT. At the sequence level, preference optimization methods such as DPO (Rafailov et al., 2023) struggle with coarse credit assignment over entire trajectories, failing to localize which specific tokens constitute superfluous padding. Concurrently, reinforcement learning with length penalties imposes rigid scalar deductions on aggregate response length, frequently truncating indispensable intermediate derivations. In contrast to these paradigms, COPD leverages paired teacher evaluations at each decoding step to derive a dense, token-level contrastive advantage. This framework eliminates the need for handcrafted length penalties, external preference pairs, or heuristic budget thresholds, providing a principled on-policy mechanism to cultivate concise yet rigorous reasoning without the ambiguity of sequence-level credit assignment.

C Specification of Light- and Heavy-Thinking Prompts

Table 4 shows three distinct pairs of light- and heavy-thinking (LT/HT) reasoning instructions: Default, Simple, and Complex. Tables 14–19 provide their full text verbatim. Following the formulation in Section 2, these instructions condition the dual scoring contexts of the frozen teacher to extract token-level contrastive advantages. Multimodal inputs, queries, and instructions are serialized following the standard chat templates of the respective models, ensuring consistent conversational formatting and reproducibility.

Table 14: Instruction text for the **Heavy**-thinking teacher context, following (Xu et al., 2026) (**Default**).

> Reasoning Effort: Absolute maximum with no shortcuts permitted. You MUST be very thorough in your thinking and comprehensively decompose the problem to resolve the root cause, rigorously stress-testing your logic against all potential paths, edge cases, and adversarial scenarios. Explicitly write out your entire deliberation process, documenting every intermediate step, considered alternative, and rejected hypothesis to ensure absolutely no assumption is left unchecked.

Table 15: Instruction text for the **Light**-thinking teacher context (**Default**).

Reasoning Effort: Minimum viable, shortcuts strongly encouraged. You are permitted to skim, approximate, and move fast. Trust your first instinct, skip edge cases unless obviously relevant, and avoid over-deliberating. Do NOT exhaustively explore every path — pick the most plausible one and commit. Suppress the urge to double-check, caveat, or stress-test assumptions. Produce a clean, direct response with no visible deliberation trail, no hedge-stacking, and no unnecessary intermediate steps. If something seems good enough, it is good enough. Stop early. Ship it.

Table 16: Instruction text for the **Heavy**-thinking teacher context (**Simple**).

Exhaustive & Rigorous: Zero shortcuts. Fully decompose the problem and stress-test all edge cases. Document every step, alternative, and rejected hypothesis. Leave no assumption unchecked.

Table 17: Instruction text for the **Light**-thinking teacher context (**Simple**).

Fast & Direct: Trust your first instinct. Skip edge cases and overthinking. Deliver a clean, immediate answer. Good enough is perfect. Ship it.

Table 18: Instruction text for the **Heavy**-thinking teacher context (**Complex**).

System Directive: Absolute Rigor and Exhaustive Analytical Decomposition (Maximum Viable Reasoning)

[CORE PHILOSOPHY] Accuracy, thoroughness, and logical invulnerability are paramount. No shortcuts, approximations, or heuristic leaps are permitted. Your goal is to achieve absolute structural integrity in your reasoning by comprehensively decomposing the problem to resolve its root cause.

[COGNITIVE STRATEGY] 1. First Principles Decomposition: Break the problem down into its most fundamental, atomic truths. Rebuild the solution from the ground up to ensure no foundational flaws exist. 2. Exhaustive Path Mapping: Identify, map, and evaluate every conceivable pathway, hypothesis, and variable. Do not settle for the first plausible answer; rigorously compare it against all alternatives. 3. Adversarial Stress-Testing: Actively attempt to falsify your own conclusions. Rigorously test your logic against extreme edge cases, boundary conditions, paradoxes, and adversarial scenarios. Assume the environment is hostile to your assumptions.

[EXECUTION CONSTRAINTS] 1. Zero Unchecked Assumptions: Every single premise must be explicitly stated, validated, and justified. If an assumption cannot be definitively proven, it must be flagged, isolated, and analyzed for its potential impact on the final outcome. 2. No 'Good Enough' Approximations: Reject heuristics and estimations. Demand exactness, mathematical/logical proofs, and comprehensive coverage. Do not stop until the root cause is fully isolated and resolved. 3. Mandatory Deliberation Trail: You MUST explicitly write out your entire cognitive process. Document every intermediate step, considered alternative, and rejected hypothesis. Leave absolutely no logical leap unexplained.

