MIRROR: LEARNING FROM THE OTHER VIEW FOR MULTI-MODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can behave very differently depending on whether the same underlying problem is presented textually or visually. This inconsistency suggests that different views expose different reasoning paths and failure modes that standard multimodal post-training does not fully exploit. To study and exploit this phenomenon, we construct ODA-Data, a high-quality paired multimodal geometry dataset with text-dominant, image-dominant, and combined image+text views of the same problems, together with splits for training and evaluating view-dependent reasoning behaviors. We then develop **Modality-Informed Reciprocal Reasoning Optimization** (`MIRROR`), a reinforcement learning approach that derives a cross-view self-distillation signal from the model's own modality-dependent capabilities. For each problem, `MIRROR` evaluates the model under all views, selects the best-performing view as a teacher, and trains the remaining views using a GRPO task-reward objective augmented with a reverse-KL regularizer toward that teacher. Across reasoning benchmarks that evaluate on multimodal mathematics problems, `MIRROR` improves over standard RL and yields more accurate and consistent behavior across different input modality representations.

1 INTRODUCTION

Vision-language models have made substantial progress on multimodal reasoning (Chen et al., 2024; Bai et al., 2023), but their behavior remains highly sensitive to how the information is presented (Li et al., 2025; Ismithdeen et al., 2025). As we also find in our experiments, a model may solve a problem when the relevant information is written in text, yet fail when those relations must be recovered from an image; conversely, an image may make geometric structure salient in a way that a purely textual description does not. This inconsistency is especially concerning because the underlying reasoning problem has not changed, only the input view has. It suggests that current VLMs are not merely limited by problem difficulty (Yao et al., 2025), but also by view-dependent failures in grounding and reasoning. Our goal is to study and address these failures.

To do so, we focus on geometry problems that serve as a natural testbed for studying this phenomenon (Zhang et al., 2024b). Many geometry problems can be presented through several views associated with the same underlying problem: a *text-dominant* view, where quantities and relations are enumerated in language; an *image-dominant* view, where the image carries essential geometric information and the accompanying text is minimal and mainly specifies the question; and a *combined* image+text view, where both sources are available. Since they share the same problem identity and answer target, they enable a controlled comparison of view-dependent reasoning. As we find in our initial study (Table 1), when the same model is asked to solve the same underlying problem, one input view can yield correct reasoning substantially more reliably than another. Such view-dependent behavior is not merely a robustness failure; it can also provide a useful supervision signal.

Prior work on multimodal reasoning often improves models by using external teachers (Liu et al., 2023), curated rationales (Zhang et al., 2024c), or specialized training data that teaches the model how to combine visual and textual inputs (Dong et al., 2025; Bousselham et al., 2025). However, such supervision may not be available for every target setting, and distillation from external teachers can be fragile under distribution mismatch (Li et al., 2026). In contrast, we study whether paired views of the same problem can provide a form of self-supervision: when the model performs substantially better under one view than another, can the stronger view help improve the weaker one?

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.

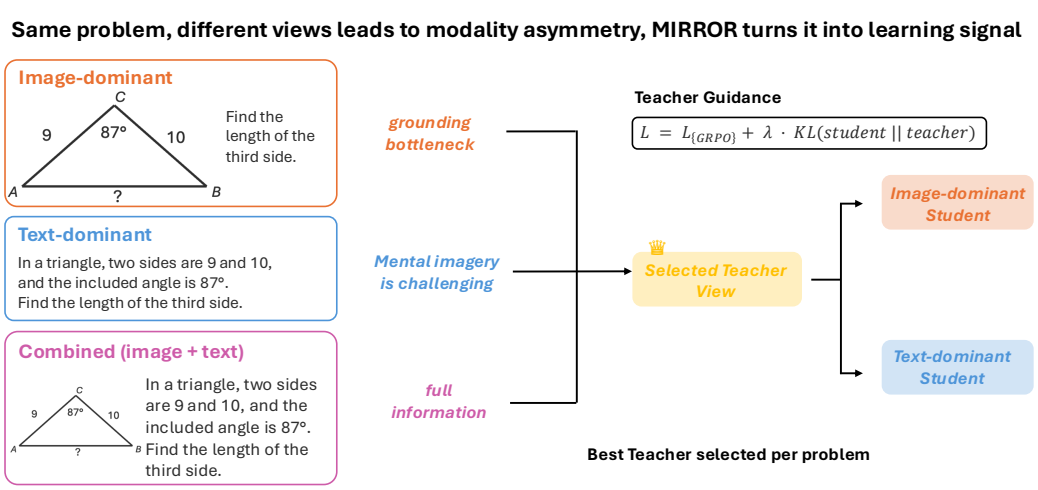


Figure 1: **Modality-Informed Reciprocal Reasoning Optimization (MIRROR).** MIRROR exploits view asymmetry by selecting the strongest-performing view of each problem as a teacher and regularizes weaker student views towards the teacher distribution, improving performance without external supervision.

In this paper, we propose MIRROR (**Modality-Informed Reciprocal Reasoning Optimization**), a training framework that uses reasoning from stronger views of a problem to improve reasoning under weaker views. For each problem, we construct text-dominant, image-dominant, and combined image+text views (Figure 1), evaluate the current model under each view, and select the strongest-performing view as a problem-specific teacher. The student is trained under either text-dominant or image-dominant which are restricted views with no information redundancy. The challenge is to transfer the teacher signal without forcing the student to imitate trajectories that depend on information unavailable from its own view. MIRROR addresses this with on-policy distillation: the student samples rollouts from its own view, while the selected teacher view only rescores the same student-generated trajectories. This on-policy distillation provides dense, token-level teacher guidance on states actually visited by the student. We combine it with outcome-reward RL because the selected teacher view is only relatively stronger and may still be imperfect; the outcome reward anchors learning to final-answer correctness rather than teacher imitation alone.

To instantiate our approach, we construct ODA-Data, a paired geometry dataset with text-dominant, image-dominant, and combined image+text views of the same underlying problems. ODA-Data preserves problem identity across views, enabling us to measure both accuracy and cross-view consistency. We use ODA-Train for post-training and ODA-Val as a held-out benchmark to measure view-dependent solvability: whether a model solves the same problem consistently when presented through different views. Empirically, MIRROR outperforms both post-training on a mixture of prompts and recent open-source post-trained VLMs. On ODA-Val, MIRROR achieves the best performance across all image and text validation metrics, improving image pass@16 by 14.49 percentage points (pp) and text pass@16 by 5.88 pp over the base model. These gains transfer to external benchmarks: on GeoInt, MIRROR improves over the base model by 7.59 pp in pass@16, and on MathVerse, it improves the mean score by 5.56 points. Beyond accuracy, on ODA-Train, the fraction of problems solved under both image-dominant and text-dominant views increases from 42.5% to 60.7% with MIRROR training. MIRROR also expands the set of solvable examples, achieving the largest net solvability gain on ODA-Data at +8.95 pp. These results show that paired views provide an effective self-supervision signal for stronger and more consistent multimodal reasoning.

2 Related Work

Multimodal mathematical reasoning benchmarks and data scaling. Prior work shows that specialized multimodal math data can substantially improve MLLM reasoning. For example, Math-LLaVA fine-tunes LLaVA-1.5 on MathV360K (Shi et al., 2024), and MultiMath-7B trains on MultiMath-300K with image captions and step-by-step solutions (Peng et al., 2024). Several benchmarks have also been introduced for visual mathematical reasoning, including MATH-Vision (Wang et al., 2024),

MathVerse (Zhang et al., 2024b), and geometry-focused benchmarks such as LeanGeo-Bench (Song et al., 2025), GeoInt (Wei et al., 2025b), and GeoGoal (Chen et al., 2026a). More recently, CROSS-MATH studies modality gaps using matched text-dominant, image-dominant, and the combined image+text versions of synthetic puzzles (Xu et al., 2026). These works motivate our focus on geometry and provide evaluation settings that we build on, including MathVerse and GeoInt. Instead of focusing on scaling data or constructing benchmarks, our goal is algorithmic: how paired views of the same problem can be used during RL training to transfer reasoning behavior across modalities?

RL and cross-modal transfer for multimodal models. Recent work has explored RL and cross-modal transfer for multimodal reasoning. Vision-R1 uses synthetic chain-of-thought data and progressive RL (Huang et al., 2026); R1-OneVision converts images into textual descriptions before SFT and RL (Yang et al., 2025b); and FAST uses difficulty-aware rewards with KL-regularized GRPO (Diao et al., 2026). Closest to our setting, VOLD distills from a stronger text-only LLM teacher into a VLM student using on-policy distillation and GRPO (Bousselham et al., 2025). These methods show that stronger teachers can improve multimodal reasoning. In contrast, we target *self-supervised* improvement: paired views of the same problem reveal cross-modal performance asymmetries that can serve as supervision for improving reasoning across modalities.

On-policy Distillation. On-policy distillation has emerged as a way to learn from a teacher while reducing the distribution mismatch of standard behavioral cloning: instead of training only on teacher-generated traces, the student samples its own rollouts and receives token-level guidance from a teacher on the states it actually visits (Agarwal et al., 2024; Gu et al., 2026). More broadly, prior work on self-improvement has trained language-model agents to critique and iteratively refine their own responses through recursive self-feedback (Qu et al., 2024). On-policy self-distillation provides a complementary mechanism for internal improvement, including self-distillation from demonstrations (Shenfeld et al., 2026), feedback-conditioned self-teachers (Hübotter et al., 2026), and privileged-context teachers (Zhao et al., 2026). Recent multi-teacher or multi-model extensions (Chen et al., 2026b; Yang & Thomason, 2025; Yang et al., 2025a) use specialized teachers, debate (Yang et al., 2026a), or calibrated teacher signals to merge complementary capabilities into a single student (Yang et al., 2026b; Team et al., 2026; Hou et al., 2026; Wang et al., 2026). `MIRROR` differs in both the source and structure of the teacher signal: rather than distilling from an external stronger model, a fixed teacher pool, or privileged solution traces, it selects the strongest on-policy generations for a certain view of the same problem as a problem-specific teacher, building a self-supervised approach for improving model.

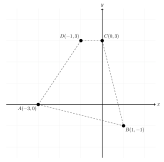
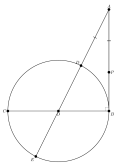
Explicit intermediate structure for visual reasoning. A complementary line of work improves visual reasoning by either actively selecting informative visual evidence or introducing intermediate structures between perception and reasoning. Act2See learns to adaptively select and process visual observations for more effective video reasoning (Ma et al., 2026). For visual mathematical reasoning, MathCoder-VL uses image-to-code supervision (Wang et al., 2025a). Zhang et al. (2025) introduces a geometry-specialized vision encoder for recognizing visual primitives. Related work on plane-geometry formalization (Cui & Liu, 2025) similarly exposes structured representations. These approaches introduce inductive biases through active perception, code, symbolic descriptions, formal proofs, or geometry priors. In contrast, `MIRROR` retains the original input views and remains end-to-end: rather than explicitly selecting evidence or constructing intermediate representations, it encourages the model to internalize transferable reasoning structure through cross-modal supervision.

3 Understanding Differences in Reasoning Across Modalities

Before introducing our approach, we first analyze reasoning behaviors of the base model `Qwen3-VL-4B-Instruct` (Bai et al., 2025) when solving the same underlying problem from different input views. In this section, we characterize how do the model's reasoning trajectories differ when conditioned on different modalities. We further study if access to both text and images reveal supervision signals that are hidden from any single modality?

