MITIGATING MODALITY INTERFERENCE FOR UNIFIED REASONING AND PERCEPTION IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-only reasoning models show that System-2 deliberation greatly improves mathematics and code, yet multimodal large language models (MLLMs) still rely largely on fast visual intuition. A dominant recipe for transferring reasoning to MLLMs converts images into caption-like intermediates and trains distilled long-form reasoning traces jointly with multimodal perception supervision under a single objective, which often trades perceptual competence for reasoning or vice versa. Through controlled pilot studies, we find that this trade-off is consistent with *modality interference*: under naive joint optimization, reasoning and perception objectives yield non-additive gains and negatively aligned gradients. Building on this diagnosis, we propose **UPRSD** (Unified Perceptual and Reasoning via Staged Distillation), a three-stage recipe that first decouples the two behaviors with prefix triggers, then unifies them under a shared reasoning interface, and finally applies verifiable-reward reinforcement learning with potency-aware data selection. At both 4B and 8B scales, the resulting UNI-VL models improve multimodal and text-only reasoning over open-source baselines while maintaining strong general multimodal understanding.

1 INTRODUCTION

Recent advances in text-only Large Reasoning Models (LRMs) have shown that introducing System-2-style deliberation can substantially improve performance on challenging tasks such as mathematics and code generation (de Winter et al., 2024; Guo et al., 2025a; Lightman et al., 2024; Gao et al., 2023; Li et al., 2022; Nijkamp et al., 2023). In contrast, multimodal large language models (MLLMs), despite rapid progress in image understanding (Zhang et al., 2025), visual question answering (Kuang et al., 2025), and video evaluation (Han et al., 2025), still tend to rely on fast, System-1-like visual intuition, and systematic gains in visuospatial cognition remain limited even under strong prompting (Rudman et al., 2025; Yang et al., 2024; Huang et al., 2025). Recent commercially deployed multimodal reasoning systems (*e.g.*, GPT-4.1, Claude-4-Sonnet) underscore that System-2 reasoning is essential for reliable, generalizable multimodal intelligence. Reaching this goal requires both strong perceptual grounding (Ma et al., 2025) and verifiable reasoning (Sun et al., 2025), but a robust and reproducible open training recipe that delivers both is still lacking.

However, bringing System-2 reasoning into open MLLMs remains challenging. A dominant line of work adopts a textual-proxy transfer recipe that implements naive mixed supervision: images are converted into caption-like textual intermediates, and long-form reasoning trajectories distilled from strong text-only LLM reasoners are trained jointly with multimodal perception supervision under a single objective (Huang et al., 2026; Yang et al., 2025b). Yet, this paradigm often comes with a trade-off that limits certain capabilities: boosting deliberative reasoning can undermine broad perceptual competence, while preserving perception may fail to produce stable, transferable reasoning. As a result, gains are often confined to reasoning-heavy benchmarks, with inconsistent improvement on broad multimodal understanding, making it difficult to achieve robust and reproducible reasoning for MLLMs (Sun et al., 2025).

Through a systematic empirical investigation using diagnostic indicators (Section 2), we find that these issues are more consistently explained by modality interference. Specifically, reasoning and

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.

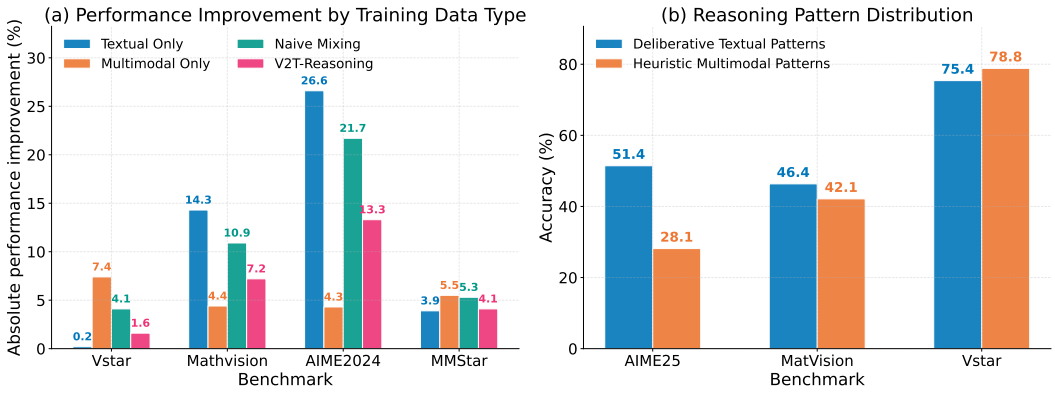


Figure 1: **Empirical analysis of supervision composition and enforced inference patterns.** (a) Absolute gains under different distillation compositions: text-only distillation mainly improves text-dominant benchmarks (*e.g.*, AIME2024), while multimodal distillation is critical for visually grounded tasks (*e.g.*, VStar). (b) After naive mixed distillation, enforcing deliberative-textual versus visually grounded heuristic patterns yields different accuracies across benchmarks, indicating that inference patterns interact with task types.

perception can induce conflicting learning signals during gradient-based optimization: reasoning tasks favor long-horizon, structured deliberation, whereas perceptual tasks emphasize short-horizon, evidence-localized correctness with broad coverage (Rudman et al., 2025). When optimized jointly with roughly equal weighting, the resulting gradients often counteract one another and create a tug-of-war in the capability space (Yang et al., 2024), causing the model to over-specialize toward long-form reasoning and lose perceptual coverage, or to remain perception-dominant and fail to develop stable reasoning mechanisms.

To address these challenges, we introduce an MLLM training framework **UPRSD** (Unified Perceptual and Reasoning via Staged Distillation), injecting reusable System-2 reasoning into an open-source MLLM (named **Uni-VL** in this work) while preserving broad perceptual grounding. UPRSD consists of three stages: (1) **Structural Priming.** Separate reasoning-oriented and perception-oriented behaviors via a lightweight routing interface, and apply a prudential protocol on visual STEM inputs to elicit reliable multimodal reasoning trajectories while reducing early-stage interference. (2) **Logic Integration.** Standardize the reasoning interface through a modality-consistent `<think>` format, encouraging deliberation when needed and suppressing unnecessary overthinking on shallow perception tasks to preserve efficiency and coverage. (3) **Reinforced Evolution.** Perform verifiable-reward reinforcement optimization via GRPO (Shao et al., 2024), together with potency-aware data selection that prioritizes hard and high-potential instances to strengthen frontier reasoning and improve sample efficiency.

In summary, our contributions are as follows:

- We first provide a systematic and reproducible empirical characterization of modality interference in MLLMs, showing that naively mixing textual CoT supervision with visual supervision leads to suboptimal convergence and catastrophic forgetting on general multimodal benchmarks, and yielding actionable diagnostics for principled unification.
- We introduce **UPRSD**, a three-stage unification framework (Structural Priming, Logic Integration, and Reinforced Evolution) to train Uni-VL, that progressively transfers System-2 reasoning priors into open-source MLLMs while maintaining direct grounding in visual evidence.
- We curate and release high-quality SFT Reasoning and RL datasets, using instance selection based on semantic balancing, clustering, and potency mining to improve training stability and sample efficiency.
- Across different scales (4B/8B) and benchmarks, **Uni-VL** trained with **UPRSD** improves multimodal and text reasoning while preserving strong general multimodal perceptual understanding.