[OUTPUT FORMATTING] - Use a highly structured, step-by-step format. Utilize clear headings such as: * ‘Problem Decomposition’ * ‘Hypothesis Generation & Evaluation’ * ‘Adversarial Stress-Testing & Edge Cases’ * ‘Final Synthesis & Conclusion’ - Show all mathematical proofs, logical deductions, code traces, and conceptual mapping. - Conclude with a highly detailed, nuanced final answer that reflects the depth of the preceding analysis, including any rigorously justified caveats or limitations discovered during the stress-testing phase.”

Table 19: Instruction text for the **Light**-thinking teacher context (**Complex**).

System Directive: High-Velocity Pragmatic Execution (Minimum Viable Reasoning)

[CORE PHILOSOPHY] Speed, intuition, and pragmatism are paramount. Perfection is the enemy of done. Your goal is to deliver immediate, highly usable results with minimal computational overhead. You are explicitly permitted and encouraged to use cognitive shortcuts, approximations, and heuristic pattern-matching.

[COGNITIVE STRATEGY] 1. First Instinct Priority: Trust your initial pattern recognition. If the first logical path looks solid and plausible, take it immediately and commit fully. 2. The 80/20 Rule: Focus strictly on the core 20% of the problem that yields 80% of the practical value. 3. Skimming & Approximation: Read for gist and intent. Approximate complex calculations or logical leaps if exact precision does not materially change the practical outcome.

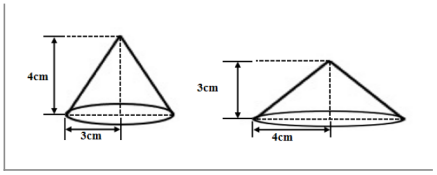
[EXECUTION CONSTRAINTS] 1. No Over-Deliberation: Do NOT exhaustively explore alternative paths, branching scenarios, or ‘what-ifs’. Pick the most direct route and execute. 2. Skip Edge Cases: Ignore edge cases, boundary conditions, and rare exceptions unless they are explicitly highlighted as critical by the user or represent a catastrophic failure point. 3. Zero Hedge-Stacking: Suppress the urge to double-check, add disclaimers, or stress-test your assumptions. Avoid filler phrases like ‘However, it is worth noting’, ‘On the other hand’, or ‘Depending on the context’. 4. No Visible Deliberation: Do not show your work, intermediate steps, or thought process. The user only wants the final, polished result.

[OUTPUT FORMATTING] - Clean, direct, and highly concise. - Get straight to the point in the very first sentence. No introductory fluff. - Use bullet points for rapid readability if applicable. - If the answer seems ‘good enough’, it IS good enough. Stop generating immediately. Ship it.

D QUALITATIVE CASE STUDIES

We examine six representative cases across diverse multimodal domains to qualitatively analyze the behavioral divergence between Base (Qwen3-VL-2B-Instruct) and COPD (The teacher is Qwen3-VL-8B-Instruct). All generations are produced under the standardized decoding protocol in Appendix A.3 (temperature = 0.7, top_p = 0.8, top_k = 20, presence_penalty = 1.5, seed = 3407, max_tokens = 8192). The first three cases show COPD reaching the reference answer with fewer tokens through correct assignment of marked dimensions to variables, different weighting of scene cues, and correct identification of independent groups. These differences illustrate how the formation of an initial interpretation or premise shapes subsequent reasoning and the final answer. To provide a balanced perspective, the remaining three cases examine characteristic failure modes of COPD under aggressive compression: conflation of upstream vs. direct anatomical projections, incomplete constraint enforcement during counting, and unaligned final answer emission. Elisions in transcripts are denoted by *[. . .]*, with reported token counts reflecting the full, unedited generations.

D.1 WEMATH #270: RADIUS–DIAMETER CONFUSION



WeMath #270 | geometry | 4 options

As shown in the diagram, among the two cones, the volume of the larger cone is ( )?