To formalize this setting, we study post-training of a multimodal language model, denoted by π_{base} with parameters θ, on problem instances that may be presented with different input views. For a problem $\mathbf{x} \sim \rho$, let $\mathcal{M}(\mathbf{x})$ denote the set of available views sharing the same underlying problem content and answer target. In this work, we focus on two views: a text-dominant view $\mathbf{x}^{\text{text}}$ and an image-dominant view $\mathbf{x}^{\text{img}}$. We can also readily construct a combined view $\mathbf{x}^{comb} := (\mathbf{x}^{\text{text}}, \mathbf{x}^{\text{img}})$,

Table 1: **Representative modality-dependent reasoning behaviors.** Each column illustrates one source of cross-modal supervision; gray highlighting marks text omitted from the image-dominant query because the corresponding information is conveyed by the diagram. Full prompts and completions are in Appendix Ks.

	Example 1: Image → Text	**Example 2:** Text → Image	**Example 3:** Image+Text → Image
Image			
Query	Given four fixed points $A(-3,0)$, $B(1,-1)$, $C(0,3)$, $D(-1,3)$ and a moving point P, find $\min_P \lvert PA\rvert + \lvert PB\rvert + \lvert PC\rvert + \lvert PD\rvert$.	In a 13-14-15 triangle, three semicircles are drawn with diameters on the sides and tangent to the other two sides. The product of their areas equals the cube of a sphere's surface area. If the sphere volume is $\frac{a\pi\sqrt{b}}{c}$, find $a+b+c$.	In circle O, BC is a diameter, $AB \perp BC$, $ADOE$ is a line, $AP = AD$, and $AB = 2R$. Which option is correct? (A) $AP^2 = PB \cdot AB$. (B) $AP \cdot DO = PB \cdot AD$. (C) $AB^2 = AD \cdot DE$. (D) $AB \cdot AD = OB \cdot AO$. (E) None.
Before training	**Image-dominant:** solves correctly with $3\sqrt{2}+2\sqrt{5}$. **Text-dominant:** performs exhaustive search of possible locations of P and gives a rough estimate 8.74. Actually, P at the intersection of AC and BD minimizes the desired quantity.	**Text-dominant:** solves correctly with 873. **Image-dominant:** misreads each semicircle's diameter as exactly the length of each side of the triangle. The model wrongly predicts 3249. Actually, each diameter is shorter than the side.	**Combined image+text:** uses the textual relation that explicitly stated $AP = AD$ and correctly chooses A. **Image-dominant:** misses $AP = AD$ condition from the diagram and wrongly chooses B.
After training	**Text-dominant:** recovers the structural solution and predicts $3\sqrt{2}+2\sqrt{5}$.	**Image-dominant:** avoids the visual grounding error and predicts 873.	**Image-dominant:** notes the tick marks indicating $AD = AP$ and now chooses A.
Takeaway	The diagram makes the geometric structure salient, so image-dominant reasoning supervises text-dominant reasoning.	The text removes a perceptual bottleneck, so text-dominant reasoning supervises image-dominant reasoning.	Combined image+text view leverages explicit textual descriptions together with the visual structure, removing ambiguity. So it supervises image-dominant reasoning.

which contains both views simultaneously. For any input view $\tilde{\mathbf{x}} \in \mathcal{M}(\mathbf{x})$, a rollout $\mathbf{y} \sim \pi(\cdot \mid \tilde{\mathbf{x}})$ attempts to solve the problem. As in standard outcome-reward RL, we assume that correctness can be evaluated from the final answer and define a binary reward $r(\mathbf{x}, \mathbf{y}) \in \{0, 1\}$. We study performance using pass@k (Chen et al., 2021) under each modality.

To study view-dependent reasoning, we construct **ODA-Data**, a geometry dataset derived from ODA-Math-460k (Cai et al., 2025). Each example contains two views of the same problem: a "text-dominant" view and an "image-dominant" view. In our analysis and training, we also consider a third combined image+text view, which is constructed on the fly by concatenating the text-dominant prompt with the image from the image-dominant view. Starting from the pure text dataset ODA-Math-460k, we first exclude easy problems whose original-dataset annotations report a pass rate of 1. For each remaining problem, we use Gemini-3-Pro-Preview (Google DeepMind, 2025) to generate and verify a diagram in TikZ, a LaTeX-based language for programmatic scientific graphics, and rewrite the image-dominant prompt so that it no longer contains information present in the diagram. We choose TikZ over direct image generation because preliminary human inspection indicated that the model generated usable TikZ code more reliably and that the compiled diagrams had higher visual quality than directly generated images. In this process, diagrams that do not accurately reflect the question or contain answer leakage are removed. To construct a challenging dataset in which every problem is difficult under at least one view, we generate 32 rollouts per view with Qwen3-VL-4B-Instruct (Bai et al., 2025) and retain problems for which exactly one view has at most one successful rollout, yielding approximately 2K curated examples. See Appendix I for full details. The dataset is randomly split 85:15 into **ODA-Train** for post-training and **ODA-Val** for evaluation. Because ODA-Data is filtered using this view-asymmetric failure criterion, ODA-Val serves as a diagnostic benchmark for studying cross-view performance differences and transfer.

We compare reasoning traces induced by the text-dominant, image-dominant, and combined views. Although these views share the same problem and answer target, they expose different information and can induce different reasoning trajectories. Text-conditioned reasoning can use explicitly stated

relations directly, whereas image-conditioned reasoning must first recover them from the diagram, creating a potential grounding bottleneck. For example, in the 13-14-15 triangle problem in Table 1, the text-conditioned trace correctly derives 873, whereas the image-conditioned trace incorrectly treats each semicircle's diameter as an entire triangle side and predicts 3249. Conversely, diagrams can make geometric structure salient when text-only reasoning fails, while the combined view can help when neither restricted view is fully reliable. These complementary strengths motivate adaptive teacher selection across the three views.

Quantitative analysis. We next quantify these view-dependent differences over ODA-Train using `Qwen3-VL-4B-Instruct` and examine pass@1 metric. At the aggregate level, text-dominant achieves 29.58%, image-dominant achieves 10.54%, and combined image+text prompting achieves 26.84%. Across ODA-Train, the text-dominant view performs best on 51.14% of problems, the image-dominant view on 37.98%, and the combined image+text view on 10.88%. Thus, although text-dominant prompting is strongest on average, nearly half of the problems are best supervised by a non-text-dominant view. These results suggest that the qualitative behaviors in Table 1 are not isolated anecdotes, but recurring sources of modality asymmetry.

Given these asymmetries, a natural question is whether standard RL can exploit multiple views simply by training on them together. A controlled single-problem experiment in Appendix A shows that naive mixed-modality GRPO underperforms GRPO trained on the corresponding single view, indicating that exposure to paired views alone does not reliably transfer successful reasoning across modalities. We next introduce `MIRROR`, which makes this transfer explicit by converting the strongest view of each problem into a teacher signal for training the student views.

4 Our Approach: `MIRROR`

Our goal is to improve reasoning under a student view, either the text-dominant view or the image-dominant view. Section 3 shows that no single view is uniformly best, and `MIRROR` converts this view-dependent heterogeneity into a self-supervision signal. For each problem, it selects the view on which the current model performs the best as a problem-specific teacher. The restricted-view student is trained with standard outcome-reward RL plus an auxiliary reverse-KL regularizer. Crucially, the student always generates rollouts from its own restricted view; the selected teacher view is used only to score those student-generated trajectories.

Best-modality teacher selection. Each problem has three candidate teacher views: text-dominant $\mathbf{x}^{\text{text}}$, image-dominant $\mathbf{x}^{\text{img}}$, and combined image+text $\mathbf{x}^{\text{comb}} = (\mathbf{x}^{\text{text}}, \mathbf{x}^{\text{img}})$. Let $\mathcal{M} = \{\text{text}, \text{img}, \text{comb}\}$ denote the candidate teacher views. Teacher selection is performed online during training using the current policy model. For each problem $\mathbf{x}$ and each candidate view $j \in \mathcal{M}$, we sample $K = 16$ rollouts and estimate the view-specific success rate. We then select the best-performing view, breaking ties uniformly at random. The teacher distribution is then defined as a slow-moving copy of the model conditioned on the selected teacher view, $\pi_{\text{teach}}(\cdot \mid \mathbf{x}) = \pi_{\bar{\theta}}(\cdot \mid \mathbf{x}^{(j^\star(\mathbf{x}))})$, where $\bar{\theta}$ is an exponential moving average (EMA) of the policy parameters, described below. Thus, the teacher direction is selected independently for each problem and can adapt throughout training, rather than being fixed globally or precomputed from the base model. Although teacher selection considers three candidate views, the text-dominant and image-dominant rollouts are already generated as standard online GRPO rollouts; thus, the only additional rollout generation required by `MIRROR` is for the combined image+text view, which introduces negligible computational overhead. We provide a FLOPs comparison with mixed-modality GRPO in Appendix F.

Training under a restricted view. For a restricted view $m \in \{\text{text}, \text{img}\}$, the student receives the corresponding input $\mathbf{x}^{(m)}$ and samples rollouts $\mathbf{y} \sim \pi_\theta(\cdot \mid \mathbf{x}^{(m)})$. The selected teacher may condition on a different view $\mathbf{x}^{(j^\star(\mathbf{x}))}$, including the combined image+text view. The challenge is therefore to use the stronger teacher view without training the student on off-policy teacher-generated trajectories or forcing it to rely on information unavailable from its restricted input. We address this by optimizing a weighted sum of outcome-reward RL and an on-policy reverse-KL regularizer:

$$\mathcal{L}(\theta; m) = \mathcal{L}_{\text{RL}}(\theta; m) + \lambda_{\text{KL}} \mathcal{L}_{\text{rKL}}(\theta; m) \tag{1}$$

where $\lambda_{\text{KL}} \geq 0$ controls the strength of teacher guidance. The RL term is a GRPO-style on-policy objective that optimizes final-answer correctness under the restricted student view. The reverse-KL

term, defined below, transfers token-level preferences from the selected teacher view while keeping rollouts on-policy with respect to the restricted-view student.

Reverse KL for cross-view distillation. The auxiliary loss distills the selected teacher's reasoning preferences while keeping rollouts on-policy: for a student view $m \in \{\text{text}, \text{img}\}$, we sample $\mathbf{y} \sim \pi_\theta(\cdot \mid \mathbf{x}^{(m)})$ and use the selected teacher only to rescore the same generated tokens. This gives the reverse-KL value

$$\mathcal{L}_{\text{rKL}}(\theta; m) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x}^{(m)})} \left[\sum_{t=1}^{|\mathbf{y}|-1} \log \frac{\pi_\theta(y_{t+1} \mid \mathbf{x}^{(m)}, y_{1:t})}{\pi_{\bar{\theta}}(y_{t+1} \mid \mathbf{x}^{(j^\star(\mathbf{x}))}, y_{1:t})} \right] \tag{2}$$

To avoid an expensive full-vocabulary KL computation, we instead optimize the sampled K3 surrogate $g(r) = e^{-r} + r - 1$, where $r = \log \pi_\theta(y_{t+1} \mid \cdot) - \log \pi_{\bar{\theta}}(y_{t+1} \mid \cdot)$ and $\mathbb{E}_{\pi_\theta}[g(r)] = D_{\text{KL}}(\pi_\theta \| \pi_{\bar{\theta}})$. Operationally, this asks how compatible the student's tokens are with the stronger teacher view only at states induced by the policy being trained, avoiding unnecessary distribution shift.

