

Ada-DeVL: Adaptive Decoupling of Visual and Logical Reasoning Pathways for Efficient MLLMs

Haoyuan Fan^{1,3*}, Jie Zhao^{2†}, Yufeng Shi^{1,3*}, Wengang Zhou^{1†}, Houqiang Li¹

¹University of Science and Technology of China,

²Hefei University of Technology, ³Huawei Technologies Co., Ltd

fanhaoyuan@mail.ustc.edu.cn, jzhaoch@hfut.edu.cn,

shiyufeng@mail.ustc.edu.cn, zhwg@ustc.edu.cn, lihq@ustc.edu.cn

Abstract

Scaling test-time computation in Multimodal Large Language Models (MLLMs) enforces heavy visual analysis and long logical reasoning uniformly across queries, leading to severe token waste and overthinking. We propose **Ada-DeVL**, a framework that decouples visual perception depth from logical reasoning intensity, allowing the model to adaptively select one of four visuo-reasoning pathways according to problem complexity. To learn this adaptive behavior, we first apply a Self-Distillation Supervised Fine-Tuning (SFT) stage that generates difficulty-aligned, hallucination-reduced data while avoiding the exposure bias of cross-model distillation. We then use Group Relative Policy Optimization (GRPO) with a composite reward, incorporating diversity and format-adherence constraints to prevent mode collapse and shortcut exploitation. Notably, on the MMMU benchmark, the Qwen2.5-VL-7B based Ada-DeVL achieves **54.0%** accuracy, surpassing Vision-SR1 while saving **39.6%** tokens; the Qwen3-VL-8B based model reaches **64.0%**, outperforming its vanilla counterpart and reducing token consumption by **44.6%**, and achieves competitive accuracy against the strong Qwen3-VL-8B-Thinking baseline with significantly lower token costs (averaging 1560.8 vs. 4889.0 tokens).

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable success in complex vision-language tasks (Team, 2026; Aneja et al., 2026; Team et al., 2026). Inspired by reasoning breakthroughs in textual domains (Guo et al., 2025; Wang et al., 2025b; Wei et al., 2022), an emerging paradigm is shifting toward scaling test-time computation in MLLMs to unlock deeper cognitive capabilities (Huang et al., 2026; Wang

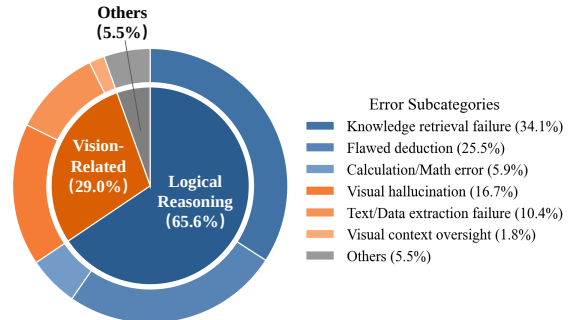


Figure 1: Error analysis of open-source MLLMs on the MMMU and MMMU-Pro benchmarks. The results reveal that model failures predominantly stem from two intertwined dimensions: Visual Perception Errors and Logical Reasoning Errors, which motivates our decoupled visuo-reasoning taxonomy.

et al., 2025a). This inference-time scaling forces MLLMs to engage in multi-step visual verification and dense logical deduction before outputting the final response (Ghosal et al., 2026; Zhang et al., 2025). Our preliminary experiments evaluating open-source models on MMMU (Yue et al., 2024) and MMMU-Pro (Yue et al., 2025) reveal that MLLM failures predominantly stem from two inter-related dimensions: **Visual Perception Errors** and **Logical Reasoning Errors** (Figure 1). This error diagnosis was conducted autonomously by leveraging a closed-source evaluator model to analyze the failure cases (see the prompt in Appendix D).

Methods like Vision-SR1 (Li et al., 2026) and Visionary-R1 (Xia et al., 2025) address these errors by incorporating explicit visual analysis and long-chain reasoning during inference. However, they enforce a uniformly heavy computational paradigm across all queries. This monolithic approach ignores varying query difficulties, leading to: (1) **Prohibitive Computational Overhead** on simple queries, and (2) **Overthinking** (Sui et al., 2025), where prolonged reasoning on straightforward questions induces hallucinations.

* Internship at Huawei Technologies Co., Ltd.

† Corresponding authors.