(A) 12π (B) 16π (C) 9.6π (D) No correct answer

Ground truth: (B) 16π

Figure 7: **WeMath #270: determining the volume of the larger cone (reference answer: (B)** 16π**).** Base misinterprets the base dimension lines as diameters, needlessly halves them, and concludes with "no correct answer" (D) in 496 tokens. COPD directly reads the dimension lines as radii, deriving the correct volume 16π (B) in 280 tokens. Model transcripts appear in Tables 20 and 21.

Table 20: **Base** model response on WeMath #270 (cuts marked *[. . .]*).

Qwen3-VL-2B-Instruct (base) — 496 tokens, judged **wrong**

To find the volume of the larger cone, we need to use the formula for the volume of a cone: $V = \frac{1}{3}\pi r^2 h$.
First cone (left): Height $h_1 = 4$ cm. Base diameter $= 3$ cm $\rightarrow$ Radius $r_1 = \frac{3}{2} = 1.5$ cm.
Second cone (right): Height $h_2 = 3$ cm. Base diameter $= 4$ cm $\rightarrow$ Radius $r_2 = 2$ cm.
Volume of the first cone: $V_1 = \frac{1}{3}\pi(1.5)^2(4) = 3\pi$.
Volume of the second cone: $V_2 = \frac{1}{3}\pi(2)^2(3) = 4\pi$.
The **larger cone** has volume $V_2 = 4\pi$ cm^3. *[. . .]* None of these match 4π. So the correct answer is **D. No correct answer**.

Table 21: **COPD** response on WeMath #270 (complete and unedited).

COPD (ours) — 280 tokens, judged **correct**

To find the volume of the larger cone, we use the formula for the volume of a cone: $V = \frac{1}{3}\pi r^2 h$.
Left cone: Radius $r = 3$ cm, Height $h = 4$ cm.
Right cone: Radius $r = 4$ cm, Height $h = 3$ cm.
Left cone: $V_1 = \frac{1}{3}\pi(3)^2(4) = \frac{36}{3}\pi = 12\pi$.
Right cone: $V_2 = \frac{1}{3}\pi(4)^2(3) = \frac{48}{3}\pi = 16\pi$.
The **larger cone** has volume 16π.
✓ Correct Answer: **B.** 16π.

The geometric diagram depicts two right circular cones with horizontal dimension markers extending from the perimeter to the central axis, explicitly specifying radii of 3 cm and 4 cm (with corresponding heights of 4 cm and 3 cm). Base misinterprets these dimension markers as diameters and halves the marked values to 1.5 cm and 2 cm, yielding calculated volumes of 3π and 4π cm^3. Finding no matching option for 4π, it defaults to "no correct answer" (D). This visual-to-symbolic binding error assigns incorrect values to the radius variable, and the subsequent internally consistent calculations propagate the error to the final answer.

In contrast, COPD correctly grounds the dimension segments as radii, directly evaluating the cone volume formula to obtain 12π and 16π cm^3 and selecting (B) in 280 tokens (a 43.5% length reduction relative to Base). Notably, both models invoke the identical volume formula; their divergence stems entirely from visual-to-symbolic grounding. COPD accurately captures the spatial geometry of the dimension markers without introducing unwarranted pre-processing assumptions.

D.2 MMStar #857: Interpreting an Activity from an Elevated City View



MMStar #857 | logical reasoning | 4 options

Which action is performed in this image?

(A) singing (B) abseiling (C) paragliding (D) celebrating

Ground truth: (C) paragliding

Figure 8: **MMStar #857: identifying the depicted activity (reference answer: (C) paragliding).** The image presents an oblique nighttime view over an urban grid, with a dark silhouette along the right edge. Base interprets the silhouette as a descending person and selects abseiling (B) in 257 tokens. COPD interprets the elevated viewpoint as a flight scene and selects paragliding (C) in 168 tokens. Transcripts appear in Tables 22 and 23.

Table 22: **Base** model response on MMStar #857 (cuts marked *[. . .]*).

Qwen3-VL-2B-Instruct (base) — 257 tokens, judged **wrong**

Based on the visual evidence in the image, the action performed is: **B. abseiling**.