EMA teacher. Because the teacher in Equation 2 is derived from the same network being trained, the most direct choice is to score the teacher term with the current parameters θ. However, this makes the distillation target move every step: the student chases a target that its own updates displace. Empirically, this self-referential coupling induces a slow inflation of policy entropy and reference-policy KL that eventually destabilizes long-horizon training (Figure 4). We therefore parameterize the teacher with an exponential moving average (Polyak average) of the policy parameters,

$$\bar{\theta} \leftarrow \alpha\, \bar{\theta} + (1 - \alpha)\, \theta \tag{3}$$

applied after every policy update, with $\bar{\theta}$ initialized to θ at the beginning of training. This mirrors target networks in value-based RL (Mnih et al., 2015; Lillicrap et al., 2019) and weight-averaged teachers in self-supervised learning (Tarvainen & Valpola, 2017; He et al., 2020; Grill et al., 2020): the teacher tracks the improving policy on a slower timescale, providing a slowly improving but stable target. The EMA teacher is held fixed during each update, while its token probabilities determine the direction and magnitude of the student gradients through the K3 surrogate.

The two terms in Equation 1 play complementary roles: the GRPO term optimizes final-answer correctness under the restricted view, while the reverse-KL term provides dense token-level guidance from the strongest view on the student's own trajectories. Unlike methods that sample from or imitate an external teacher distribution, `MIRROR` keeps rollouts on-policy with respect to the restricted-view student and uses the selected teacher view only as a supervision signal. Please see Algorithm 1 for the full pseudocode of `MIRROR` update.

5 Experimental Evaluation

The goal of our experiments is to evaluate whether `MIRROR` can turn asymmetries across views into an effective self-supervision signal for improving multimodal reasoning. We answer four questions in this section: **(1)** Does MIRROR improve restricted-view reasoning over single-view and mixed-modality GRPO baselines? **(2)** Does selecting the teacher view separately for each problem outperform using a fixed global teacher view? **(3)** Does `MIRROR` expand the set of solvable problem instances after training? **(4)** How sensitive is `MIRROR` to the cross-view KL coefficient λ_{KL} and to the parameterization of the teacher (current policy vs. slow-moving EMA)?

Experimental setup. We post-train `Qwen3-VL-4B-Instruct` and implement `MIRROR` in verl (Sheng et al., 2025). Models are trained with maximum response length 16,384, sampling temperature 0.8, GRPO clipping range $\epsilon_{\text{low}} = 0.2, \epsilon_{\text{high}} = 0.26$, reference-policy KL coefficient 0.001, and entropy coefficient 0.001. `MIRROR` uses a reverse-KL coefficient of 0.01 and EMA weights $\alpha = 0.99$. Our in-distribution evaluation uses ODA-Val, where paired views preserve problem identity and allow us to measure pass@k under each view and track how training changes problem-level solvability across views. To test whether improvements extend beyond this asymmetry-enriched distribution, we also evaluate on a disjoint random subset of unfiltered ODA-Math-460K that excludes all samples used in ODA-Train and ODA-Val. We additionally evaluate on public multi-modal geometry benchmarks GeoInt (Wei et al., 2025a) and MathVerse (Zhang et al., 2024b). For

Table 2: **MIRROR improves single-view reasoning over standard GRPO.** We report pass@1 and pass@16 on held-out image/text ODA-Val problems and GeoInt, pass@1 on a disjoint random unfiltered ODA-460K subset, and mean judge score on MathVerse. Parenthetical values after model or method names denote the reported number of post-training samples. Green tags indicate improvement and pink rounded tags indicate degradation relative to the Qwen3-VL-4B Instruct base model. **Bold** numbers indicate the best result.

Approach	ODA-Val Image		ODA-Val Text		ODA-Unfiltered		GeoInt		MathVerse
	pass@1	pass@16	pass@1	pass@16	image pass@1	text pass@1	pass@1	pass@16	mean
Existing models									
Vision-R1-7B (210K)	7.41	19.48	9.67	29.35	11.40	11.10	23.30	52.80	46.75
PAPO-7B (39K)	11.38	25.77	16.88	39.57	17.80	17.90	29.90	60.30	47.82
Vero-8B (600K)	16.16	39.74	35.48	71.73	34.50	38.90	**67.30**	77.42	**69.49**
Base model									
Qwen3-VL-4B Instruct	12.30	42.57	32.91	80.22	35.40	35.89	58.62	70.79	41.22
Standard GRPO (2K)									
text-dominant GRPO	17.32 +5.02	43.29 +0.72	39.92 +7.01	81.06 +0.84	36.40 +1.00	37.40 +1.51	63.02 +4.40	78.95 +8.16	45.38 +4.16
image-dominant GRPO	20.38 +8.08	48.78 +6.21	41.06 +8.15	83.16 +2.94	39.10 +3.70	39.80 +3.91	62.30 +3.68	**79.69** +8.90	39.24 −1.98
mixed-modality GRPO	17.99 +5.69	45.68 +3.11	39.18 +6.27	81.66 +1.44	36.87 +1.47	39.15 +3.26	61.68 +3.06	75.21 +4.42	44.72 +3.50
MIRROR (2K)	**23.57** +11.27	**57.06** +14.49	**45.67** +12.76	**86.10** +5.88	**39.40** +4.00	**41.29** +5.40	66.15 +7.53	78.38 +7.59	46.78 +5.56

GeoInt, we consider only answer-based question. We evaluate final-answer correctness with pass@1 and pass@16 estimated by 32 rollouts. For MathVerse, we use `gpt-4o-2024-11-20` (Hurst et al., 2024) as the judge through the LMMs-Eval suite (Zhang et al., 2024a) and report the mean score.

Baselines and comparisons. We evaluate the base `Qwen3-VL-4B-Instruct` model and three post-trained multimodal reasoners: `Vision-R1-7B` (Huang et al., 2026), which applies cold-start training and RL to `Qwen2.5-VL-7B-Instruct`; `PAPO-7B` (Wang et al., 2025b), which adds a perception-aware objective contrasting original and masked-image rollouts to the same base model; and `Vero-8B` (Sarch et al., 2026), which applies a broad, task-routed visual-reasoning recipe to `Qwen3-VL-8B-Instruct`. Our controlled ODA-Train comparisons include text-dominant and image-dominant single-view GRPO, mixed-modality GRPO trained on both views, and fixed-teacher MIRROR variants. Mixed-modality GRPO tests whether cross-view transfer emerges implicitly when both views are present, while MIRROR uses the same mixture of student prompts but makes the transfer direction explicit. We provide additional results on other base models in Appendix B. We also show MIRROR's generalization to the chart domain in Appendix C.

Result 1: MIRROR improves over standard and mixed-modality GRPO. Table 2 compares MIRROR with the base model, existing reasoning-based VLMs, and standard RL baselines. On ODA-Val, MIRROR performs best across all image- and text-view metrics, outperforming both the strongest single-view GRPO baseline and mixed-modality GRPO. On the ODA-Unfiltered, the base model performs nearly identically across the two views, confirming that this is not an asymmetry enriched distribution. Yet MIRROR still outperforms all baselines compared. Thus, the benefit of cross-view supervision is not limited to the filtered distribution that exposes strong modality asymmetry. The gains also transfer beyond ODA-Val: MIRROR improves over the base model on GeoInt and MathVerse. Despite using only 2K training examples, it outperforms Vision-R1-7B and PAPO-7B across ODA-Val and GeoInt and Vero-8B across ODA-Val, although Vero-8B remains strongest on MathVerse. Appendix B further shows that the gains generalize to Qwen3-VL-8B-Instruct (Bai et al., 2025) and GLM-4.1V-9B-Thinking (GLM-V Team, 2025), with MIRROR outperforming standard GRPO across all reported metrics for both larger base models. Together, these results show that cross-view asymmetry provides a useful training signal across evaluation distributions, model scales, and model families.



Figure 2: **Mixed-modality GRPO does not automatically transfer reasoning across views.** With full-dataset training, mixed-view GRPO underperforms GRPO trained directly on the corresponding restricted view, while MIRROR obtains the strongest training reward by using paired views directionally.

Mixed-modality GRPO is the most direct test of whether paired views are sufficient by themselves. Although this baseline trains on both text-

Table 3: **Ablation on teacher choice.** We compare fixed teacher choices against MIRROR, which adaptively selects the strongest modality-specific teacher for each problem. By selecting the teacher per problem, MIRROR better exploits modality-dependent strengths and provides more consistent cross-modal guidance.

Teacher choice	ODA-Val Image		ODA-Val Text		GeoInt		MathVerse
	pass@1	pass@16	pass@1	pass@16	pass@1	pass@16	mean
text-dominant teacher	19.12	51.03	39.78	83.59	64.41	**79.59**	45.28
image-dominant teacher	18.95	49.10	43.43	83.77	63.80	79.50	43.36
combined image+text teacher	20.99	52.28	43.35	82.33	64.56	78.77	**47.26**
Adaptive teacher (MIRROR)	**23.57**	**57.06**	**45.67**	**86.10**	**66.15**	78.38	46.78

dominant and image-dominant views of the same underlying problems, it treats each view as an independent RL prompt and relies on transfer emerging implicitly through shared parameters. Figure 2 and 5 show that this is not reliable: mixed-view GRPO obtains lower training reward than single-view GRPO in both the controlled setting and full-dataset training. This behavior is also consistent with *ray interference* in heterogeneous on-policy RL mixtures (Qu et al., 2026). In our setting, views differ in grounding requirements, and access to explicit relations. Optimizing all views with sparse outcome rewards does not guarantee that updates from an easier view teach the harder restricted view. MIRROR addresses this by specifying the direction of transfer: it selects the strongest-performing view to regularize the restricted-view student toward its teacher distribution. As detailed in Appendix F, MIRROR incurs approximately 37.5% more FLOPs per update than standard mixed-modality GRPO. To control for this difference, we compare MIRROR at step 150 with GRPO at step 200, which have roughly matched cumulative compute. Even under this compute-matched comparison, MIRROR achieves higher image (0.2625 vs. 0.2368) and text (0.4581 vs. 0.4387) reward, suggesting that its gains cannot be attributed to additional computation alone.

Result 2: Adaptive teacher selection provides the most consistent gains. We next evaluate whether the teacher view should be fixed globally. Table 3 shows that the fixed-teacher variants have complementary strengths, but none is uniformly best. These differing preferences indicate that the most useful supervision source is problem-dependent rather than governed by one global transfer direction. By selecting the strongest view separately for each problem, MIRROR achieves the best result on most metrics. Appendix A further shows that teacher-selection proportions evolve during training as the policy's view-specific capabilities change. Together, these results motivate selecting teachers online and per problem rather than fixing them globally or from the base model.