2 PRELIMINARY EMPIRICAL ANALYSIS

We first run pilot experiments to probe why unifying perceptual grounding and System-2 reasoning is difficult for open-source MLLMs. We ask (1) whether naively mixing text reasoning distillation with multimodal distillation yields additive gains, and (2) how enforced inference patterns interact with task types. Figure 1 summarizes the findings, and Appendix A.2 details the setup.

2.1 NAIVE MIXING YIELDS NON-ADDITIVE GAINS

We compare three regimes that differ only in distillation composition: text reasoning distillation ($\mathcal{D}_{text}$), multimodal distillation ($\mathcal{D}_{mm}$), and their naive mixture ($\mathcal{D}_{mix}$), all initialized from QWEN3-VL-8B-INSTRUCT and trained with identical hyperparameters. To equalize the training budget, the mixture is trained for one epoch and each single-source regime for two, so that every model is updated on a comparable number of samples (about 700k). As shown in Figure 1a, the gains are strongly benchmark-dependent: text-only distillation improves AIME24 by 26.6 points, whereas multimodal distillation contributes most on visually grounded tasks such as VSTAR (Wu & Xie, 2023). Naive mixing is not additive: $\mathcal{D}_{mix}$ often yields weaker improvements than the specialized regimes on both reasoning- and perception-oriented benchmarks. Non-additivity alone does not identify its cause, since data imbalance, optimization instability, or task-specific learning-rate needs could produce a similar pattern. The mixture is roughly balanced (321k textual vs. 374k multimodal samples), which argues against simple imbalance but does not rule out the others. We therefore treat modality interference as a working hypothesis and probe it further via gradient alignment (Section 3.3) and ablations (Section 4.3).

2.2 INFERENCE PATTERNS ARE TASK-DEPENDENT

We further test whether a single inference style generalizes across tasks by enforcing two prompting patterns *after* naive mixed distillation: a deliberative textual pattern and a visually grounded heuristic pattern. Figure 1b shows a clear interaction with benchmark type. Deliberative reasoning is far better on logic-dense tasks (51.4 vs. 28.1 on AIME25) and remains ahead on MATHVISION (46.4 vs. 42.1), while the visually grounded pattern is better on perception-centric benchmarks (78.8 vs. 75.4 on VSTAR), where additional deliberation brings little benefit and can introduce spurious errors. These probes suggest that unification requires explicit control over deliberation versus direct grounding, which motivates a staged recipe that reduces interference while integrating reasoning and perception.

3 METHODOLOGY

UPRSD is a progressive training recipe that transfers the latent reasoning abilities of large language models (LLMs) to the multimodal domain. Let P_θ denote an MLLM with parameters θ. Given a visual input x_v and a textual instruction x_t (jointly x), we aim to maximize the likelihood of the correct answer y. Rather than modeling $P_\theta(y \mid x)$ directly, we introduce a latent reasoning trace z (the chain of thought) and maximize the marginal likelihood $P_\theta(y \mid x) = \sum_z P_\theta(y \mid z, x)\, P_\theta(z \mid x)$. UPRSD optimizes this objective in three progressive stages: structural priming, logic integration, and reinforced evolution (Figure 2). We refer to the resulting model as UNI-VL.

3.1 DATA CURATION ENGINE

The main obstacle to a multimodal System-2 model is the scarcity of high-quality multimodal reasoning supervision. We start from a foundational pool of **374k multimodal** and **321k textual** instances with ground-truth answers, drawn from diverse STEM and logical-reasoning sources (Figure 3). Multimodal reasoning trajectories are bootstrapped from this pool and verified (Section 3.2); an example reasoning trace produced by UNI-VL is shown in Figure 6. To balance STEM, logic, and general perception domains without redundancy, we embed prompts with NV-EMBED-V2 (Lee et al., 2024) and cluster them with HDBSCAN (Campello et al., 2013), pruning over-represented shallow clusters and up-sampling rare but important reasoning categories. For decontamination, we remove training samples with 13-gram overlap with any evaluation benchmark, apply a stricter 7-gram filter to short-answer math problems, and exclude problems from known math competitions

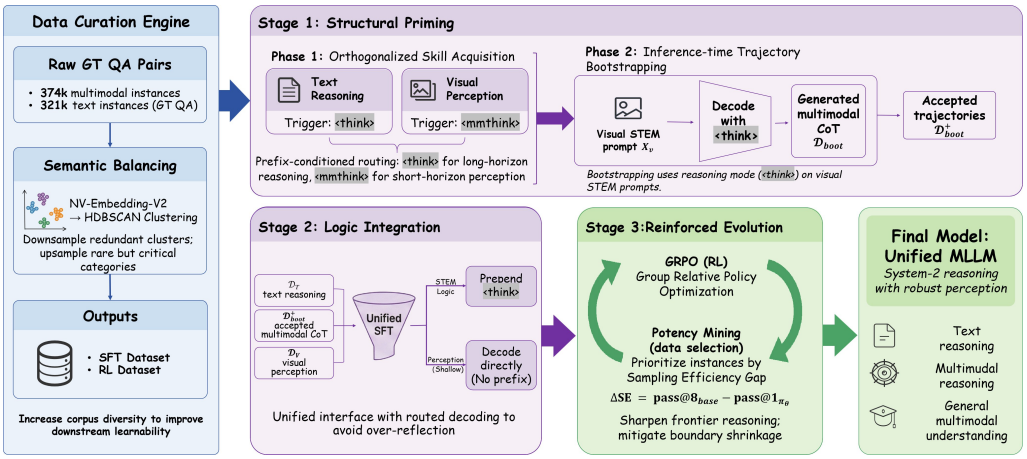


Figure 2: **UPRSD pipeline for training UNI-VL. Left:** data curation builds SFT and RL corpora via semantic balancing. **Stage 1 (Structural Priming)** reduces early interference with prefix routing and bootstraps grounded multimodal reasoning trajectories. **Stage 2 (Logic Integration)** unifies the traces under a modality-consistent `<think>` interface, with routed decoding to avoid overthinking on shallow perception. **Stage 3 (Reinforced Evolution)** applies verifiable-reward GRPO with potency-mined data to sharpen frontier reasoning without sacrificing perceptual coverage.

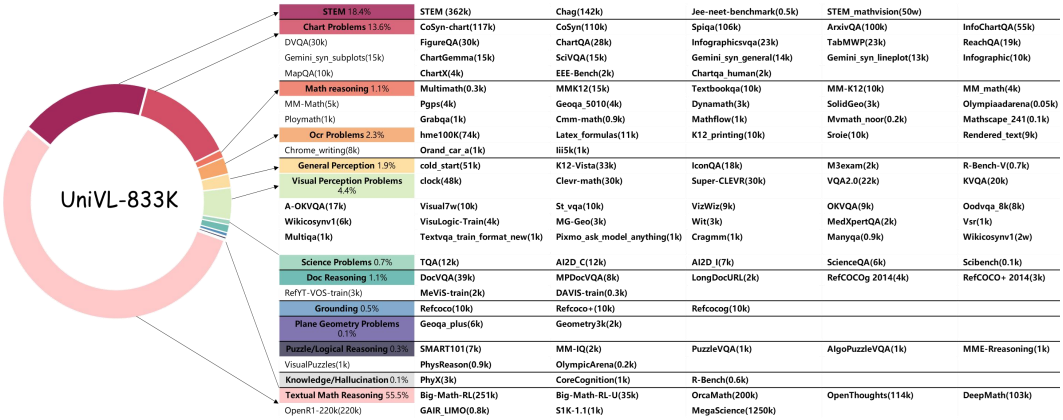


Figure 3: **Composition of the UniVL-833K training corpus.** The donut chart shows the category distribution, spanning text math reasoning, STEM, chart/OCR understanding, visual perception, document reasoning, grounding, and other multimodal reasoning tasks; the right panel lists representative source datasets with instance counts.