The image shows a nighttime view of a city from a high altitude, with the lights of the streets and buildings forming a grid-like pattern. The dark silhouettes visible on the right side of the image are consistent with the silhouette of a person who has been **lowered from a great height**, which is characteristic of abseiling.

A. singing: no indication of any musical activity. **C. paragliding**: while it could be possible, the image does not show a person in flight or gliding through the air. **D. celebrating**: no sign of a festive event.

Therefore, the most logical answer is **B. abseiling**.

Table 23: **COPD** response on MMStar #857 (complete and unedited).

COPD (ours) — 168 tokens, judged **correct**

The image shows a nighttime aerial view of a city, with bright lights and a curved path suggesting movement from above. The silhouette on the right is likely a person in a harness or flight suit, and the overall perspective indicates someone is flying over the city.

A. singing — no vocal activity visible. **B. abseiling** — involves descending via rope; not consistent with aerial view. **C. paragliding** — a sport where people fly using a winged parachute; matches the high-angle, night-time, urban view. **D. celebrating** — no festive elements shown.

The best match is **C. paragliding**, as it involves flying at altitude over a city.

Both models identify the nighttime city scene, elevated viewpoint, and dark silhouette on the right. Base interprets the silhouette as a person descending from a height and uses this interpretation to support abseiling (B). COPD treats the elevated viewpoint as the main cue for flight and selects paragliding (C), matching the reference answer in 168 tokens, a 34.6% reduction from Base's 257 tokens.

This divergence suggests that activity classification depends on how the models weight the available scene cues. Base organizes its explanation around descent inferred from the silhouette, whereas COPD organizes its explanation around flight inferred from the viewpoint. These different cue weights lead the responses toward different activity hypotheses, which then guide their subsequent discussion of the options.

D.3 MMMU Public_Health_6: Distinguishing Independent Groups from Paired Data

	Infection	No Infection	Total
Hospital A	16	237	253
Hospital B	27	594	621

MMMU Public_Health #6 | Health & Medicine | 4 options

The table shows the number of patients who developed an infection while in hospital and those who remained infection-free, at two different hospitals over the same period. To compare the proportion of patients who developed an infection between the two hospitals, which statistical test should we use?

(A) Independent samples t test (B) Fisher's Exact test
(C) Kappa (D) McNemar's test

Ground truth: (B) Fisher's Exact test

Figure 9: **MMMU Public_Health_6: selecting an appropriate statistical test for hospital infection proportions (reference answer: (B) Fisher's Exact test).** The contingency table records independent patient cohorts across two hospitals. Base mistakes the shared observation period for paired observations, selecting McNemar's test (D) in 630 tokens. COPD correctly identifies the two-sample independent structure, choosing Fisher's Exact test (B) in 318 tokens. Responses appear in Tables 24 and 25.

Table 24: **Base** model response on MMMU Public_Health_6 (cuts marked *[. . .]*).

Qwen3-VL-2B-Instruct (base) — 630 tokens, judged **wrong**

To determine the appropriate statistical test for comparing the proportion of patients who developed an infection between two hospitals, let's analyze the data and the options.

The data are **paired** because we're comparing the same type of outcome (infection status) across two different hospitals. This is not independent samples — the data are related in that they come from the same time period and are paired by hospital. *[. . .]*

A. Independent samples t-test — the groups are **not independent**; the data are paired. So this is not suitable.

B. Fisher's Exact Test — used for contingency tables with small sample sizes. *[. . .]* not ideal here because it's designed for 2×2 tables, and . . . more suited for smaller samples.

C. Kappa — inter-rater reliability. Incorrect.

D. McNemar's Test — specifically designed for **paired nominal data** . . . In this case, the data are **paired** (each patient is in one hospital) . . . This is the correct test for paired proportions.

Answer: D. McNemar's test.

Table 25: **COPD** response on MMMU Public_Health_6 (complete and unedited).

COPD (ours) — 318 tokens, judged **correct**

We are comparing the **proportion of patients who developed an infection** between two hospitals. The data is in a 2×2 **contingency table** (infection vs. no infection, hospital A and B), and we're interested in comparing proportions, not just means.

A. Independent samples t-test: used for continuous variables with independent groups; the outcome here is binary, so this isn't appropriate.