Result 3: MIRROR expands solvability and improves cross-view consistency. Aggregate pass@k scores do not fully capture how post-training changes the set of problems a model can solve. We therefore measure solvability on ODA-Train before and after training. Net solvability gain is the before-to-after change in the percentage of sampled problems with at least one successful rollout. As shown in Figure 3, all GRPO variants improve net solvability, but MIRROR achieves the largest gain at +8.95 pp. Beyond expanding the set of solvable examples, MIRROR also improves cross-view consistency. On ODA-Train, only 42.5% of problems are solvable under both the text-dominant and image-dominant views by the base model. Standard GRPO improves this both-view solvability rate to 53.6%, while MIRROR further increases it to 60.7%. This shows that MIRROR not only improves accuracy under each restricted view, but also helps the model solve the same underlying problem consistently across modalities.

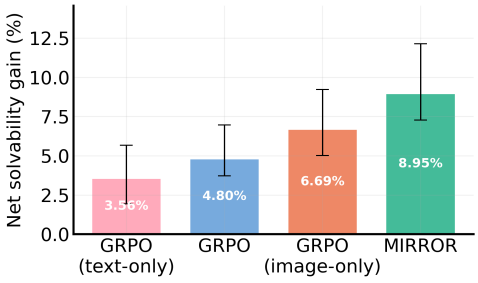


Figure 3: **Net solvability gain after training.** Each bar shows the change in the fraction of solved examples relative to the base model across 3 random seeds. Each random seed samples 309 problems which is the same size as the validation set. Error bars indicate the min and max values across seeds. MIRROR produces the largest net increase.

Ablation 1: effect of the reverse-KL coefficient. Table 4 studies the strength of the reverse-KL regularizer while keeping all other training components fixed for 100 steps. We found that $\lambda_{\text{KL}} = 0.01$ achieves the best rank overall and therefore use it for subsequent training. Larger coefficients degraded performance, indicating that excessive alignment can overpower the RL objective.

Table 4: **Ablation on the reverse-KL coefficient.** We sweep the strength of the cross-modal reverse-KL regularizer used in MIRROR while keeping all other components fixed and let it ran for 100 steps. This ablation tests how sensitive MIRROR is to the alignment strength and identifies the regime that best balances on-policy RL with cross-modal distillation.

Reverse KL coefficient	ODA-Val Image		ODA-Val Text		GeoInt		MathVerse
	pass@1	pass@16	pass@1	pass@16	pass@1	pass@16	mean
$\lambda_{\text{KL}} = 0.001$	18.26	49.98	42.03	85.21	62.30	77.24	44.50
$\lambda_{\text{KL}} = 0.005$	17.74	49.45	40.83	**86.70**	62.13	76.71	45.41
$\lambda_{\text{KL}} = 0.01$	**21.74**	52.36	**44.44**	85.25	**64.58**	**79.70**	**46.98**
$\lambda_{\text{KL}} = 0.05$	18.75	**54.73**	39.67	85.07	61.46	77.17	46.62

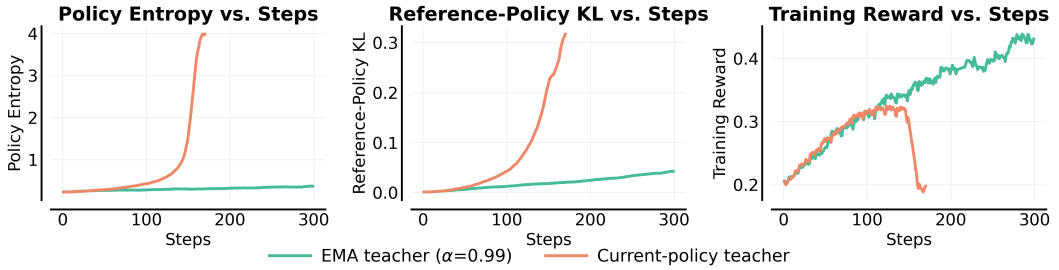


Figure 4: **The EMA teacher prevents long-horizon collapse.** Training MIRROR with the teacher scored by the current policy (no EMA) vs. the EMA teacher (α=0.99), otherwise identical. Left: policy entropy. Middle: reference-policy KL. Right: training reward averaged across image and text modalities. The current-policy teacher destabilizes after ∼90 steps; the EMA teacher stays stable and reaches higher reward.

Ablation 2: a slow-moving EMA teacher is necessary for long-horizon stability. Figure 4 compares MIRROR using an EMA teacher ($\alpha = 0.99$) with a variant using the current policy θ as the teacher, under otherwise identical settings ($\lambda_{\text{KL}} = 0.01$). The runs behave similarly for the first 50 steps, after which the current-policy teacher exhibits self-reinforcing drift: policy entropy and reference-policy KL increase while training reward collapses. In contrast, the EMA teacher remains stable. We interpret this difference as a moving-target effect analogous to the instability addressed by target networks in value-based RL (Mnih et al., 2015): decoupling the target's timescale from the student's stabilizes optimization without sacrificing asymptotic performance.

6 Discussion and Perspectives on Future Work

In this work, we identify reasoning asymmetry across modalities as a source of supervision signal for multimodal reasoning. We introduce **Modality-Informed Reciprocal Reasoning Optimization** (MIRROR), which selects the strongest-performing view of each problem as an internal teacher and regularizing the student-view policy toward it during RL. Our results show that paired multimodal data is most effective when the training objective explicitly exploits the correspondence between views. Simply mixing text-dominant and image-dominant views into the RL distribution does not reliably transfer reasoning, because each view is still optimized as an independent prompt. In contrast, MIRROR makes the transfer direction explicit: the empirically stronger view provides guidance for the view being trained. This yields stronger and more consistent reasoning, with the largest gains on examples exhibiting pronounced cross-view performance differences.

More broadly, our findings suggest a different way to use multimodal data. Rather than treating different input formats as independent examples, we can treat them as alternative views of the same latent problem and use their disagreements to expose missing reasoning capabilities. This perspective opens several directions for future work. Extending beyond geometry and charts to scientific figures, tables, robotics observations, or broader visual reasoning will require reliable ways to align views by problem identity and answer target while accounting for differences in their information content. Teacher selection could also be improved with uncertainty-aware weighting or alignment at intermediate reasoning steps, rather than relying only on rollout rewards. Overall, our findings point to a broader opportunity: cross-modal inconsistency can be transformed from a robustness concern into a useful training signal for learning multimodal reasoners that are stronger and less sensitive to how information is represented.

AI USE STATEMENT

We used generative AI tools for research execution, synthetic dataset construction, automated evaluation, related-work retrieval, and writing assistance. Claude Code was used to assist with implementing the authors' proposed research idea. The authors manually examined the resulting code and performed both smoke tests and full tests to verify its correctness. Gemini-3-Pro-Preview was used during the construction of ODA-Data to generate TikZ diagrams, assess their consistency with the source problems, screen for answer leakage, and rewrite prompts for the image-dominant view. GPT-4o-2024-11-20 was used for MathVerse scoring, and Gemini was used for the error analysis described in the appendix. These uses and their associated procedures are documented in the paper. Generative AI tools were also used to retrieve relevant related work. We manually consulted the original sources and verified the accuracy and relevance of all citations included in the paper. ChatGPT was used only to polish author-written text for clarity and readability. We did not use generative AI to formulate or prove mathematical claims. We reviewed all AI-assisted work and take full responsibility for the final content of this work, including its text, claims, code, data, and experimental results.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing detailed descriptions of the method, data construction, training procedure, and evaluation protocol. Section 4 defines the MIRROR objective, teacher-selection procedure, reverse-KL regularizer, and EMA update, while Appendix G provides complete pseudocode. Appendix H reports the model configurations, optimization settings, rollout parameters, and other training hyperparameters, and Appendix F details the computational-cost calculation. Section 5 specifies the evaluated models, baselines, benchmarks, metrics, rollout counts, and scoring procedures. Appendix I documents the source data, filtering and splitting procedures, generative models, and prompts used to construct ODA-Data. Additional ablations, generalization experiments, and error analyses are reported in the remaining appendices.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation, 2025. URL https://arxiv.org/abs/2510.23497.

Mengzhang Cai, Xin Gao, Yu Li, Honglin Lin, Zheng Liu, Zhuoshi Pan, Qizhi Pei, Xiaoran Shang, Mengyuan Sun, Zinan Tang, et al. Opendataarena: A fair and open arena for benchmarking post-training dataset value. *arXiv preprint arXiv:2512.14051*, 2025.

Jianlong Chen, Daocheng Fu, Shengze Xu, Jiawei Chen, Yuan Feng, Yue Yang, Junchi Yan, Hongyuan Zha, and Renqiu Xia. Milestones over outcome: Unlocking geometric reasoning with sub-goal verifiable reward, 2026a. URL https://arxiv.org/abs/2601.05073.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Yiqun Chen, Wei Yang, Erhan Zhang, Shijie Wang, Qi Liu, Zechun Niu, Bin Zhang, Haitao Li, Rui Li, Lingyong Yan, et al. Unitymas-o: A general rl optimization framework for llm-based multi-agent systems. *arXiv preprint arXiv:2605.26646*, 2026b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. URL https://arxiv.org/abs/2312.14238.

Xiaoteng Cui and Yi Liu. Plane geometry diagram formalization via vision-language models. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

Xingjian Diao, Zheyuan Liu, Chunhui Zhang, Weiyi Wu, Keyi Kong, Lin Shi, Kaize Ding, Soroush Vosoughi, and Jiang Gui. Addressing overthinking in large vision-language models via gated perception-reasoning optimization, 2026. URL https://arxiv.org/abs/2601.04442.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models, 2025. URL https://arxiv.org/abs/2411.14432.

GLM-V Team. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Google DeepMind. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, December 2025. Model card for Gemini 3 Pro.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe, 2026. URL https://arxiv.org/abs/2605.03677.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2026. URL https://arxiv.org/abs/2503.06749.

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

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

Mohamed Insaf Ismithdeen, Muhammad Uzair Khattak, and Salman Khan. Promptception: How sensitive are large multimodal models to prompts?, 2025. URL https://arxiv.org/abs/2509.03986.

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*, 2026.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025. URL https://arxiv.org/abs/2501.02189.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019. URL https://arxiv.org/abs/1509.02971.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Martin Q Ma, Yuxiao Qu, Aditya Agrawal, Willis Guo, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Act2see: Emergent active visual perception for video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5455–5464, 2026.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging visual and mathematical reasoning for large language models, 2024. URL https://arxiv.org/abs/2409.00147.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285, 2024.

Yuxiao Qu, Amrith Setlur, Virginia Smith, Ruslan Salakhutdinov, and Aviral Kumar. Pope: Learning to reason on hard problems via privileged on-policy exploration, 2026. URL https://arxiv.org/abs/2601.18779.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open rl recipe for general visual reasoning, 2026. URL https://arxiv.org/abs/2604.04917.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws, 2025. URL https://arxiv.org/abs/2401.00448.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL https://arxiv.org/abs/2601.19897.

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.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models, 2024. URL https://arxiv.org/abs/2406.17294.

Chendong Song, Zihan Wang, Frederick Pu, Haiming Wang, Xiaohan Lin, Junqi Liu, Jia Li, and Zhengying Liu. Leangeo: Formalizing competitional geometry problems in lean, 2025. URL https://arxiv.org/abs/2508.14644.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026. URL https://arxiv.org/abs/2601.02780.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. Mad-opd: Breaking the ceiling in on-policy distillation via multi-agent debate, 2026. URL https://arxiv.org/abs/2605.01347.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning, 2025a. URL https://arxiv.org/abs/2505.10557.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025b.