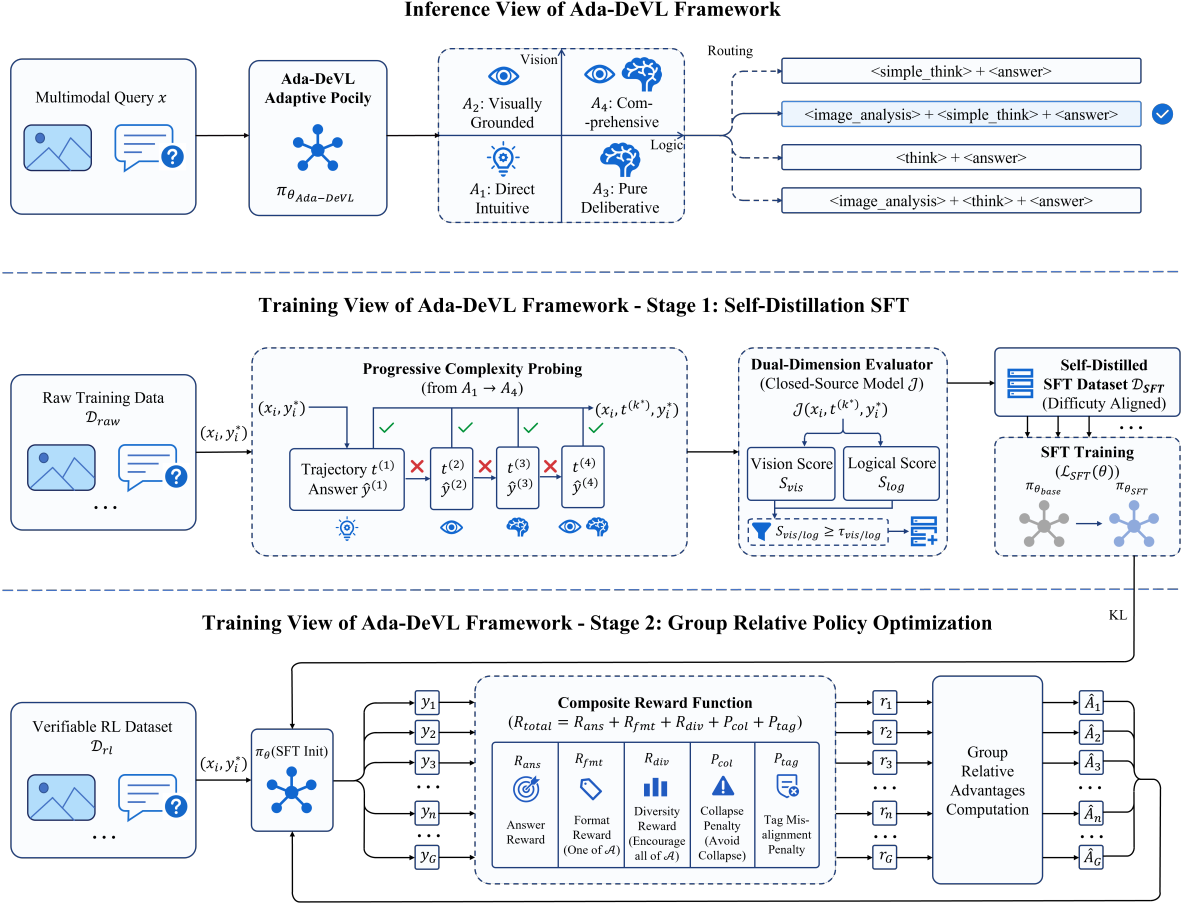


Figure 2: The overall architecture of the Ada-DeVL framework, illustrating the decoupled visuo-reasoning pathways and the two-stage training pipeline.

To this end, we propose **Ada-DeVL**, which conceptualizes the reasoning space as a 2×2 taxonomy along two complementary axes: visual perception depth and logical reasoning intensity. This yields four decoupled visuo-reasoning pathways tailored to varying complexities: Direct Intuitive Reasoning (Low Visual, Low Logic), Visually-Grounded Intuition (High Visual, Low Logic), Pure Deliberative Logic (Low Visual, High Logic), and Comprehensive Visuo-Reasoning Deliberation (High Visual, High Logic).

To instill this adaptive behavior, we design a robust two-stage training paradigm: (1) **Self-Distillation SFT**: We mitigate exposure bias by using the student model itself to generate trajectories, applying a dual-dimension scoring mechanism (visual and logical) to retain only difficulty-aligned, hallucination-reduced data. (2) **GRPO-based Alignment**: To prevent mode collapse and format dishonesty, we optimize decision-making using GRPO (Shao et al., 2024) with a composite reward featuring a novel diversity constraint and a

tag-misalignment penalty.