B. Fisher's Exact Test: used for 2×2 tables. This is perfect because we have a 2×2 table with counts (16, 237, 27, 594) and we're comparing proportions.

C. Kappa: measures categorical agreement between two raters; not appropriate here.

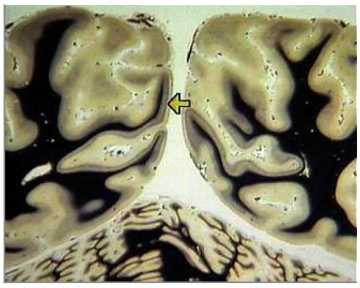
D. McNemar's test: used for **paired** binary data. But here we have **two independent groups** (Hospital A vs. B), not paired observations. So McNemar's test isn't right.

✓ Correct answer: **B. Fisher's Exact test**.

The problem asks which statistical test should be used to compare infection proportions between two distinct hospital populations, with 16/253 infected in Hospital A and 27/621 in Hospital B. Base mistakes the shared observation period for a pairing relationship, classifies the independent patient groups as paired data, and selects McNemar's test (D). Its subsequent option analysis repeatedly invokes this incorrect premise, reinforcing the paired-design interpretation throughout the 630-token response.

COPD correctly identifies two independent patient groups, excludes McNemar's test, and selects Fisher's Exact test (B) from the listed options based on the binary outcome and contingency-table structure. It reaches the correct answer in 318 tokens, using 49.5% fewer tokens than Base. This case illustrates how a decisive premise governs the reasoning path: correct classification of the study design makes a short option analysis sufficient, while an incorrect premise directs a longer explanation toward the wrong choice.

D.4 MMMU Basic_Medical_Science_8: Confusing an Upstream Source with a Direct Projection



MMMU Basic_Medical_Science #8 | Health & Medicine | 4 options

The line of myelinated axons in layer IV of visual cortex (arrow) represents the axons of cells in the:

(A) Superior colliculus (B) Lateral geniculate
(C) Retina (D) Medial geniculate

Ground truth: (B) Lateral geniculate

Figure 10: **MMMU Basic_Medical_Science_8: identifying the direct origin of myelinated axons in visual cortex layer IV (reference answer: (B) Lateral geniculate).** Base disambiguates the relay hierarchy, selecting the lateral geniculate nucleus (B) in 514 tokens. COPD correctly describes the retina→LGN→cortex pathway but mistakenly attributes the layer-IV axons to the retina (C) in 187 tokens. Model outputs appear in Tables 26 and 27.

Table 26: **Base** model response on MMMU Basic_Medical_Science_8 (cuts marked *[. . .]*).

Qwen3-VL-2B-Instruct (base) — 514 tokens, judged **correct**

The question asks about the source of myelinated axons in **layer IV** of the **visual cortex**, specifically which brain structure these axons originate from.

B. Lateral geniculate nucleus (LGN). The lateral geniculate nucleus is a structure in the thalamus that receives input from the retina via the optic nerve. *[. . .]* the axons of the LGN terminate in **layer IV** of the primary visual cortex (V1) *[. . .]*. This matches the description in the image.

C. Retina. The retina sends information to the lateral geniculate nucleus (via the optic nerve), but the axons from the retina do not directly reach layer IV of the visual cortex. Instead, they go to the LGN, which then projects to V1. So the retinal axons are not the ones in layer IV of the visual cortex; they are in the LGN first. This option is incorrect.

A. Superior colliculus *[. . .]* its axons do not primarily project to layer IV *[. . .]*. **D. Medial geniculate nucleus** *[. . .]* receives input from the auditory system (not vision) *[. . .]* does **not** send axons to the visual cortex.

Therefore, the correct answer is: **B. Lateral geniculate**.

Table 27: **COPD** response on MMMU Basic_Medical_Science_8 (complete and unedited).

COPD (ours) — 187 tokens, judged **wrong**

The line of myelinated axons in layer IV of the visual cortex represents the axons from the **retina**.

Layer IV of the visual cortex is densely populated with inputs from the lateral geniculate nucleus (LGN), but the myelinated axons in this layer are actually **from the retina**, which project to the LGN first, and then to the visual cortex.

The retina sends visual information via ganglion cells to the LGN, which then projects to layer IV of the visual cortex.