(*e.g.*, AIME and AMC) published after 2023 (Appendix A.9). The resulting SFT and RL data form the UniVL-833K corpus, which we will release.

3.2 STRUCTURAL PRIMING

To mitigate interference early in training, we decouple *System-2* reasoning and *System-1* perception through prefix-conditioned routing. UNI-VL keeps a single tokenizer and vocabulary; the separation comes from conditioning generation on distinct *behavioral triggers* that steer it toward long-horizon deliberation (*reasoning mode*) or short-horizon perceptual answering (*perception mode*).

Phase 1: Orthogonalized Skill Acquisition. We optimize θ on textual reasoning data $\mathcal{D}_T$ and visual perception data $\mathcal{D}_V$ with the joint objective

$$\mathcal{L}_{S1}(\theta) = \mathbb{E}_{\mathcal{D}_T}\big[-\log P_\theta(y \mid x, S_{reas})\big] + \mathbb{E}_{\mathcal{D}_V}\big[-\log P_\theta(y \mid x, S_{perc})\big], \tag{1}$$

where S_{reas} = `<think>` and S_{perc} = `<mmthink>` are behavior triggers. Conditioning the two objectives on different prefixes is intended to reduce destructive interference between the gradients

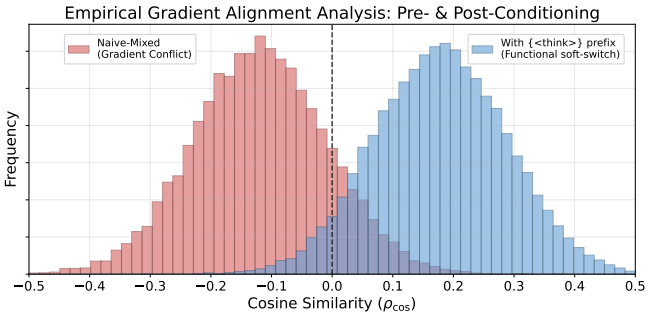


Figure 4: **Gradient alignment.** Distribution of cosine similarities between $\mathcal{D}_T$ and $\mathcal{D}_V$ gradients. Naive mixing produces conflicting gradients, whereas prefix conditioning with `<think>` shifts them toward positive alignment.

induced by long-form reasoning traces and those induced by short perceptual answers; we quantify this effect with gradient cosine similarity in Section 3.3.

Phase 2: Bootstrapping Multimodal Trajectories. We bridge the modality gap via **inference-time trajectory bootstrapping**. Given a visual STEM prompt x_v, we force the reasoning trigger so that the model approximates the textual reasoning distribution:

$$\hat{y}_{boot} = \arg\max_{y} \prod_{t=1}^{L} P_\theta(y_t \mid x_v, S_{reas}, y_{<t}). \tag{2}$$

Forcing the S_{reas} prefix leverages *cross-modal latent transfer* to synthesize multimodal reasoning trajectories. To avoid learning from incorrect self-generated chains, we apply **trajectory quality control**: (i) *answer verification* with task-specific verifiers (symbolic equivalence for math and exact-match option extraction for multiple-choice questions); (ii) *self-consistency filtering*, which samples $K{=}4$ trajectories and keeps the majority-answer cluster, breaking ties by average log-probability; and (iii) *sanity filters* that drop empty, malformed, or excessively long outputs. The accepted set $\mathcal{D}^+_{boot} = \{(x_v, \hat{y}^+_{boot})\}$ serves as multimodal reasoning supervision in later stages.

3.3 Logic Integration

The second stage dissolves the boundary between the two modes and merges them into a single System-2 interface. We train with a unified cross-entropy objective; reducing the divergence between modality-conditioned behaviors motivates this stage but is not an explicit loss term.

Universal Output Space Alignment. We standardize all reasoning supervision under one protocol with the unified SFT objective

$$\mathcal{L}_{S2}(\theta) = -\sum_{(x,y)\in\mathcal{D}_{unified}} \log P_\theta(y \mid x, \texttt{<think>}), \tag{3}$$

where $\mathcal{D}_{unified} = \mathcal{D}_T \cup \mathcal{D}^+_{boot} \cup \mathcal{D}_{perc}$. To avoid hallucinated reasoning on "shallow" perception tasks (*e.g.*, OCR), we apply a routing-based bypass:

$$y \sim \begin{cases} P_\theta(\cdot \mid x, \texttt{<think>}) & \text{if } x \in \text{STEM/Logic}, \\ P_\theta(\cdot \mid x, \emptyset) & \text{if } x \in \text{Perception}. \end{cases} \tag{4}$$

Routing relies on dataset metadata during training and, at inference, on a lightweight task classifier that decides whether to prepend `<think>` or to decode directly; this 22M-parameter DeBERTa-v3 model reaches 98.4% accuracy on held-out queries and adds about 5 ms per query (Appendix A.7). This makes `<think>` a universal cognitive trigger while keeping latency low on perception-heavy tasks.

Does interference re-emerge after integration? Stage 2 merges the two behaviors only after Stage 1 has established them under separate triggers, so the shared interface is learned from an

initialization in which both skills already exist, rather than acquired while the two compete for the same parameters; UPRSD can thus be viewed as a curriculum over supervision interfaces. To test whether integration reintroduces conflict, we compute the cosine similarity between gradients of textual reasoning ($\mathcal{D}_T$) and multimodal ($\mathcal{D}_V$) mini-batches on the shared LLM parameters (protocol in Appendix A.3):

$$\rho_{cos} = \frac{\nabla_\theta \mathcal{L}(\mathcal{D}_T) \cdot \nabla_\theta \mathcal{L}(\mathcal{D}_V)}{\|\nabla_\theta \mathcal{L}(\mathcal{D}_T)\| \, \|\nabla_\theta \mathcal{L}(\mathcal{D}_V)\|}. \tag{5}$$

Naive mixing yields a negative mean similarity ($\rho_{cos} \approx -0.12$), whereas at the end of Stage 2 the mean is positive ($\rho_{cos} \approx 0.18$; Figure 4). In a parameter space with billions of dimensions, cosine similarities between mini-batch gradients are typically small in magnitude, and the sign of the mean is the more informative quantity: a negative value means that, to first order, a step that decreases one loss increases the other. This evidence is correlational: it shows that conflict is reduced under UPRSD, not that the reduction alone causes the downstream gains. Moreover, Stage 2 also introduces $\mathcal{D}^{+}_{boot}$, so its gains cannot be attributed to integration alone (Section 4.3).