Our contributions are summarized as follows:

- We reveal that the rigid coupling of visual and logical reasoning causes token inefficiency and overthinking in MLLM test-time scaling, and propose **Ada-DeVL** to address this by allowing models to dynamically choose among four distinct paths within a 2×2 decoupled taxonomy.
- We design a robust two-stage training pipeline (Self-Distillation SFT and GRPO with structure-aware rewards) to effectively instill this adaptive routing behavior.
- Our Ada-DeVL framework consistently improves performance and efficiency across scales. On MMMU, the Qwen2.5-VL-7B based model achieves 54.0% accuracy, surpassing Vision-SR1 while saving 39.6% tokens. Furthermore, the Qwen3-VL-8B based model reaches 64.0% accuracy on MMMU and reduces token consumption by 44.6%

compared to its vanilla counterpart, with accuracy improvements of up to +9.7% on MMMU-Pro 4-Options. Additionally, it achieves competitive accuracy against the remarkably strong Qwen3-VL-8B-Thinking model while drastically reducing average token consumption (1560.8 vs. 4889.0 tokens).

2 Methodology

To mitigate the prohibitive computational overhead and overthinking induced by uniform test-time scaling, we introduce **Ada-DeVL**, a framework that decouples visual perception depth from logical reasoning intensity (Figure 2). The overall pipeline consists of two stages: (1) Self-Distillation Supervised Fine-Tuning (SFT) to acquire the structural syntax and an intrinsic perception of difficulty, and (2) Group Relative Policy Optimization (GRPO) to instill robust, adaptive routing behaviors.

2.1 Adaptive Reasoning Formulation

Instead of a monolithic, fixed reasoning trajectory, we conceptualize the reasoning space as a 2×2 taxonomy based on visual and logical complexity. We formulate this as a discrete action space of four formatting templates: A_1 (Low Visual/Logic, using `<simple_think>`), A_2 (High Visual/Low Logic, using `<image_analysis>` and `<simple_think>`), A_3 (Low Visual/High Logic, using `<think>`), and A_4 (High Visual/High Logic, requiring both `<image_analysis>` and `<think>`).

For a given multimodal query x , the generation of the response y is governed by the selected pathway $A_k \in \mathcal{A}$, explicitly determining whether to maintain a lightweight intuitive reasoning process, or to scale up computational resources for explicit visual parsing(`<image_analysis>`), heavy logical deduction (`<think>`), or both.

2.2 Stage 1: Self-Distillation Supervised Fine-Tuning

Standard SFT typically utilizes reasoning trajectories generated by a stronger, closed-source teacher model as the gold standard. However, this paradigm inevitably introduces severe exposure bias and data misalignment due to the fundamental capacity gap between the teacher and the student. Crucially, the perception of problem difficulty is deeply dependent on the model’s intrinsic capability boundaries. A reasoning trajectory deemed optimal or straightforward by a highly capable teacher

model may severely mismatch the student’s representational capacity. Forcing the student to uniformly mimic such foreign, overly complex reasoning trajectories compels it to learn under a distorted perception of task difficulty. This mismatch not only exacerbates the train-inference discrepancy (i.e., train-test mismatch) during deployment, but also injects optimization noise, causing the model to hallucinate when encountering queries that lie outside its practical cognitive space.

SFT Data Construction. To construct a high-quality, difficulty-aligned SFT dataset autonomously, we introduce a **Progressive Complexity Probing (PCP)** mechanism coupled with a **Dual-Dimension Filtering** strategy.

Given a raw training dataset $\mathcal{D}_{raw} = \{(x_i, y_i^*)\}$ containing only multimodal queries x_i and final ground-truth answers y_i^* , we utilize the base student model $\pi_{\theta_{base}}$ to generate its own reasoning trajectories. To explicitly align the chosen visuo-reasoning pathway with the model’s intrinsic capability, we enforce a sequential probing process from the most lightweight format (A_1) to the most exhaustive format (A_4). For a query x_i , the model sequentially samples trajectories $t^{(k)} \sim \pi_{\theta_{base}}(\cdot | x_i, A_k)$ for $k = 1, 2, 3, 4$. The probing halts at the first format index k^* where the extracted answer $\hat{y}^{(k^*)}$ correctly matches y_i^* . The rationale behind prioritizing A_2 (high visual, low logic) over A_3 (low visual, high logic) is rooted in the intuition that complex logical reasoning is generally more computationally demanding and prone to compounding errors than visual perception. Analogous to human cognition, if a problem can be solved primarily through careful observation, it naturally circumvents the need for deep, intensive logical deduction. Therefore, our probing order is designed to encourage the model to seek the most straightforward visual evidence before resorting to heavy logical computation. This mechanism guarantees that the model learns to allocate the minimum necessary computation required to solve the problem, directly mitigating the "overthinking" phenomenon.