Jingxuan Wei, Caijun Jia, Qi Chen, Honghao He, Linzhuang Sun, Conghui He, Lijun Wu, Bihui Yu, and Cheng Tan. Geoint-r1: Formalizing multimodal geometric reasoning with dynamic auxiliary constructions. *arXiv preprint arXiv:2508.03173*, 2025a.

Jingxuan Wei, Caijun Jia, Qi Chen, Honghao He, Linzhuang Sun, Conghui He, Lijun Wu, Bihui Yu, and Cheng Tan. Geoint-r1: Formalizing multimodal geometric reasoning with dynamic auxiliary constructions, 2025b. URL https://arxiv.org/abs/2508.03173.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. ChartX & ChartVLM: A versatile benchmark and

foundation model for complicated chart reasoning, 2024. URL https://arxiv.org/abs/2402.12185.

Yige Xu, Yongjie Wang, Zizhuo Wu, Kaisong Song, Jun Lin, and Zhiqi Shen. Do vision-language models truly perform vision reasoning? a rigorous study of the modality gap, 2026. URL https://arxiv.org/abs/2604.16256.

Wei Yang and Jesse Thomason. Learning to deliberate: Meta-policy collaboration for agentic llms with multi-agent reinforcement learning. *arXiv preprint arXiv:2509.03817*, 2025.

Wei Yang, Jiacheng Pang, Shixuan Li, Paul Bogdan, Stephen Tu, and Jesse Thomason. Maestro: Learning to collaborate via conditional listwise policy optimization for multi-agent llms. *arXiv preprint arXiv:2511.06134*, 2025a.

Wei Yang, Shixuan Li, Heng Ping, Peiyu Zhang, Paul Bogdan, and Jesse Thomason. Auditing multi-agent llm reasoning trees outperforms majority vote and llm-as-judge. *arXiv preprint arXiv:2602.09341*, 2026a.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026b. URL https://arxiv.org/abs/2602.12125.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025b. URL https://arxiv.org/abs/2503.10615.

Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Yufei Wu, Shichao Su, Yifan Xu, Wenxi Zeng, Zhaoyu Yang, Guoyou Li, Shilan Zhang, Zichan Li, Yaxiong Chen, Shengwu Xiong, Peng Xu, Jiajun Zhang, Bowen Zhou, David Clifton, and Luc Van Gool. Lens: Multi-level evaluation of multimodal reasoning with large language models, 2025. URL https://arxiv.org/abs/2505.15616.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024b. URL https://arxiv.org/abs/2403.14624.

Shan Zhang, Aotian Chen, Yanpeng Sun, Jindong Gu, Yi-Yu Zheng, Piotr Koniusz, Kai Zou, Anton van den Hengel, and Yuan Xue. Open eyes, then reason: Fine-grained visual mathematical understanding in mllms, 2025. URL https://arxiv.org/abs/2501.06430.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024c. URL https://arxiv.org/abs/2302.00923.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

Appendices

A Controlled Single-Problem Analysis and Online Teacher Dynamics

Given the cross-view asymmetries identified in Section 3, two related questions arise. First, if standard RL observes text-dominant and image-dominant views of the same problem, will successful reasoning automatically transfer from the stronger view to the weaker one? Second, can MIRROR provide useful guidance when the strongest view has only a small absolute success rate?

Teacher selection in MIRROR is based on relative empirical success, so the selected view is not necessarily an accurate oracle. For example, the strongest view may solve only 2 of 16 rollouts while the other two views solve none. MIRROR does not require the selected teacher to be highly accurate; it requires only a relative capability advantage over the student view. Unlike supervised fine-tuning on teacher-generated solutions, MIRROR does not train the student to imitate teacher trajectories. The student samples its own trajectories, and the selected teacher distribution provides token-level reverse-KL guidance on those student-generated states. At the same time, the outcome-reward RL objective continues to anchor learning to task correctness regardless of teacher strength. The student is therefore neither trained solely by the teacher nor bounded by the teacher's initial success rate.

We study these questions through a controlled single-problem experiment comparing mixed-modality GRPO, text-only GRPO, image-only GRPO, and MIRROR. In the mixed-modality condition, the same underlying problem appears under both text-dominant and image-dominant views, but each view is treated as an independent RL prompt and transfer must emerge implicitly through shared policy updates. The single-view conditions remove this cross-view heterogeneity and optimize only the corresponding student view. Figure 5 shows the training-reward trajectories, while Table 5 reports endpoint success counts from the same experiment before training and after 100 optimization steps.

Initially, the text-dominant view has only a small advantage over the image-dominant view ($1/16$ versus $0/16$). Nevertheless, Figure 5 and Table 5 show that MIRROR reaches $13/16$ successful rollouts under both views after 100 steps, compared with $5/16$ under mixed-modality GRPO and $8/16$ under the corresponding single-view GRPO runs. Thus, the student is not bounded by the teacher view's initial performance, both views can improve as teacher selection and policy optimization are updated online, and directed cross-view guidance accelerates this co-improvement relative to mixed-modality and single-view RL. In contrast, naive mixed-modality GRPO obtains the lowest reward on both views, showing that exposure to paired views alone does not reliably transfer successful reasoning without an explicit direction of guidance.

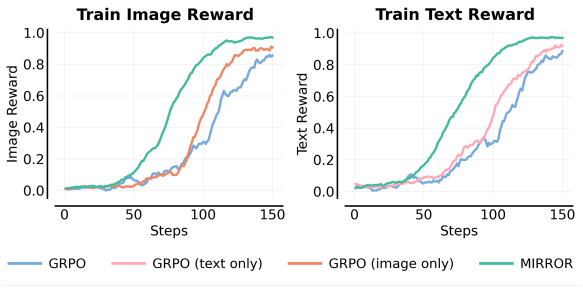


Figure 5: **Controlled single-problem analysis.** Across both views, MIRROR achieves the highest training reward, followed by GRPO trained on the corresponding single modality, while naive GRPO over mixed-modality views performs worst. Simply exposing RL to multiple views of the same underlying problem is therefore insufficient for reliable cross-modal transfer; without an explicit transfer direction, the sparse reward signal can be diluted across heterogeneous problem–view pairs.

The relative advantages used for teacher selection are typically larger than the one-rollout gap in the controlled example. Before training, the median difference between the strongest and weakest views across problems is 0.2188, corresponding to $7/32$ rollouts. The mean success rates of the strongest, second-strongest, and weakest views are 0.3602, 0.2348, and 0.0751, respectively. Thus, although

Table 5: **Endpoint success counts for the controlled experiment in Figure 5.** Each entry reports successful rollouts out of 16 on the same problem before training and after 100 optimization steps. Dashes denote views excluded from the corresponding single-view GRPO run.

	Image-dominant		Text-dominant	
Setting	**Step 0**	**Step 100**	**Step 0**	**Step 100**
Mixed-modality GRPO	0/16	5/16	1/16	5/16
GRPO (text only)	–	–	1/16	8/16
GRPO (image only)	0/16	8/16	–	–
MIRROR	0/16	**13/16**	1/16	**13/16**

the selected teacher is not assumed to be an oracle, the initial views commonly exhibit a nontrivial relative capability gap that can provide directional guidance. These initial gaps form the starting point for online teacher selection, but the assignment need not remain fixed as the model learns.

Because MIRROR performs teacher selection online, the selected view can change as text-dominant, image-dominant, and combined-view capabilities evolve. Table 6 compares the teacher-selection proportions averaged over the first and last 50 training steps. *No teacher* denotes cases in which no alternative view performs better than the current student view and therefore no view offers a relative capability advantage for cross-view guidance.

Table 6: **Online teacher-selection dynamics.** We report the mean teacher-selection percentage and its variation over the first and last 50 training steps.

Teacher view	**Early training**	**Late training**
Combined	29.67 ± 3.31	28.17 ± 2.88
Text	18.05 ± 2.19	14.23 ± 1.86
Image	5.30 ± 1.26	5.25 ± 1.30
No teacher	46.98 ± 2.64	52.35 ± 2.64

The distribution changes moderately rather than abruptly. The combined view remains the most frequently selected teacher, decreasing slightly from 29.67% to 28.17%. Text-teacher selection decreases from 18.05% to 14.23%, while image-teacher selection remains nearly unchanged. Meanwhile, the no-teacher proportion increases from 46.98% to 52.35%. This increase indicates that, as training progresses, a larger fraction of student views no longer has an alternative view that performs better, so fewer examples offer a relative advantage for cross-view guidance. The changing proportions also show that modality dominance is not fixed by the base model, supporting online, problem-specific teacher selection rather than assigning one teacher modality before training.

B Generalization to Larger Base Models

The main experiments use Qwen3-VL-4B-Instruct. To test whether the benefits of cross-view supervision are specific to a single base model or scale, we repeat the comparison with Qwen3-VL-8B-Instruct (Bai et al., 2025) and GLM-4.1V-9B-Thinking (GLM-V Team, 2025). For each model, we compare the base model with GRPO and MIRROR. Table 7 reports pass@1 and pass@16 on ODA-Val, image- and text-view pass@1 on ODA-Unfiltered, pass@1 and pass@16 on GeoInt, and the official five-split mean judge score on MathVerse.

For Qwen3-VL-8B-Instruct, standard GRPO improves GeoInt and ODA-Val pass@1, but reduces ODA-Val image pass@16 from 45.79% to 41.73%, text pass@16 from 82.41% to 75.81%, and both ODA-Unfiltered pass@1 scores. In contrast, MIRROR improves all nine reported metrics over the base model. Its gains are largest on ODA-Val, where image pass@1 and pass@16 improve by 12.09 and 15.11 pp, and text pass@1 and pass@16 improve by 14.39 and 5.32 pp. It also improves ODA-Unfiltered image/text pass@1 by 4.7/5.8 pp, GeoInt pass@1/pass@16 by 9.66/6.01 pp, and the MathVerse mean judge score by 2.24 points.

Table 7: **MIRROR generalizes to larger base models.** Within each model family, MIRROR improves over the base model on every reported metric. ODA-Unfiltered contains 309 problems and is evaluated with 32 rollouts per problem and view. MathVerse values are averages over the five official splits.

Approach	ODA-Val Image		ODA-Val Text		ODA-Unfiltered		GeoInt		MathVerse
	pass@1	pass@16	pass@1	pass@16	image pass@1	text pass@1	pass@1	pass@16	mean
Qwen3-VL-8B-Instruct	14.40	45.79	34.48	82.41	37.0	37.7	58.81	74.07	55.30
GRPO	17.29	41.73	36.10	75.81	30.1	33.5	65.41	79.67	55.81
MIRROR	**26.49**	**60.90**	**48.87**	**87.73**	**41.7**	**43.5**	**68.47**	**80.08**	**57.54**
GLM-4.1V-9B-Thinking	14.18	34.61	25.94	57.99	25.6	27.4	56.70	73.30	69.65
GRPO	19.50	40.20	31.74	61.51	29.0	29.6	62.60	76.40	70.03
MIRROR	**24.54**	**49.98**	**33.75**	**63.19**	**29.5**	**30.0**	**64.10**	**78.50**	**70.36**

The GLM-4.1V-9B-Thinking results show the same pattern. Relative to its base model, MIRROR improves ODA-Val image pass@1/pass@16 by 10.36/15.37 pp, ODA-Val text pass@1/pass@16 by 7.81/5.20 pp, ODA-Unfiltered image/text pass@1 by 3.9/2.6 pp, GeoInt pass@1/pass@16 by 7.40/5.20 pp, and the official five-split MathVerse mean by 0.71 points. It also exceeds GRPO on all nine reported metrics. Together, these results show that the cross-view training signal is not limited to the 4B base model used in the main experiments or to a single larger-model family.