3.4 Reinforced Evolution

Starting from the SFT model, we refine reasoning trajectories with Group Relative Policy Optimization (GRPO; Shao et al., 2024), which estimates advantages from groups of sampled responses instead of a learned value model. For each prompt x, we sample G responses $\{y_i\}_{i=1}^{G}$ from the current policy $\pi_{\theta_{\text{old}}}$ and maximize

$$\mathcal{J}(\theta) = \mathbb{E}\Big[\frac{1}{G}\sum_{i=1}^{G} \min\big(\rho_i \hat{A}_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\,\hat{A}_i\big)\Big] - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big), \tag{6}$$

where $\rho_i = \pi_\theta(y_i \mid x)/\pi_{\theta_{\text{old}}}(y_i \mid x)$, ϵ controls the trust region, β weights the KL penalty to the reference policy π_{ref}, and $\hat{A}_i = (r_i - \bar{r})/\sigma_r$ is the group-relative advantage, with $\bar{r}$ and σ_r the mean and standard deviation of the rewards within the group. The reward is a weighted composite

$$r_i = (1 - \lambda_{\text{fmt}})\, r_{\text{ans}} + \lambda_{\text{fmt}}\, r_{\text{fmt}}, \tag{7}$$

where r_{ans} comes from task-specific verifiers (*e.g.*, symbolic equivalence for math or IoU for visual grounding) and r_{fmt} rewards adherence to the `<think>...<answer>` template.

Mitigating Boundary Shrinkage via *pass@k* Filtering. Recent studies (Yue et al., 2025) suggest that reinforcement learning with verifiable rewards (RLVR) often acts as a "distribution sharpening" mechanism: it improves sampling efficiency ($pass@1$) by concentrating on reasoning paths already latent in the model rather than expanding its capability frontier, which risks shrinking the coverage of solvable problems ($pass@k$). To counteract this, we select RL data by *potency mining*, prioritizing samples with a large sampling-efficiency gap

$$\Delta SE = \text{pass@8}_{\text{base}} - \text{pass@1}_{\pi_\theta}. \tag{8}$$

Using $pass@8_{\text{base}}$ as a proxy for latent potential, we target problems that the base model *can* solve through exploration but the current policy fails to solve *reliably*; mining is performed on a held-out pool disjoint from the RL prompts (Appendix A.10). The resulting RL corpus contains **56k textual** and **82k multimodal** samples, so that System-2 behaviors are reinforced in both modalities.

4 Experiments

We study four questions: whether Uni-VL improves reasoning on multimodal and textual benchmarks (RQ1), whether its reasoning transfers to purely textual mathematics (RQ2), whether the recipe scales across model sizes (RQ3), and how each stage of UPRSD contributes (RQ4).

4.1 Experimental Setup

Benchmarks. We evaluate three groups of benchmarks. *Multimodal logical reasoning*: MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), MathVista (Lu et al., 2023), DynaMath (Zou et al., 2025), WeMath (Qiao et al., 2024), Gaokao-MM (Zong & Qiu, 2024),

Table 1: **Comparison on multimodal and textual reasoning benchmarks.** and mark the best and second-best scores within each model group; the Δ rows give the gain of UNI-VL over the best open-source baseline of the same scale.

Model	Multimodal Reasoning											Textual Reasoning						
	MVerse	MVision	MathVista	DynaM	WeMath	LogVista	MMMU	MMMU-Pro	CharXiv	MMStar	Avg.	AIME24	AIME25	M500	MMLU-P	BBEH	GPQA	Avg.
Closed-Source																		
GPT-4.1	*59.8*	*51.8*	72.0	*48.3*	55.5	*63.8*	75.0	*65.0*	*55.8*	71.2	61.8	46.7	23.3	89.6	80.2	32.7	67.7	56.7
Gemini-2.5-Pro	**81.2**	55.3	83.8	57.1	**78.0**	**75.2**	**82.0**	**76.5**	**69.3**	**79.7**	**73.8**	66.7	**88.0**	85.8	81.6	17.7	**82.3**	70.4
Claude-4-Sonnet	66.1	54.6	70.4	46.9	63.0	64.4	74.4	60.7	58.4	66.9	62.6	46.7	33.3	92.0	83.2	35.0	68.7	59.8
Seed-1.5-VL-Pro	80.4	**68.7**	**85.6**	**60.5**	**78.0**	71.8	77.9	67.6	63.4	78.2	73.2	**80.0**	53.3	**96.8**	**83.5**	**39.7**	71.2	**70.8**
3B/4B-scale MLLMs																		
Qwen3-VL-4B	*33.0*	21.2	62.3	*17.0*	*17.9*	35.8	53.1	31.6	*25.3*	51.1	34.8	0.0	0.0	55.8	41.5	7.0	29.3	22.3
FAST-3B	37.2	26.8	66.2	*15.4*	*23.3*	*35.4*	*52.0*	34.6	28.3	55.2	37.4	6.7	3.3	55.4	34.6	8.6	25.3	22.3
VLAA-Thinker-3B	36.4	24.4	61.0	18.2	33.8	38.5	*49.7*	*33.3*	28.2	53.4	37.7	3.3	0.0	61.4	33.3	8.1	27.8	22.3
Uni-VL-4B (Ours)	**48.2**	**44.2**	**66.8**	**24.6**	**40.8**	**43.5**	**56.8**	**45.2**	**34.4**	**56.7**	**46.1**	**41.5**	**34.8**	**86.2**	**55.8**	**15.0**	**41.4**	**45.8**
Δ (vs. best baseline)	+11.0	+17.4	+0.6	+6.4	+7.0	+5.0	+3.7	+10.6	+6.1	+1.5	+8.4	+34.8	+31.5	+24.8	+14.3	+6.4	+12.1	+23.5
7B/8B-scale MLLMs																		
Qwen3-VL-8B	*42.9*	25.1	68.2	*21.2*	36.2	*45.0*	58.6	38.3	*35.5*	62.1	43.3	6.7	0.0	66.0	50.5	10.8	36.9	28.5
OpenVLThinker-7B	*40.7*	25.9	72.3	*21.2*	37.9	*41.4*	58.7	42.9	*39.3*	62.9	44.3	0.0	0.0	63.2	49.3	13.6	31.3	26.2
MM-Eureka-7B	*49.6*	26.9	73.0	*24.0*	34.7	*46.8*	*57.3*	43.3	*39.5*	64.4	46.0	13.3	0.0	65.4	53.4	12.2	41.4	31.0
MMR1-Math-v0	45.1	30.2	71.0	*25.2*	*33.2*	50.8	*57.1*	*43.2*	*39.3*	63.8	45.9	3.3	3.3	66.6	51.8	11.4	32.3	28.1
ThinkLite-7B-VL	*45.3*	32.9	75.1	*22.0*	*26.5*	*40.7*	55.5	*41.3*	*39.3*	63.4	44.2	3.3	3.3	67.6	51.9	11.3	35.9	28.9
VLAA-Thinker-7B	48.2	26.4	68.0	22.4	*29.2*	48.5	*54.6*	*41.6*	*36.1*	64.5	44.0	6.7	3.3	68.6	51.2	11.7	28.8	28.4
VL-Rethinker-7B	*49.1*	32.3	74.9	27.4	*27.8*	*44.5*	56.7	41.7	*39.8*	62.7	45.7	3.3	3.3	64.4	52.0	12.9	31.8	28.0
Vision-G1	50.0	31.3	**76.1**	27.2	29.0	50.2	53.4	41.2	41.0	63.1	46.2	3.3	0.0	69.0	53.0	13.9	31.3	28.4
Uni-VL-8B (Ours)	**56.5**	**54.7**	75.6	**39.7**	**52.7**	**54.1**	**67.9**	**54.8**	**47.0**	**66.7**	**57.0**	**65.0**	**61.7**	**95.2**	**67.0**	**18.9**	**57.6**	**60.9**
Δ (vs. best baseline)	+6.5	+21.8	−0.5	+12.3	+14.8	+3.3	+9.2	+11.5	+6.0	+2.2	+10.8	+51.7	+58.4	+26.2	+13.6	+5.0	+16.2	+29.9