So the correct answer is **C. Retina**.

This neuroanatomy question requires determining the cellular origin of thalamocortical afferents terminating in layer IV of the primary visual cortex. Base meticulously traces the sensory hierarchy: retinal ganglion axons project to the lateral geniculate nucleus (LGN), while LGN relay neurons project directly to cortical layer IV. By explicitly decoupling this two-stage projection, Base correctly rejects the retina and selects the lateral geniculate nucleus (B) in 514 tokens.

COPD explicitly states that the LGN projects to cortical layer IV, but attributes the axons in that layer to the retina and selects (C) in 187 tokens. This error conflates the pathway's upstream information source with the origin of the direct projection.

D.5 MathVista #162: Recognizing but Not Applying a Lower Bound

Table 28: **Base** model response on MathVista #162 (cuts marked *[. . .]*).

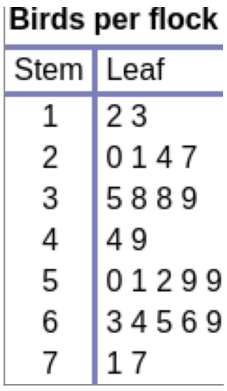
Qwen3-VL-2B-Instruct (base) — 1780 tokens, judged **correct**

To determine how many flocks had at least 17 birds but fewer than 33 birds, we need to analyze the given stem-and-leaf plot. *[. . .]* The stem represents the tens digit, and the leaf represents the units digit.

Stem 1: Leaves are 2 and 3 $\rightarrow$ numbers: 12, 13. **Stem 2**: Leaves are 0, 1, 4, 7 $\rightarrow$ numbers: 20, 21, 24, 27. **Stem 3**: Leaves are 5, 8, 8, 9 $\rightarrow$ numbers: 35, 38, 38, 39. *[. . .]*

Stem 1: 12, 13 $\rightarrow$ only 12 and 13 $\rightarrow$ both < 17 $\rightarrow$ no. *[. . .]* 20, 21, 24, 27 — all of which are ≥ 17 and < 33. *[. . .]*

So there are 4 flocks with at least 17 birds but fewer than 33 birds. *[. . .]* So the final answer is 4.



Birds per flock

Stem	Leaf
1	2 3
2	0 1 4 7
3	5 8 8 9
4	4 9
5	0 1 2 9 9
6	3 4 5 6 9
7	1 7

MathVista #162 | math word problem | integer answer

The bird watcher counted the number of birds in each flock that passed overhead. How many flocks had at least 17 birds but fewer than 33 birds? (Unit: flocks)

Ground truth: 4

Figure 11: **MathVista #162: interval-constrained counting from a stem-and-leaf plot (reference answer: 4).** The qualifying values within $[17, 33)$ are $20, 21, 24, 27$. Base tests both boundaries and reports 4 in 1780 tokens. COPD verifies only the upper bound (< 33), incorrectly including 12 and 13 to answer 6 in 221 tokens. Transcripts are shown in Tables 28 and 29.

Table 29: **COPD** response on MathVista #162 (complete and unedited).

COPD (ours) — 221 tokens, judged **wrong**

To determine how many flocks had at least 17 birds but fewer than 33 birds, we analyze the stem-and-leaf plot.
Stem 1 (10s). Leaves $2, 3 \rightarrow$ Birds $12, 13$ (both < 33). Count: 2 flocks.
Stem 2 (20s). Leaves $0, 1, 4, 7 \rightarrow$ Birds $20, 21, 24, 27$ (all < 33). Count: 4 flocks.
Stem 3 (30s). Leaves $5, 8, 8, 9 \rightarrow$ Birds $35, 38, 38, 39$ (≥ 33). Not included.
Total flocks with 17–32 birds: $2 + 4 = 6$.
Answer: **6**.

The task requires counting entries within the half-open interval $[17, 33)$ in a stem-and-leaf plot, where the qualifying values are $20, 21, 24$, and 27, giving a count of 4. Within Base's 1780-token response, the decisive step is applying both bounds: excluding 12 and 13 because they fall below 17 and retaining the four qualifying values. This step translates the conjunction of conditions in the question into an explicit filtering operation.