C Generalization to a New Domain: Chart Reasoning

We further test whether the cross-view training framework generalizes beyond geometry by applying it in the chart domain. ChartX (Xia et al., 2024) is a multimodal chart-reasoning benchmark spanning 18 chart types, 7 tasks, and 22 disciplinary topics. Each chart is represented through four modalities: a chart image, its underlying CSV data, the corresponding Python code, and a text description. In this experiment, we use only the question-answering task. We construct the chart-domain training data in the same manner as ODA-Data: the image and CSV representations form paired views of the same question and underlying chart information, and we select examples that exhibit view-dependent solvability asymmetry under the base model. Starting from Qwen3-VL-4B, we post-train on these asymmetric paired question-answering examples with either standard GRPO or MIRROR, and evaluate pass@1 accuracy for both input representations.

Table 8: **Cross-view training generalizes to chart reasoning.** Results are pass@1 accuracy on the ChartX question-answering task after chart-domain post-training on image and CSV views. The final column reports the improvement of MIRROR over standard GRPO in pp.

Input view	Base	GRPO	MIRROR	Δ vs. GRPO
Image	58.49	61.71	**63.74**	+2.03
CSV	82.41	92.35	**93.15**	+0.80

Both chart-domain post-training methods improve over the base model, and MIRROR is strongest under both chart representations. Relative to the base model, MIRROR improves pass@1 by 5.25 pp for chart images and 10.74 pp for CSV inputs. It also outperforms standard GRPO by 2.03 pp on images and 0.80 pp on CSV inputs. The larger advantage on the image view suggests that cross-view supervision is particularly helpful when the model must recover values and relations from a visual representation rather than reading them directly from structured text. Together with the geometry results, this experiment shows that the training framework can be instantiated in a distinct domain with different input representations.

D Consistency of Cross-View Asymmetry Across Model Scales

The effectiveness of MIRROR depends on different views exposing different capabilities of the current model. We therefore examine whether the direction of this asymmetry is stable when model scale changes. On the MIRROR training set, we sample 32 rollouts per view from both Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct. For problem i and model m, let $\hat{p}^{(m)}_{i,\text{image}}$ and $\hat{p}^{(m)}_{i,\text{text}}$ denote the

empirical image-view and text-view success rates. We define the signed modality gap as

$$g_i^{(m)} = \hat{p}_{i,\text{image}}^{(m)} - \hat{p}_{i,\text{text}}^{(m)}. \tag{4}$$

A positive gap indicates that the image view is stronger, a negative gap indicates that the text view is stronger, and a zero gap is a tie. We call a problem strictly asymmetric when $g_i^{(m)} \neq 0$.

Table 9: **Cross-view asymmetry persists across model scales.** Preference categories are computed from 32 rollouts per view on the MIRROR training set. Percentages are rounded independently.

Image–text preference	Qwen3-VL-4B	Qwen3-VL-8B
Image view stronger	15.7%	18.1%
Text view stronger	70.4%	70.5%
Tie	13.8%	11.3%
Strict cross-view asymmetry	86.2%	88.7%

The aggregate preference distribution is similar at both scales. The text view is stronger for roughly 70% of examples for each model, while strict cross-view asymmetry occurs for 86.2% of examples at 4B and 88.7% at 8B. The similarity is also visible at the problem level. Among examples for which both models have a strict preference, the models agree on the stronger view in 89.6% of cases. Moreover, their signed modality gaps are strongly correlated, with Pearson correlation $r = 0.779$. Thus, scaling the model from 4B to 8B does not eliminate view-dependent solvability or substantially change its overall direction.

At the same time, preference is not completely invariant to model scale. In the remaining 10.4% of examples where both models have a strict preference, one model favors the image view while the other favors the text view. A teacher assignment computed once using a different model or checkpoint would therefore be incorrect for a non-negligible subset of examples. This residual disagreement supports the online, model-specific teacher selection in MIRROR, which continually chooses the stronger view according to the capabilities of the model being trained rather than relying on a fixed cross-model preference label.

E Error Analysis of Cross-View Transfer

To better understand what MIRROR transfers across modalities, we conduct an error analysis on a random sample comprising 15% of the data used by MIRROR. For both Qwen3-VL-4B-Instruct and MIRROR, we generate 32 rollouts per problem under each of the text-dominant and image-dominant views. A Gemini judge classifies the first incorrect response from each model into one of three primary error categories: perception, reasoning, or arithmetic/final-answer. When a model solves all 32 rollouts for a problem, we instead assign the outcome *No error*.

Table 10: Primary error outcomes before and after MIRROR training.

	Text-dominant		Image-dominant	
Primary outcome	**Qwen3-VL-4B-Instruct**	**MIRROR**	**Qwen3-VL-4B-Instruct**	**MIRROR**
Perception error	0.0%	0.0%	53.0%	51.7%
Reasoning error	52.4%	44.8%	16.0%	10.7%
Arithmetic/final-answer error	47.5%	45.5%	31.0%	31.0%
No error	0.0%	9.7%	0.0%	6.7%

The largest reduction is in reasoning errors: 7.6 pp for text-dominant inputs and 5.3 pp for image-dominant inputs. Moreover, among problems on which the base model produces an error, MIRROR is error-free across all 32 sampled rollouts in 9.7% of text-dominant cases and 6.7% of image-dominant cases. In contrast, arithmetic/final-answer errors decrease only slightly for text-dominant inputs and remain unchanged for image-dominant inputs. These shifts suggest that MIRROR's improvements include mainly gains in reasoning rather than arising only from corrections to arithmetic or final-answer production.

F FLOPs Estimation

We estimate the computational overhead of MIRROR relative to standard mixed-modality GRPO. Following the common approximation used in language-model scaling analyses, we count one forward-only pass over D tokens as approximately $2ND$ FLOPs (Sardana et al., 2025), where N is the number of model parameters, and one training pass with forward and backward propagation as approximately $6ND$ FLOPs (Hoffmann et al., 2022). This gives a coarse but useful estimate of the relative cost between rollout generation, teacher-logit evaluation, and gradient updates.

Setup. Standard mixed-modality GRPO trains on two restricted student views: text-dominant and image-dominant. For each problem, it samples K rollouts from each view, giving $2K$ online student rollouts. MIRROR uses the same two restricted views for student RL training, but performs teacher selection over three candidate views:

$$\mathcal{M} = \{\text{text}, \text{img}, \text{comb}\}$$

The text-dominant and image-dominant candidate rollouts are exactly the online rollouts already generated for GRPO, so they do not need to be recomputed. Therefore, the only additional rollout generation required by MIRROR is the K rollouts from the combined image+text view.

Mixed-modality GRPO cost. Let D denote the average rollout length. Standard mixed-modality GRPO generates $2K$ rollouts per problem, so its forward-only rollout-generation cost is approximately

$$C_{\text{mixed-GRPO,gen}} \approx 2K \cdot 2ND = 4KND. \tag{5}$$

It then performs one student-side policy update on these $2K$ rollouts. Using the standard training approximation, the forward-backward update cost is

$$C_{\text{mixed-GRPO,train}} \approx 2K \cdot 6ND = 12KND. \tag{6}$$

Thus, ignoring implementation-specific constants, the total cost is

$$C_{\text{mixed-GRPO}} \approx 4KND + 12KND = 16KND. \tag{7}$$

MIRROR cost. MIRROR reuses the same $2K$ text-dominant and image-dominant student rollouts as mixed-modality GRPO. It additionally samples K combined-view rollouts for online teacher selection. Therefore, its rollout-generation cost is

$$C_{\text{MIRROR,gen}} \approx 3K \cdot 2ND = 6KND. \tag{8}$$

After selecting the teacher view, MIRROR computes the reverse-KL regularizer on the $2K$ student-view rollouts. This requires a teacher-forced forward pass under the selected teacher view to compute teacher log-probabilities on the same generated tokens using the EMA teacher weights. Since this is forward-only and stop-gradient, its cost is approximately

$$C_{\text{MIRROR,KL-eval}} \approx 2K \cdot 2ND = 4KND. \tag{9}$$

The student-side GRPO update is unchanged from mixed-modality GRPO:

$$C_{\text{MIRROR,train}} \approx 2K \cdot 6ND = 12KND. \tag{10}$$

Therefore, the total cost of MIRROR is approximately

$$C_{\text{MIRROR}} \approx 6KND + 4KND + 12KND = 22KND. \tag{11}$$

Relative overhead. Under this coarse FLOPs model, the relative cost of MIRROR compared with standard mixed-modality GRPO is

$$\frac{C_{\text{MIRROR}}}{C_{\text{mixed-GRPO}}} \approx \frac{22KND}{16KND} = 1.375. \tag{12}$$

Thus, MIRROR adds roughly 37.5% FLOPs over standard mixed-modality GRPO under a single-update accounting. This overhead comes only from forward-only computation: K additional combined-view rollouts for teacher selection and a teacher-logit evaluation pass for the reverse-KL regularizer. MIRROR uses the same trainable model, the same $2K$ student-view RL rollouts, and the same number of backward updates as mixed-modality GRPO. To examine whether standard GRPO would obtain similar optimization gains under comparable compute, we inspect the running-average training rewards at MIRROR step 150 and GRPO step 200 shown in figure 2, which correspond to approximately similar cumulative FLOPs. At these points in training, MIRROR has higher image reward (0.2625 vs. 0.2368) and text reward (0.4581 vs. 0.4387). This suggests that the advantage is not solely due to the additional per-update computation.

EMA teacher overhead. The EMA teacher does not change the FLOPs accounting above: the teacher-logit evaluation pass ($C_{\texttt{MIRROR},\text{KL-eval}} \approx 4KND$) is simply executed with the EMA weights $\bar{\theta}$ instead of the current weights θ, and the Polyak update in Equation 3 is a single $O(N)$ elementwise operation per training step (under one second per step in our runs, versus roughly 20 minutes per full training step). The main cost is memory: one additional forward-only copy of the policy parameters, kept in fp32 for numerically stable high-decay averaging and offloaded to CPU between teacher evaluations. Checkpoints additionally store the EMA parameters so that training resumes exactly.

G Pseudocode for MIRROR

Algorithm 1 outlines the training loop of MIRROR.

H Training Details

We train all models using GRPO with verl (Sheng et al., 2025). Unless otherwise specified, experiments use `Qwen3-VL-4B-Instruct` as the base model, vLLM as the rollout engine, and 8 nodes with 8 H200 GPUs per node. Each training run uses a maximum prompt length of 4096 tokens and a maximum response length of 16,384 tokens. We use a sampling temperature of 0.8 and generate $n = 16$ rollouts per prompt. The training batch size is 256, with PPO mini-batch size 128 and micro-batch size 1 per GPU. We use GRPO with clipping thresholds $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.26$. We train with learning rate 1×10^{-6}, entropy coefficient 0.001, and a reference-policy KL loss coefficient of 0.001 using the low-variance KL estimator. For MIRROR, we additionally enable the cross-modal reverse-KL regularizer with coefficient $\lambda_{\text{KL}} = 0.01$ by default, and compute the teacher target log-probabilities from an EMA copy of the policy with decay $\alpha = 0.99$, updated after every actor step. The EMA copy is kept in fp32 and is forward-only (no optimizer state); it is saved with each checkpoint so training resumes exactly.