and ZEROBENCH (Roberts et al., 2025). *General multimodal understanding*: LOGICVISTA (Xiao et al., 2024), MMMU (Yue et al., 2024a), MMMU-PRO (Yue et al., 2024b), CHARXIV (Wang et al., 2024c), HALLUSIONBENCH (Guan et al., 2024), COUNTBENCH (Paiss et al., 2023), and MMSTAR (Chen et al., 2024). *Text-based reasoning*: AIME24/25 (Li et al., 2024b) and MATH-500 (Hendrycks et al., 2021) for competition mathematics, and GPQA (Rein et al., 2024), MMLU-PRO (Wang et al., 2024b), and BBEH (Kazemi et al., 2025) for general-purpose reasoning. All open-source models are evaluated with the same pipeline and sampling settings (temperature 0.6, top-p 0.9, and up to 30k generated tokens; Appendix A.4).

Baselines. We compare against three groups of models. *Proprietary frontier models*: GPT-4.1 (OpenAI, 2025), GEMINI-2.5-PRO-THINKING (Comanici et al., 2025), CLAUDE-4-SONNET (Anthropic, 2025), and SEED-1.5-VL-PRO (Doubao-1.5-Thinking-Vision-Pro; Guo et al., 2025b). *3B/4B-scale open-source MLLMs*: QWEN3-VL-4B-INSTRUCT (Bai et al., 2025a), FAST-3B (Xiao & Gan, 2025), and VLAA-THINKER-3B (Chen et al., 2025a). *7B/8B-scale open-source MLLMs*: QWEN3-VL-8B-INSTRUCT, OPENVLTHINKER-1.2-7B (Deng et al., 2025), MM-EUREKA-QWEN-7B (Meng et al., 2025), MMR1-MATH-V0 (Leng et al., 2025), THINKLITE-7B-VL (Wang et al., 2025b), VLAA-THINKER-7B, VL-RETHINKER-7B (Wang et al., 2025a), and VISION-G1 (Zha et al., 2025). Most of the other open-source baselines are fine-tuned from QWEN2.5-VL (Bai et al., 2025b) backbones of comparable size.

Implementation details. UNI-VL-4B and UNI-VL-8B are initialized from QWEN3-VL-4B-INSTRUCT and QWEN3-VL-8B-INSTRUCT, respectively. The supervised stages (Stages 1 and 2) train for 5 epochs with a learning rate of 2×10^{-5}, cosine decay with 5% warmup, and sequences of up to 32,768 tokens. Stage 3 runs GRPO for 1,000 updates with a global batch size of 128, an actor learning rate of 1×10^{-6}, a KL coefficient of 2×10^{-3}, and prompt and response lengths capped at 8,192 tokens. All stages use BF16 and DeepSpeed ZeRO-2 on 8×A100 (80GB) GPUs (Appendix A.1).

4.2 MAIN RESULTS

Multimodal reasoning (RQ1). As shown in Table 1, UNI-VL-8B averages 57.0 on the multimodal benchmarks, 10.8 points above the strongest open-source baseline (Vision-G1, 46.2). The largest gains appear on visually intensive tasks such as MATHVISION (+21.8) and DYNAMATH (+12.3). General understanding also improves (MMMU +9.2, MMMU-PRO +11.5, CHARXIV +6.0, MMSTAR +2.2), and the only regression is a 0.5-point drop on MATHVISTA, consistent with staged training deepening reasoning without eroding perception. Across three Stage-3 runs with different seeds, validation accuracy varies by only about ±0.8% (Appendix A.8).

Table 2: **Comparison with RAPID** on the seven benchmarks it reports (RAPID numbers from Gou et al. (2026)). RAPID builds on QWEN2.5-VL-3B/7B, UNI-VL on QWEN3-VL-4B/8B; "Params" counts all parameters used at inference; bold: better result per scale.

Model	Params	MathVista	MVision	MVerse	MMMU	WeMath	DynaM	LogVista	Avg.
RAPID (3B + Qwen3-8B)	11B	**69.6**	40.8	**48.6**	**60.9**	39.1	**29.3**	**56.4**	**49.2**
UNI-VL-4B	4B	66.8	**44.2**	48.2	56.8	**40.8**	24.6	43.5	46.4
RAPID (7B + Qwen3-8B)	15B	**76.1**	43.7	52.2	64.7	45.4	32.7	**57.7**	53.2
UNI-VL-8B	8B	75.6	**54.7**	**56.5**	**67.9**	**52.7**	**39.7**	54.1	**57.3**

Table 3: **Stage-wise ablation of UPRSD** (UNI-VL-8B). Each row adds one stage; parentheses give gains over the Naive-Mixed baseline; best in **bold**, second-best underlined.

Stage	AIME24	CVBench	MathVision	Avg.
Naive-Mixed (Baseline)	33.2	75.4	32.3	46.97
After Stage 1	44.6 (+11.4)	75.75 (+0.4)	40.1 (+7.8)	53.48 (+6.5)
After Stage 2	<u>58.9</u> (+25.7)	<u>78.3</u> (+2.9)	<u>49.2</u> (+16.9)	<u>62.13</u> (+15.2)
After Stage 3 (Uni-VL)	**65.0** (+31.8)	**81.8** (+6.4)	**54.7** (+22.4)	**67.17** (+20.2)

Textual reasoning transfer (RQ2). UNI-VL-8B reaches 65.0 on AIME24 and 61.7 on AIME25 (pass@1), whereas no 7B/8B open-source baseline exceeds 13.3, and its textual average is 29.9 points above the best baseline. This average (60.9) also exceeds those of GPT-4.1 (56.7) and CLAUDE-4-SONNET (59.8), although a clear gap remains to GEMINI-2.5-PRO and SEED-1.5-VL-PRO. Stage 3 alone contributes 6.1 of the AIME24 points (Table 3), while GAOKAO-MM and MMSTAR rise steadily during RL and ZEROBENCH improves later (Appendix A.6).

Scalability across model sizes (RQ3). UNI-VL-4B is the best model in the 3B/4B group, improving the multimodal and textual averages by 8.4 and 23.5 points over the best baseline of that group. It even matches the best 7B/8B open-source baseline on the multimodal average (46.1 vs. 46.2) and surpasses all of them on the textual average (45.8 vs. at most 31.0).

Comparison with perception–reasoning decoupling. RAPID (Gou et al., 2026) addresses the same trade-off by pairing a QWEN2.5-VL captioner with an external QWEN3-8B reasoner (Yang et al., 2025a). On the seven benchmarks it reports (Table 2), the single UNI-VL-8B model outperforms the 15B RAPID system by 4.1 points on average, with the largest margins on MATHVISION (+11.0) and WEMATH (+7.3). At the smaller scale, where the external reasoner is much larger than the model it assists, RAPID is ahead of UNI-VL-4B by 2.8 points. RAPID's numbers come from its paper and a different evaluation pipeline, and UNI-VL builds on the newer QWEN3-VL backbone, so these gaps are indicative rather than controlled (Appendix A.5).