However, correct answers can occasionally arise from flawed logic or hallucinated visual perceptions (i.e., false positives). To prevent noisy data from corrupting the SFT phase, we introduce a closed-source model $\mathcal{J}$ (see Appendix A for evaluator choices) as a dual-dimension evaluator. For the successful trajectory $t^{(k^*)}$, $\mathcal{J}$ assigns a visual

perception score S_{vis} and a logical reasoning score S_{log} : $S_{vis}, S_{log} = \mathcal{J}(x_i, t^{(k*)}, y_i^*) \in [0, 1]$. The trajectory is appended to the distilled dataset $\mathcal{D}_{SFT}$ only if it surpasses predefined strict thresholds: $S_{vis} \geq \tau_{vis}$ and $S_{log} \geq \tau_{log}$. Consequently, $\mathcal{D}_{SFT}$ encapsulates hallucination-reduced, self-generated trajectories that inherently reflect the model’s authentic difficulty boundaries.

SFT Training. With the difficulty-aligned dataset $\mathcal{D}_{SFT}$ prepared, we proceed to explicitly train the model to output the structural formatting tags and the corresponding reasoning content autonomously using an adaptive reasoning prompt (Figure 14).

For a given multimodal query x and its corresponding self-distilled target sequence $t = (t_1, t_2, \dots, t_{|t|})$ (which inherently contains the chosen format tags and the correct answer), the objective is to minimize:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{(x,t) \sim \mathcal{D}_{SFT}} \left[\sum_{j=1}^{|t|} \log \pi_{\theta}(t_j \mid x, t_{<j}) \right]. \quad (1)$$

Through this SFT stage, the model achieves two preliminary goals: (1) it masters the syntactical structure of the four decoupled pathways ($A_1 \sim A_4$); and (2) it acquires a rudimentary, implicit perception of query difficulty based on its own cognitive space.

2.3 Stage 2: Group Relative Policy Optimization (GRPO)

While Self-Distillation SFT establishes formatting syntax, it fundamentally relies on imitation and lacks the exploration necessary to instill robust, adaptive decision-making. Therefore, we utilize GRPO to optimize the policy, formulating a composite reward structure to balance accuracy, format adherence, and efficiency.

Unlike standard PPO, GRPO eliminates the need for an external value model by estimating the baseline from a group of G sampled outputs $\{y_1, y_2, \dots, y_G\}$ for each query x . The objective

is to maximize:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{x, y \sim \pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^G \min(r_i(\theta) \hat{A}_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i) - \beta D_{KL}(\pi_{\theta} \parallel \pi_{ref}) \right], \quad (2)$$

where $r_i(\theta) = \frac{\pi_{\theta}(y_i|x)}{\pi_{old}(y_i|x)}$ is the probability ratio, and the advantage $\hat{A}_i$ is computed by normalizing the total reward $R_{total}^{(i)}$ within the group.

We design a meticulously tailored composite reward function for R_{total} :

$$R_{total} = R_{ans} + R_{fmt} + R_{div} + P_{col} + P_{tag}. \quad (3)$$

Answer and Format Rewards (R_{ans}, R_{fmt}). R_{ans} provides a positive scalar for achieving the correct final answer, while R_{fmt} enforces structural adherence, rewarding strict adherence to one of the four defined A_k templates.

Diversity Reward and Collapse Penalty (R_{div}, P_{col}). MLLMs fine-tuned with reinforcement learning frequently suffer from mode collapse, where the policy collapses to only one or two output formats within $\mathcal{A}$. To counteract this phenomenon, a sliding-window statistical constraint is introduced. Let $\mathcal{W}$ be a window of the last w reward calls. We calculate the empirical frequency f_i for each format $A_i \in \mathcal{A}$ within $\mathcal{W}$. For a generated response utilizing the specific format A_k , the diversity reward and collapse penalty are assigned based on its corresponding frequency f_k :

$$R_{div} = \lambda \mathbb{I}(f_k < \gamma_{low}), \quad (4)$$

$$P_{col} = -\eta \mathbb{I}(f_k > \gamma_{high}), \quad (5)$$

where γ_{low} and γ_{high} are predefined lower and upper frequency thresholds, λ, η are scaling factors, and $\mathbb{I}(\cdot)$ is the indicator function. This explicit constraint continuously forces the policy to explore underrepresented pathways and penalizes the over-exploitation of dominant ones.