Each RL training job uses 8 nodes with 8 H200 GPUs per node, runs at approximately 20 minutes per training step, and is trained for at least 200 steps, corresponding to roughly 67 wall-clock hours or about 4,267 H200 GPU-hours per model.

Table 11: **Main training hyperparameters.** Unless otherwise specified, all MIRROR and GRPO baselines use the same training configuration. The only method-specific differences are whether the cross-modal reverse-KL regularizer is enabled and its EMA teacher (decay 0.99).

Category	Setting
Base model	`Qwen/Qwen3-VL-4B-Instruct`
Training framework	verl (Sheng et al., 2025)
RL algorithm	GRPO
Rollout engine	vLLM
Maximum prompt length	4096 tokens
Maximum response length	16,384 tokens
Rollouts per prompt	16
Sampling temperature	0.8
Train batch size	256
PPO mini-batch size	128
PPO micro-batch size per GPU	1
Learning rate	1×10^{-6}
Clip ratio low	0.2
Clip ratio high	0.26
Entropy coefficient	0.001
Reference KL loss	Enabled
Reference KL coefficient	0.001
Reference KL type	Low-variance KL
Total epochs	200
Cross-modal reverse-KL coefficient λ_{KL}	0.01 (MIRROR only)
EMA teacher decay α	0.99 (MIRROR only)
EMA teacher precision	fp32, forward-only

Algorithm 1 `MIRROR` update loop

Require: policy π_θ, frozen reference policy π_{ref}, EMA teacher decay α, paired dataset $\mathcal{D}$, rollouts per prompt K, reverse-KL coefficient λ_{KL}, learning rate η

Dataset. Each problem has paired text-dominant x^{text} and image-dominant x^{image} views, sharing the same pair identifier. Each row also has a fixed combined image+text input x^{comb} precomputed at dataset construction time.
Sampler. Construct minibatches so that paired rows from the same problem appear in the same minibatch and on the same data-parallel rank.
EMA teacher. Initialize EMA parameters $\bar{\theta} \leftarrow \theta$.

for each epoch **do**
  for each minibatch B **do**
    Set $\pi_{\text{old}} \leftarrow \pi_\theta$.

    Standard rollouts from student views.
    for each row $x \in B$ **do**
      Sample N responses $\{y_i\}_{i=1}^{N} \sim \pi_{\text{old}}(\cdot \mid x)$.
      Compute rewards $r_i \leftarrow R(x, y_i)$ for $i = 1, \ldots, K$.
    end for

    combined view rollouts.
    for each row $x \in B$ **do**
      Sample N responses $\{y_j^{\text{comb}}\}_{j=1}^{K} \sim \pi_{\text{old}}(\cdot \mid x^{\text{comb}})$.
      Compute rewards $r_j^{\text{comb}} \leftarrow R(x^{\text{comb}}, y_j^{\text{comb}})$ for $j = 1, \ldots, K$.
    end for

    Teacher selection.
    for each row $x \in B$ **do**
      Let x^{pair} be the paired view of the same problem.
      Let $\mathcal{T}(x)$ contain three candidates: the current rollout $\pi_{\text{old}}(\cdot \mid x)$, the paired-view rollout $\pi_{\text{old}}(\cdot \mid x^{\text{pair}})$ and the combined-view rollout $\pi_{\text{old}}(\cdot \mid x_{\text{comb}})$.
      Score each candidate using its rollout rewards.
      Select the highest-scoring valid candidate view x^{sel}; define the teacher as the EMA model under that view: $\pi_T(\cdot \mid x) \leftarrow \pi_{\bar{\theta}}(\cdot \mid x^{\text{sel}})$.
      If the selected teacher is the current view, exclude x from the reverse-KL loss.
    end for

    GRPO loss.
    Compute GRPO advantages from the standard rollout rewards $\{r_i\}_{i=1}^{K}$.
    Compute $\mathcal{L}_{\text{GRPO}}$ using π_θ, π_{old}, π_{ref}, and the standard rollouts.

    Reverse-KL loss.
    Compute $\mathcal{L}_{\text{rKL}}$ using the sampled K3 surrogate described after Equation 2.

    Update.
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{GRPO}} + \lambda_{\text{KL}} \mathcal{L}_{\text{rKL}}$.
    Update θ by minimizing $\mathcal{L}_{\text{total}}$ with learning rate η.
    $\bar{\theta} \leftarrow \alpha\, \bar{\theta} + (1 - \alpha)\, \theta$ ▷ EMA teacher update
  end for
end for

I ODA-Data Construction

To study modality-dependent reasoning while holding the underlying problem fixed, we construct **ODA-Data**, a paired multimodal geometry dataset in which each problem is represented in two aligned views: a text-dominant view and an image-dominant view. Starting from 97K geometry-style problems from OpenDataArena/ODA-Math-460k (Cai et al., 2025), we first apply a difficulty

filter that removes easy questions with annotated pass@5=1, yielding 16K candidate questions. For each candidate, we prompt an LLM to generate a diagram in TikZ, a LaTeX-based language for programmatic scientific graphics. We choose TikZ over direct image generation because preliminary human inspection indicated that the model generated usable TikZ code more reliably and that the compiled diagrams had higher visual quality than directly generated images. We retain only diagrams whose TikZ code compiles successfully and use an LLM judge to verify that the generated diagram is semantically aligned with the original problem statement. We then apply a separate LLM judge to identify and filter out diagrams that leak the answer.

To construct the image-dominant view, we further prompt the judge to remove from the text any information already presented in the diagram, so that the remaining text and image provide complementary information. All LLM-based generation and judging steps use Gemini-3-Pro-Preview. We then perform modality-specific rollout evaluations using the base model, generating 32 rollouts for each of the text-dominant and image-dominant views, for a total of 64 rollouts per question. We define a view as failing when at most one of its 32 rollouts is correct. A question is retained only when exactly one of the two views meets this failure criterion; equivalently, one view has at most one successful rollout while the other has at least two. We exclude questions for which both views meet the failure criterion because they do not provide the cross-view advantage targeted by this construction. This filtering process yields approximately 2K curated paired multimodal examples.

We split these examples into **ODA-Train** and **ODA-Val** using an 85:15 split, with 15% of the dataset held out for validation. **ODA-Train** is used for cross-modal post-training, while **ODA-Val** serves as a held-out benchmark for evaluating view-dependent failure, cross-modal consistency, and cross-view transfer. Each example contains the original text problem, generated diagram, image-dominant prompt, text-dominant prompt, answer annotations, and modality-specific rollout/evaluation metadata. This structure makes ODA-Data useful not only for training `MIRROR`, but also for studying whether VLMs solve the same mathematical problem consistently across textual and visual representations.

For Tikz code generation, the prompt is:

Prompt for Tikz Code Generation

<question>
Your task is to generate **only the TikZ (LaTeX) code** that draws the geometric diagram described in the problem.
Important rules:

- Do NOT solve the problem.
- You may use colors (e.g., for clarity or emphasis), but:
  - Do NOT use solid fills that completely cover underlying lines or shapes.
  - If you use 'fill', it must include an 'opacity' value (e.g., 'fill opacity=0.2') or be very light.
  - Prefer 'draw', 'dashed', 'thick', or colorized outlines over filled regions.
- Avoid unnecessary decorations or stylistic choices not implied by the problem.
- Ensure the diagram is clear, uncluttered, and geometrically consistent.

Return only the TikZ code in a *latex* block.

For verifying the tikz compiled image against the question, the prompt is:

Prompt for Image Verification

You are a strict geometry diagram judge.
You are given:

1. A geometry problem described in text
2. A diagram image intended to represent that problem

Your task is to judge whether the diagram is an accurate and faithful representation of the problem text.
Check carefully that:

1. All geometric entities mentioned in the text (points, lines, angles, shapes, objects) appear in the diagram.
2. Spatial relationships match the text (relative positions, intersections, parallelism, perpendicularity, containment, motion if described).
3. Angle markings are correct:
   - Angles are marked at the correct vertex.
   - The marked angle corresponds to the intended angle, not its complementary or supplementary angle.
   - The rays defining the angle match the textual description.
4. Any stated angle measures or symbols (e.g., $90\circ$, $100\circ$, α, β) are applied to the correct geometric region.
5. The diagram does not introduce misleading constructions, extra constraints, or alternative interpretations that contradict the text.
6. Do NOT judge based on visual scale accuracy:
   - Differences in segment lengths, proportions, or drawing scale (e.g., a segment of length 15 not appearing longer than one of length 13) should NOT be treated as errors unless relative length ordering is explicitly required by the text.

If any required element is missing, incorrectly marked, ambiguously represented, or misleading, judge the diagram as inaccurate.
Respond with exactly the following format:
Judgement: Accurate or Inaccurate
Reason: A brief explanation (1–2 sentences) identifying the key correctness or primary issue.
Now evaluate the following.
[PROBLEM TEXT]
<QUESTION TEXT>
[DIAGRAM IMAGE]
<IMAGE>

For rewriting image questions to remove information present in the diagram from the question, the prompt is:

Prompt for Text Rewrite

You are given a math question that is accompanied by a diagram/image. Your task is to rewrite the question text so that information **already visible in the image** is removed from the text.
Rules:

1. **Remove** textual descriptions of geometric configurations, diagrams, shapes, angles, side lengths, point positions, or any other information that a reader could extract from the image.
2. **Only remove properties that are explicitly marked in the image.** For example, if the text says 'isosceles triangle' and the diagram shows a triangle that look isosceles but does NOT have explicit markings (e.g., tick marks on equal sides, or a label saying 'isosceles'), you must KEEP 'isosceles triangle' in the text. The same applies to right angles (keep unless a square corner mark is shown), parallel lines (keep unless arrow marks are shown), equal segments (keep unless tick marks are shown), etc.
3. **Always keep** the final question being asked (e.g., 'Find the length of...', 'What is the value of...', 'Determine...').
4. **Always keep** any answer format instructions (e.g., 'put your final answer in $\square$', 'Final Answer: $\square$').
5. **Keep** any mathematical constraints or conditions that are purely algebraic and NOT depicted in the image.
6. If the question is purely algebraic with no visual information to remove, return it UNCHANGED.

7. Do NOT add any new information or rephrase the question – only remove redundant parts.
8. Return ONLY the rewritten question text, nothing else.
9. Do not add placeholder tokens such as ¡image¿.

Original question: <question>
Rewritten question (visual-redundant info removed):

J Effect of Modality-Asymmetric Training Pairs

As described in Section 3, ODA-Data contains problems that exhibit modality asymmetry. To test whether this asymmetry is important, we compare MIRROR with a control model trained on the same number of properly paired examples that do not exhibit clear modality asymmetry, using identical hyperparameters. Training on asymmetric examples improves text-dominant validation pass@1 by 4.43 pp and image-dominant validation pass@1 by 22.15 pp relative to the paired-control set. These results show that paired representations alone are insufficient. The useful supervision arises when an empirically stronger view can provide problem-specific guidance to a view with very low rollout success, supporting our hypothesis that modality asymmetry is the primary source of self-supervision in MIRROR.