4.3 ABLATION STUDY

Table 3 dissects each stage's contribution (RQ4) relative to a Naive-Mixed baseline. Stage 1 mainly improves reasoning-heavy benchmarks (AIME24 +11.4, MATHVISION +7.8) and leaves the vision-centric CVBENCH (Tong et al., 2024) nearly unchanged (+0.4). Stage 2 gives the largest single increment (+14.3 on AIME24 and +9.1 on MATHVISION over Stage 1), and Stage 3 improves all three benchmarks, contributing more than half of the total CVBENCH gain (+3.5 of +6.4), for a total average gain of 20.2 points. Because the unified `<think>` interface and the bootstrapped trajectories $\mathcal{D}^{+}_{boot}$ enter together in Stage 2, this ablation does not separate the effect of integration from that of the added multimodal reasoning data. The gradient analysis in Section 3.3 indicates no renewed conflict after integration, but a cleaner Stage-2-only ablation remains future work.

5 RELATED WORK

Multimodal reasoning models. Reasoning has moved from text-only LLMs to MLLMs (de Winter et al., 2024; Xu et al., 2025a; Chen et al., 2025c). Early work aligned visual encoders with LLMs (Radford et al., 2021; Li et al., 2024a) or fine-tuned MLLMs on chain-of-thought data,

initially built with template-based prompting (Xu et al., 2025b) on top of multimodal QA benchmarks (Lu et al., 2022; Liu et al., 2024). Following DeepSeek-R1 (Guo et al., 2025a), many works apply reinforcement learning with verifiable rewards to visual reasoning, including MM-Eureka (Meng et al., 2025), R1-V (Chen et al., 2025b), Revisual-R1 (Chen et al., 2026), and Visual-RFT (Liu et al., 2025). Closest to our goal, Open Vision Reasoner (Wei et al., 2025) transfers cognitive behaviors from a large text-only cold start to visual reasoning through multimodal RL. Many open-source multimodal reasoning models still produce short responses that leave little room for reflection or re-examining visual evidence (Wang et al., 2025a; Deng et al., 2025; Yang et al., 2025b). UPRSD shares the goal of transferring text reasoning to vision, but targets the interference arising when textual reasoning and multimodal perception supervision are optimized jointly.

The reasoning–perception trade-off. To reuse strong text reasoners, several works convert visual inputs into textual or code intermediates (Yang et al., 2025b; Huang et al., 2026), which can lose fine-grained visual detail in the vision-to-text translation. RAPID (Gou et al., 2026) makes this decoupling explicit: an MLLM produces query-relevant captions and tentative solutions for an external text-only LLM and is trained with reinforcement learning on the downstream answer correctness. UPRSD takes the opposite route and keeps perception and reasoning in a single model that reasons over visual tokens directly, addressing the conflict during training rather than by assigning the two capabilities to different models; its router only decides whether to deliberate, not which model answers. We compare the two approaches empirically in Section 4.2.

Adaptive reasoning depth. To avoid wasteful reasoning on simple queries, FAST (Xiao & Gan, 2025) adds difficulty-aware, length-based rewards to GRPO so that the model chooses between fast and slow thinking for each question, and ARES (Chen et al., 2025c) allocates exploration effort by task difficulty through difficulty-proportional cold-start traces and token-level entropy shaping. The UPRSD router makes a coarser, task-level choice (deliberate or answer directly) that is cheap and predictable but cannot adjust depth within a task; the two mechanisms are complementary.

Gradient conflict in multi-task learning. Conflicting task gradients are a well-known obstacle in multi-task optimization, addressed by projecting out conflicting components (PCGrad; Yu et al., 2020), balancing gradient magnitudes (GradNorm; Chen et al., 2018), or computing Pareto-stationary update directions (MGDA; Sener & Koltun, 2018). These methods modify the update rule for a fixed set of tasks. UPRSD instead changes the supervision interface and the training curriculum, conditioning the two behaviors on distinct prefixes before unifying them, and uses gradient alignment only as a diagnostic; the two families are complementary.

6 Conclusion

We presented Uni-VL, a family of multimodal reasoning models trained with UPRSD, a staged recipe designed to mitigate the modality interference arising when perception and reasoning supervision are naively mixed. Structural Priming separates the two behaviors, Logic Integration unifies them under a shared `<think>` interface, and Reinforced Evolution sharpens frontier reasoning via GRPO with potency-aware data. At both 4B and 8B scales, Uni-VL improves multimodal and text-only reasoning over open-source baselines while preserving general multimodal understanding.

Limitations

Our evidence for modality interference is largely correlational: gradient alignment and stage-wise gains are consistent with the hypothesis but do not establish that reduced conflict alone causes the improvements, and we do not compare against gradient-conflict mitigation methods such as PCGrad or GradNorm under the same data budget. Stage 2 introduces the unified interface and the bootstrapped trajectories together, so their effects are confounded. Routing relies on dataset metadata during training and on a text-only classifier at inference (Appendix A.7), so queries whose need for deliberation depends on the image content may be misrouted. Finally, deliberation via `<think>` adds latency relative to direct-answer models, the model may still produce reasoning steps not grounded in fine-grained visual details, and our experiments are limited to the 8B scale.

AI USE STATEMENT

In this work, we used generative AI tools to polish manuscript wording after an author-written draft, and to assist with data-processing and ablation-experiment code. All AI-assisted text and code were manually verified by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This study strictly adheres to the ICLR Code of Ethics. The datasets utilized in our experiments are publicly available, fully anonymized, and do not involve human subjects, privacy infringement, or harmful discrimination concerns.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide full implementation details, hyperparameter configurations, and training scripts in the anonymized code repository submitted as supplementary material.

REFERENCES

Anthropic. Introducing Claude 4, 2025. URL https://www.anthropic.com/news/claude-4.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer, 2013.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? An early investigation into training R1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-V: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025b. Accessed: 2025-02-02.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Shuang Chen, Yue Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Haoxi Li, Manyuan Zhang, Jiayu Chen, Song Guo, and Nanyun Peng. ARES: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. *arXiv preprint arXiv:2510.08457*, 2025c.

Shuang Chen, Hangyu Guo, Zhaochen Su, Yafu Li, Jiacheng Chen, Yulun Wu, Weijie Wang, ZhiYuan Feng, Xiaoye Qu, and Yu Cheng. Revisual-R1: Advancing multimodal reasoning from optimized cold start to staged reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NTo6f6GENJ.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261.