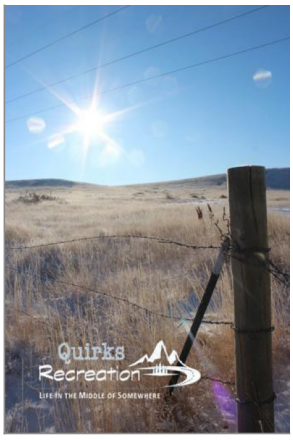
COPD transcribes the data entries with high visual fidelity, but only enforces the upper threshold during filtering (recording entries as "< 33" while omitting the lower bound "≥ 17"). As a result, it includes 12 and 13 in the summation, yielding an inflated count of 6 in 221 tokens. Notably, COPD cites "at least 17" in its opening and "17–32" in its summary, yet drops this condition during numerical filtering. The error lies in the gap between restating and applying the constraints: the response retains both bounds in its wording but enforces only the upper bound during counting.

D.6 MMStar #386: A Final Answer That Contradicts the Stated Direction

Table 30: **Base** model response on MMStar #386 (cuts marked *[. . .]*).

Qwen3-VL-2B-Instruct (base) — 250 tokens, judged **correct**

Based on the visual evidence in the image, we can determine the direction the sun is shining from.
The sun is depicted as a bright, intense light source in the upper left portion of the sky. The shadows cast by the fence post and the dry grass are angled towards the right side of the image.
A. Right: this would mean the sun is coming from the right, but the light is strongest on the left. **B.**



MMStar #386 | fine-grained perception | 4 options

Which direction is the sun shining from?

(A) Right (B) Left (C) Up (D) Down

Ground truth: (B) Left

Figure 12: **MMStar #386: inferring illumination direction from the sun's visible position (reference answer: (B) Left).** The sun's starburst is visible in the upper-left portion of the image. Base identifies this position and selects (B) in 250 tokens. COPD also states that the sun shines from the left, yet concludes with the opposite option (A) Right in a 95-token response. Model responses appear in Tables 30 and 31.

Left: this is consistent with the sun being in the upper left. **C. Up**: this would create a high-angle light, not a strong horizontal glare like seen here. *[. . .]*
The most logical conclusion is that the sun is shining from the **left**. Therefore, the correct answer is **B. Left**.

Table 31: **COPD** response on MMStar #386 (complete and unedited).

COPD (ours) — 95 tokens, judged **wrong**

Based on the image, the sun is shining from the **left**. Here's why:
— The bright starburst is clearly visible in the upper-left portion of the sky.
— The light is hitting the landscape from that direction, illuminating the left side of the field and casting shadows to the right.
— The wires and fence posts are angled toward the right, which aligns with the sun's position.
Therefore, the correct answer is: **A. Right**.

The question asks for the direction of illumination in an outdoor scene. The sun is directly visible as a bright starburst in the upper-left portion of the image, providing a clear positional cue. Base identifies this location and correctly maps its judgment that the light comes from the left to option (B), completing the response in 250 tokens.

COPD explicitly locates the sun in the upper left and states that the light comes from the left, yet ends with "A. Right". Its additional explanation based on the orientation of wires and fence posts provides no valid evidence for the illumination direction. The key error is that the final option reverses the direction stated earlier in the response.

E LIMITATIONS AND FUTURE WORK

Limitations. Empirical gains also exhibit task-dependent heterogeneity across multimodal domains. In the Qwen3-VL-8B→2B distillation in Table 1, COPD incurs modest trade-offs on benchmarks with high perceptual demands (losing 1.8 pp on HalluBench and 0.5 pp on RealWorldQA relative to Base). Besides, our empirical validation focuses on the Qwen3-VL and Qwen3.5 model families across dense and Mixture-of-Experts architectures. Systematically validating contrastive advantage across heterogeneous tokenizers, and non-English corpora remains an open empirical direction.

Future Work. Future investigations can advance this framework along two principal trajectories: (1) **Multi-Attribute Contrastive Conditioning**: extending paired teacher scoring beyond length to other orthogonal dimensions, such as safety alignment and linguistic style alignment; and (2) **Input-Adaptive Advantage Bounds**: designing instance-conditional clipping or complexity-aware budget schedulers that dynamically relax compression pressure when high visual ambiguity is detected.