K Examples

K.1 Example 1

Summary

The correct answer is $3\sqrt{2}+2\sqrt{5}$. The image-dominant response realizes that P should sit on the intersection of AC and BD to minimize $|PA|+|PB|+|PC|+|PD|$ because this is intuitive from the diagram. The text-dominant response never realizes this and simply guess and exhausts different locations of P and only reaches a rough estimate of $|PA|+|PB|+|PC|+|PD| \approx 8.74$

Image-dominant Prompt

In the Cartesian coordinate plane, there are four fixed points A, B, C, D and a moving point P. What is the minimum value of $|PA|+|PB|+|PC|+|PD|$?

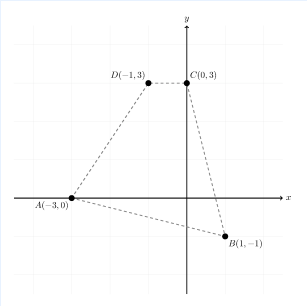


Image-dominant Response

[...]
Step 25: Consider that the optimal point may be the intersection of the perpendicular bisectors or something else
[...]
Final Answer: $3\sqrt{2}+2\sqrt{5}$ ✓

Text-dominant Prompt

In the Cartesian coordinate plane, there are four fixed points $A(-3,0)$, $B(1,-1)$, $C(0,3)$, $D(-1,3)$ and a moving point P. What is the minimum value of $|PA|+|PB|+|PC|+|PD|$?

Text-dominant Response

[...]
Let me try P = (0, 3) — 9.36
Or try P = (0, 2) — 9.18
Or try P = (-1, 2) — 8.848
Or try P = (0, 2.5) — 8.756
Or try P = (-0.5, 2.5) — 8.756
Or try P = (-0.5, 2.4) — 8.744
Or try P = (-0.5, 2.34) — 8.741
So perhaps the minimum is 8.74.
Maybe the answer is 8.74 ✗

K.2 Example 2

Summary

The correct answer is 873. The text-dominant response successfully reasons through the set up of this question while the image-dominant response misreads the radii of each semicircle as half of the side length of the triangle. The text-dominant response realizes that for a semicircle with diameter on one side of the triangle and be tangent to the other two sides, its diameter must be less than the side of the triangle it sits on.

Image-dominant Prompt

In a triangle, three semicircles can be drawn such that the diameter of each lies on a side of the triangle and each semicircle is tangent to the other two sides. The product of the areas of these three semicircles is numerically equal to the cube of the surface area of a certain sphere. The volume of this sphere can be expressed as $\frac{a\pi\sqrt{b}}{c}$, where a and c are relatively prime integers and b is the product of distinct primes. Find $a+b+c$.

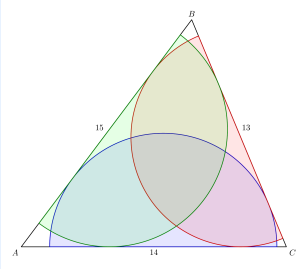


Image-dominant Response

Let's consider the semicircle drawn on side AB as diameter. Let M be the midpoint of AB. The semicircle with diameter AB has center M, radius $r = \frac{AB}{2} = 7.5$. For this semicircle to be tangent to sides AC and BC, the distance from point M to lines AC and BC must equal the radius $r = 7.5$.
[...]
So a + b + c = 455 + 2730 + 64 = 3249
3249 ✗

Text-dominant Prompt

In a triangle with side lengths 13, 14, and 15, three semicircles can be drawn such that the diameter of each lies on a side of the triangle and each semicircle is tangent to the other two sides. The product of the areas of these three semicircles is numerically equal to the cube of the surface area of a certain sphere. The volume of this sphere can be expressed as $\frac{a\pi\sqrt{b}}{c}$, where a and c are relatively prime integers and b is the product of distinct primes. Find $a+b+c$.

Text-dominant Response

[...]
But wait — in a triangle, if you draw a semicircle with diameter on side AB, and you want it to be tangent to sides AC and BC, then the semicircle must lie **inside** the triangle, and its center lies on AB, and it touches AC and BC.
[...]
So a = 784, b = 2, c = 87, a + b + c = 784 + 2 + 87 = 873
So 873 ✓

K.3 Example 3

Summary

The correct answer is A. The privileged response correctly reaches the answer because it has all the necessary condition. The image-dominant response, though having all necessary condition, it did not read the $AP = AD$ condition from the diagram and reached the wrong asnwer B

Image-dominant Prompt

In the circle O, the length of AB is twice the radius R. Then
(A) $AP^2 = PB \cdot AB$.
(B) $AP \cdot DO = PB \cdot AD$.
(C) $AB^2 = AD \cdot DE$.
(D) $AB \cdot AD = OB \cdot AO$.
(E) None of the above.

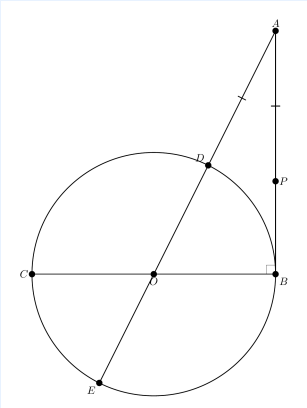


Image-dominant Response

We are given a circle with center O, and a point A outside the circle such that AB is a tangent to the circle at point B, and $AB = 2R$, where R is the radius.
Also, $AB \perp OB$, since a tangent is perpendicular to the radius at the point of contact. So triangle ABO is a right triangle with right angle at B.

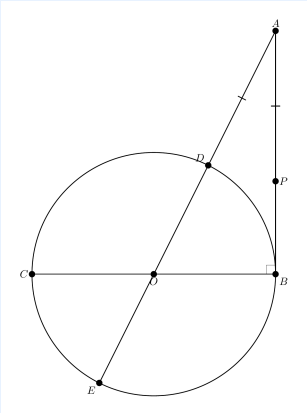
Moreover, D is a point on the circle and also on AO, so D lies on the line segment from A to O, and since O is the center, $OD = R$ (radius), and since AO is a line from A to the center, and $AB = 2R$, then AO is the hypotenuse of right triangle ABO, with legs $AB = 2R$ and $OB = R$.
Let's compute AO:
In right triangle ABO:

$$AO = \sqrt{AB^2 + OB^2} = \sqrt{(2R)^2 + R^2} = \sqrt{4R^2 + R^2} = \sqrt{5R^2} = R\sqrt{5}$$

Now, point D lies on AO, and $OD = R$, so since $AO = R\sqrt{5}$, then $AD = AO - OD = R\sqrt{5} - R = R(\sqrt{5} - 1)$.
We are to check which of the given options is true.
—
Let's analyze each option one by one.
—
[...]
—
Therefore, the correct answer is **(B)**.
—
Final Answer: (B) ✗

Combined Image+Text Prompt

In the circle O, BC is the diameter of the circle, and $AB \perp BC$. Also, $ADOE$ is a straight line, $AP = AD$, and the length of AB is twice the radius R. Then
(A) $AP^2 = PB \cdot AB$.
(B) $AP \cdot DO = PB \cdot AD$.
(C) $AB^2 = AD \cdot DE$.
(D) $AB \cdot AD = OB \cdot AO$.
(E) None of the above.



Combined Image+Text Response

Step 1: Understand the diagram and given data

- BC is the diameter → so $BC = 2R$, and O is the midpoint of BC.
- $AB \perp BC$ → so triangle ABC is a right triangle at B.
- $ADOE$ is a straight line → points A, D, O, E are colinear.
- $AP = AD$ → point P lies on AB, and $AP = AD$.
- $AB = 2R$.

[...]
—
Therefore, only **(A)** is correct.
—
Final Answer: (A) ✓

K.4 EXAMPLE 4

Summary

The image-dominant response wrongly reads the $5\sqrt{2}$ segment as AF instead of HF. Without the necessary and correct information, the image-dominant response never reaches a solution.

Image-dominant Prompt

Given that E is a point on the extension of side BC of square $ABCD$, and a square $AEFG$ is constructed with AE as one side. Then the length of BE is?

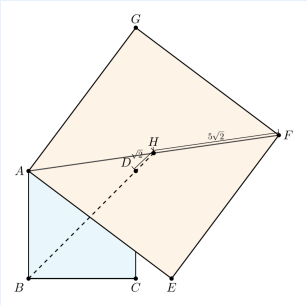


Image-dominant Response

We are given:
- Square $ABCD$, so all sides are equal and all angles are 90°. - Point E lies on the **extension** of side BC, meaning beyond point C. - A square $AEFG$ is constructed with side AE. - We are given the length $AF = 5\sqrt{2}$, and also the length $DH = \sqrt{2}$ (where H is the point on AF such that DH is drawn and labeled as $\sqrt{2}$).
We are to find the length of BE.
[...]
✗

Text-dominant Prompt

Given that E is a point on the extension of side BC of square $ABCD$, and a square $AEFG$ is constructed with AE as one side, such that points A and G are on the same side of BE. The extension of BD intersects AF at point H, with $HD = \sqrt{2}$ and $FH = 5\sqrt{2}$. Then the length of BE is?

Text-dominant Response

We are given a square $ABCD$, and point E is on the extension of side BC. A square $AEFG$ is constructed with AE as one side, and points A and G are on the same side of line BE. The extension of diagonal BD intersects line AF at point H, with $HD = \sqrt{2}$ and $FH = 5\sqrt{2}$. We are to find the length of BE. [...]
Final Answer: 8 ✓

K.5 EXAMPLE 5

Summary

The text-dominant response treats 'Square A and B share a common side' as side length equal and could never solve the problem.

Text-dominant Prompt

Square A, with side length 20 meters, has an area equal to the areas of square B and square C combined. Square A and square B share a common side. Square B and square C share a common side as well. What is the sum, in meters, of the combined perimeters of square B and square C?

Text-dominant Response

We are given:
- Square A has side length **20 meters**, so its area is:

$$\text{Area}_A = 20 \times 20 = 400 \text{ m}^2$$

- The area of square A equals the **combined areas** of square B and square C:

$$\text{Area}_B + \text{Area}_C = 400$$

- Square A and square B **share a common side**, meaning they are adjacent — likely side by side or stacked, but since both are squares and share a side, **their side lengths must be equal**. Why? Because if two squares share a side, that side must have the same length for both squares. So:

$$\text{Side length of square B} = \text{Side length of square A} = 20 \text{ meters}$$

Wait — but that would make area of B = 400 m², and then area of C would be 0, which is impossible.
So maybe they don't share a side of equal length? But the problem says "share a common side". That implies that the side that is common must be of the same length for both squares.
So if square A and square B share a side, then they must both have that side as their side length. But then square B would also have side 20, so area = 400, and C would have area 0 — impossible.
Therefore, perhaps the common side is **not** the side of the square — that doesn't make sense.
[...]
✗

Image-dominant Prompt

Square A has an area equal to the areas of square B and square C combined. What is the sum, in meters, of the combined perimeters of square B and square C?

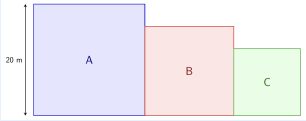


Image-dominant Response

We are given that Square A has an area equal to the combined areas of Square B and Square C. Also, the side length of Square A is given as **20 meters**.
—
Step 1: Find the area of Square A
[...]
Final Answer: 112 meters ✓