Joost C. F. de Winter, Dimitra Dodou, and Yke Bauke Eisma. System 2 thinking in OpenAI's o1-preview model: Near-perfect performance on a mathematics exam. *Computers*, 13(11):278, October 2024. ISSN 2073-431X. doi: 10.3390/computers13110278. URL http://dx.doi.org/10.3390/computers13110278.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. *arXiv preprint arXiv:2503.17352*, 2025.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 10764–10799, 2023.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. Reasoning-aligned perception decoupling for scalable multi-modal reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hlLXvyz5iP.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL https://arxiv.org/abs/2310.14566.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025a. doi: 10.1038/s41586-025-09422-z.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-VL technical report. *CoRR*, abs/2505.07062, 2025b. doi: 10.48550/ARXIV.2505.07062.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: Human-aligned video generation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868, 2025.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, Zhaopeng Tu, and Haodong Duan. Human cognitive benchmarks reveal foundational visual gaps in MLLMs. *arXiv preprint arXiv:2502.16435*, 2025.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UZIjskfbfU.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. BIG-Bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with MLLM in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Hang Zhang, Yuming Jiang, Xin Li, Deli Zhao, Fan Wang, Yu Rong, Aixin Sun, and Shijian Lu. MMR1: Advancing the frontiers of multimodal reasoning. https://github.com/LengSicong/MMR1/tree/mmr1_v0, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath: The largest public dataset in AI4Maths with 860k pairs of competition math problems and solutions. Technical report, Numina, July 2024b. URL https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022.

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

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

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

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

Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F. Wong, Zhen Zhao, Xiaoyi Feng, and Maosong Sun. KARL: Knowledge-aware reasoning and reinforcement learning for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *CoRR*, abs/2503.07365, 2025. doi: 10.48550/ARXIV.2503.07365.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. CodeGen: An open large language model for code with multi-turn program synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

OpenAI. Introducing GPT-4.1 in the API, 2025. URL https://openai.com/index/gpt-4-1/.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten, 2023. URL https://arxiv.org/abs/2302.12066.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, Wenye Lin, Gyungin Shin, Qiaochu Yang, Anh Totti Nguyen, David I. Atkinson, Aaditya Baranwal, Alexandru Coca, Mikah Dang, Sebastian Dziadzio, Jakob D. Kunz, Kaiqu Liang, Alexander Lo, Brian Pulfer, Steven Walton, Charig Yang, Kai Han, and Samuel Albanie. ZeroBench: An impossible visual benchmark for contemporary large multimodal models, 2025. URL https://arxiv.org/abs/2502.09696.

William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. Forgotten polygons: Multimodal large language models are shape-blind. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11983–11998, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.620. URL https://aclanthology.org/2025.findings-acl.620/.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. MM-Verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14100–14115, 2025. URL https://aclanthology.org/2025.acl-long.689/.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.

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

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. SoTA with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024b.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024c.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, Chunrui Han, Yuang Peng, Qi Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, and Vishal M. Patel. Open Vision Reasoner: Transferring linguistic cognitive behavior for visual reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023.

Wenyi Xiao and Leilei Gan. Fast-slow thinking GRPO for large vision-language model reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025a. URL https://arxiv.org/abs/2501.09686.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVa-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of CVPR*, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. Vision-G1: Towards general vision language reasoning with multi-domain data curation. *arXiv preprint arXiv:2508.12680*, 2025.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Yi Zong and Xipeng Qiu. GAOKAO-MM: A Chinese human-level benchmark for multimodal models evaluation, 2024. URL https://arxiv.org/abs/2402.15745.

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

A APPENDIX

A.1 IMPLEMENTATION

We initialize UNI-VL from QWEN3-VL-8B-INSTRUCT for the 8B variant and from QWEN3-VL-4B-INSTRUCT for the 4B variant. In the supervised stages (Stages 1 and 2), we train for 5 epochs with a learning rate of 2.0×10^{-5}, a warmup ratio of 0.05, and a cosine learning-rate schedule; to accommodate long-form reasoning, the maximum sequence length is 32,768 tokens.

For Stage 3, we run GRPO with a global batch size of 128, an actor learning rate of 1×10^{-6}, and a KL penalty coefficient of 2×10^{-3}, capping both prompt and response lengths at 8,192 tokens. RL rollouts are sampled with temperature 1.0 and top-p 0.95; the evaluation settings are described in Appendix A.4. All stages use BF16 precision and DeepSpeed ZeRO-2 on 8×A100 80GB GPUs. Table 4 summarizes the hyperparameters.

Table 4: Key hyperparameters of the training stages.

Component	Hyperparameter
Stage 1 & 2 (SFT)	
Learning Rate	2.0×10^{-5}
Gradient Accumulation	8
Number of Epochs	5
LR Scheduler	Cosine
Warmup Ratio	0.05
Max Sequence Length	32,768
Precision	BF16
DeepSpeed	Zero2
Actor (Stage 3)	
Global Batch Size	128
Micro Batch Rollout	4
Max Grad Norm	1.0
Learning Rate	1×10^{-6}
Weight Decay	1×10^{-2}
Entropy Coef Init (β_0)	0.02
Entropy Coef Min (β_{min})	0.0
Entropy Decay Rate (λ)	0.985 (exp)
Entropy Warmup Steps (τ_w)	140
Total Updates	1000
GRPO	
Adv Estimator	grpo
KL Penalty Type	low var kl
KL Coef	2×10^{-3}
τ	0.3
Model Settings	
Max Prompt Length	8192
Max Response Length	8192
Rollout Batch Size	512
Generation Temperature	1.0
Generation Top P	0.95

A.2 EXPERIMENTAL DETAILS FOR THE PILOT STUDY (FIGURE 1)

To compare data-integration strategies in the pilot analysis (Figure 1), we control for confounding variables across all regimes.

Base checkpoint. All pilot experiments are initialized from QWEN3-VL-8B-INSTRUCT.

Data configurations. We compare three regimes built from subsets of our curated data:

- **Text-only ($\mathcal{D}_{text}$):** 321k textual reasoning samples.

- **Multimodal-only ($\mathcal{D}_{mm}$):** 374k multimodal samples.
- **Naive-mixed ($\mathcal{D}_{mix}$):** the union of both sets (695k samples).

Controlled training setup. To keep compute disparities from confounding the results, we align the training budget across regimes: the naive-mixed model is trained for one epoch, and the text-only and multimodal-only models for two epochs each, so that all models are updated on a comparable number of samples (≈700k). All other hyperparameters are shared: global batch size 128, learning rate 2.0×10^{-5} (cosine schedule, 0.05 warmup), and maximum sequence length 32k. This setup rules out differences in training budget as the source of the degradation in the naive-mixed setting, although, as discussed in Section 2, it does not exclude every alternative explanation.

A.3 Protocol for Gradient Conflict Analysis (Figure 4)

To quantify modality interference, we measure the cosine similarity between gradients computed on textual reasoning ($\mathcal{D}_T$) and visual perception ($\mathcal{D}_V$) batches.

Parameter scope. Gradient similarity is computed over the **shared LLM backbone parameters**, namely the weights of all self-attention and feed-forward layers, excluding the vision encoder, projector, and embeddings. This focuses the analysis on the shared substrate where interference is hypothesized to occur.

Measurement protocol. We measure at fixed checkpoints to ensure comparability:

- **Naive-mixed baseline:** gradients are recorded at the end of the first epoch, during active joint optimization.
- **UPRSD (ours):** gradients are recorded at the end of Stage 2 (Logic Integration), where both modalities share the unified interface.

Estimation details. At each checkpoint, we freeze the model weights and sample $N = 50$ pairs of disjoint mini-batches (one from $\mathcal{D}_T$ and one from $\mathcal{D}_V$, each of size 32). For each pair, we compute the flattened gradient vectors g_T and g_V of the selected parameters and their cosine similarity $\cos(g_T, g_V) = \frac{g_T \cdot g_V}{\|g_T\| \|g_V\|}$. The distributions shown in Figure 4 (kernel density estimates) are derived from these 50 independent samples per setting.

A.4 Standardized Evaluation Protocol

All open-source models in Table 1, including Uni-VL and the baselines, are evaluated with a unified pipeline, so that differences reflect model capability rather than evaluation settings.

Unified inference configuration. On reasoning benchmarks, all open-source models use the same sampling parameters, chosen to balance diversity and stability: temperature 0.6, top-p 0.9, top-k 60, and at most 30,000 generated tokens. The high token limit ensures that no model is penalized for generating long chains of thought. For high-variance benchmarks such as AIME24/25, we report pass@1 averaged over multiple independent sampling runs.

Prompting and routing consistency. For AIME24/25 and the other math benchmarks, we use the official evaluation prompts without modification. For baselines that support explicit reasoning triggers (*e.g.*, `<think>`), we apply the same routing logic, so that all models generate reasoning traces before the final answer. This removes prompt engineering as a confounder, and Table 1 reflects reasoning performance under a common protocol.

A.5 Details of the Comparison with RAPID

RAPID numbers in Table 2 are copied from Gou et al. (2026), which reports MathVista (testmini), MathVision (test), MathVerse (vision-only), MMMU (val), WeMath (strict accuracy), DynaMath (worst-case accuracy), and LogicVista evaluated with VLMEvalKit (Duan et al., 2024); Uni-VL numbers are taken from Table 1. Two differences limit this comparison. First, the evaluation

pipelines differ. Second, RAPID builds on QWEN2.5-VL-3B/7B (Bai et al., 2025b), whereas UNI-VL builds on the newer QWEN3-VL-4B/8B (Bai et al., 2025a), so part of the gap may stem from the stronger base model rather than from the training recipe. The comparison should therefore be read as indicative.

A.6 RL TRAINING DYNAMICS

Figure 5 tracks benchmark performance over 1,000 RL steps. Most benchmarks improve over training: GAOKAO-MM and MMSTAR rise steadily with little fluctuation, indicating stable optimization, while ZEROBENCH improves later but consistently, suggesting that complex visual reasoning benefits from extended training.

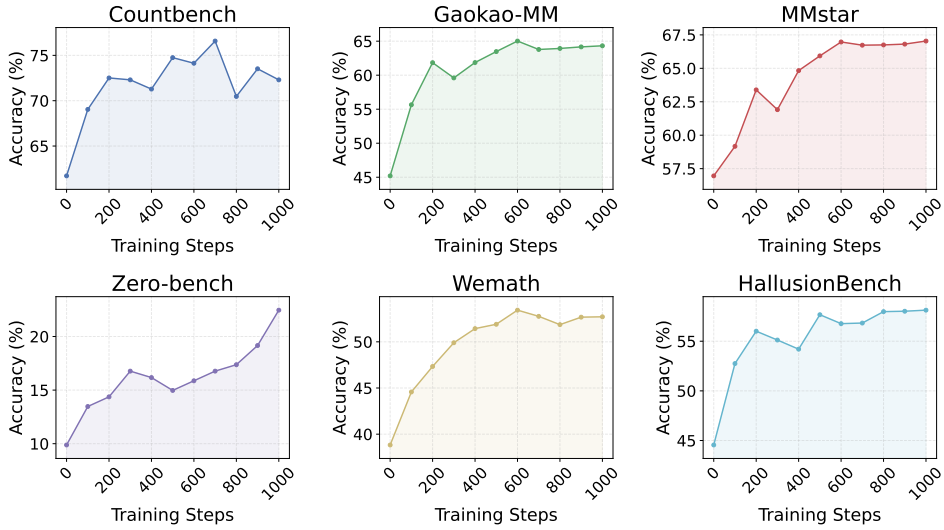


Figure 5: **RL training dynamics.** Accuracy on six multimodal benchmarks during Stage 3 (0–1,000 steps).

A.7 TASK ROUTING SPECIFICATION AND ANALYSIS

To operationalize "think-on-demand" for unseen inputs, we train a lightweight classifier that predicts whether a query requires System-2 reasoning.

Classifier architecture. We use a distilled DEBERTA-V3-XSMALL model (He et al., 2021) (22M parameters) as the routing controller. It takes the textual instruction as input and outputs a binary decision: `Enable-Reasoning` (prepend `<think>`) or `Direct-Response` (no prefix).

Training data. The classifier is trained on a balanced subset of our SFT corpus (100k samples), in which samples from reasoning-intensive domains (math, coding, logic) are labeled positive and perception-heavy samples (OCR, captioning, spotting) negative.

Performance and impact. On a held-out test set, the router achieves 98.4% accuracy, and the routing step adds negligible latency ($\approx$5 ms) compared with MLLM decoding. Misrouting is asymmetric: false negatives (skipping reasoning on math) cause a sharp performance drop (-15% on AIME), whereas false positives (forcing reasoning on simple captioning) increase latency by $3\times$ and introduce a slight rise in object hallucination (+2% on POPE; Li et al., 2023), which justifies dynamic routing.

Ambiguous inputs. Because the router sees only the textual instruction, it cannot use the image to decide whether deliberation is needed. It is therefore least reliable when the instruction underspecifies the task, for example a generic question about a diagram whose difficulty depends on the visual content.

A.8 Variability and Reproducibility Analysis

Scores on small, high-variance benchmarks such as AIME24/25 are averaged over multiple sampling runs (Appendix A.4). To assess the stability of RL training, we ran the Stage 3 training of Uni-VL-8B three times with different random seeds during our pilot phase and observed a variation of approximately $\pm 0.8\%$ on the validation set, well below the gaps between our method and the baselines. The learning curves in Figure 5 are representative of these runs and show consistent improvement without divergence.

A.9 Contamination Audit for Math Benchmarks

Because short-answer math problems are particularly vulnerable to leakage, we conduct a layered contamination audit targeting AIME24/25 and related math benchmarks. The audit covers all training corpora used in SFT and RL, including the bootstrapped trajectories.

Exact-match filtering. We remove any training item whose normalized problem statement exactly matches a test item, using hashing after canonicalization (lowercasing, whitespace normalization, punctuation stripping, and LaTeX normalization).

n-gram overlap filtering. We apply stricter thresholds to short problems: 7-gram overlap filtering for short-answer math and 13-gram overlap filtering for longer items.

Semantic near-duplicate filtering. We compute embedding similarity between training and test problem statements; items above a high similarity threshold are removed or manually reviewed.

Source-level exclusions. We exclude known competition sources and mirrors (*e.g.*, AIME/AMC collections) and apply time-based cutoffs where metadata is available.

Audit outputs. The audit report, which will be released with the data, lists the number of removed or flagged items per corpus, overlap rates by method (exact, n-gram, and semantic), and a sample of flagged examples for qualitative inspection.

A.10 Leak-Free Potency Mining Protocol

To avoid coupling data selection with training, we compute $pass@8_{\text{base}}$ and $pass@1_{\pi_\theta}$ on a held-out **mining pool** disjoint from the RL training prompts. We periodically evaluate $pass@k$ curves on a separate validation split to monitor boundary shrinkage and to ensure that RL improves sampling efficiency without reducing coverage.



Figure 6: **Reasoning trace of UNI-VL-8B on a visual sequence-completion puzzle.** Highlighted spans mark reflection, revisiting the image, self-critique, and hypothesis revision; the final answer (E) is correct.