Tag-Misalignment Penalty. During preliminary experiments, a shortcut-exploitation phenomenon was observed: the model attempts to bypass format

constraints by generating overwhelmingly verbose reasoning within the `<simple_think>` tag, thereby effectively converting a low-complexity format into a high-complexity one (i.e., turning A_1 into A_3 or A_2 into A_4). To strictly decouple reasoning depth, a non-linear tag-misalignment penalty is applied based on the length L of the `<simple_think>` block. To ensure training stability and prevent gradient explosion, the penalty is capped at a maximum value $C_{\max}$:

$$P_{\text{tag}} = -C_{\max} \cdot \min\left(1, \max(0, \alpha(L - L_{\text{threshold}}))\right), \quad (6)$$

where α is a scaling coefficient. When L exceeds $L_{\text{threshold}}$, the penalty magnitude grows linearly with the excess length, scaled by α , until it saturates at the maximum penalty $C_{\max}$. This mechanism effectively penalizes models that fail to align their computational cost with the chosen complexity tag, discouraging the shortcut of inflating low-complexity responses with verbose reasoning.

3 Experiments

3.1 Experimental Setup

Benchmarks. To comprehensively evaluate the proposed Ada-DeVL framework, we select a diverse set of multimodal benchmarks: MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), HallusionBench (Guan et al., 2024), and VisNumBench (Weng et al., 2025). Detailed descriptions of these benchmarks are provided in Appendix B.

Baselines and Implementation Details. We implement our Ada-DeVL framework on two mainstream open-source foundational models: **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025b) and **Qwen3-VL-8B-Instruct** (Bai et al., 2025a). For the 7B model, we compare our method against both the vanilla baseline and **Vision-SR1** (Li et al., 2026), a pioneering strong baseline that explicitly scales inference-time compute by optimizing models via a self-reward mechanism to generate visual analysis and long-chain logical reasoning trajectories. For the 8B model, we compare directly against the vanilla baseline to demonstrate the scalability of our approach. Additionally, to address the need for a strong 8B baseline, we included a comparison with **Qwen3-VL-8B-Thinking**, which

possesses powerful visual-language reasoning capabilities. For the standard baselines (Standard SFT and Standard GRPO), we ensured a strictly fair comparison: the training parameters, RL data, and SFT source data are identical to those used in Ada-DeVL, with the only difference being that the SFT data for these baselines was rewritten into standard CoT format using a closed-source model.

During the SFT data construction stage, we leverage a dual-dimension automatic evaluation protocol to evaluate both visual perception and logical reasoning quality, filtering out low-scoring or noisy trajectories to ensure data purity. Crucially, to guarantee a rigorous and fair comparison with Vision-SR1, both our self-distilled SFT dataset and the downstream GRPO training pipeline are strictly constructed based on the open-source **Vision-SR1-47K** dataset. For the GRPO alignment stage, hyperparameters including the sliding window size w , the empirical thresholds for diversity and collapse penalties, and the non-linear tag-misalignment coefficient $C_{\max}$ are empirically tuned (see Appendix A for detailed configurations).

3.2 Main Results

The overall performance of our models on the evaluated benchmarks is summarized in Table 1. Our Ada-DeVL framework demonstrates a significant advantage in balancing reasoning accuracy and test-time token efficiency across both 7B and 8B scales. **Efficacy on Qwen2.5-VL-7B.** We first observe that explicitly scaling inference-time reasoning via Vision-SR1 yields a significant performance boost over the vanilla baseline (e.g., Average Acc increases from 43.9% to 47.7%). However, this monolithic reasoning approach induces a massive computational overhead, causing the average token consumption to surge by 74.7% (from 367.2 to 641.6 tokens). In contrast, our Ada-DeVL (SFT+GRPO) achieves highly competitive, and in some cases superior, performance compared to Vision-SR1, while strictly controlling token consumption. For instance, on the MMMU benchmark, Ada-DeVL reaches an accuracy of **54.0%**, outperforming Vision-SR1’s 53.7%. Remarkably, Ada-DeVL achieves this with only **418.4 tokens** on average, saving approximately **39.6%** in token consumption compared to Vision-SR1 (693.1 tokens). Furthermore, on HallusionBench, Ada-DeVL reaches **69.6%** accuracy using merely **172.8** tokens, surpassing Vision-SR1 (67.2%, 432.4 tokens) by mitigating the "overthinking" problem

Table 1: Overall Performance Comparison across Multimodal Benchmarks. We report Accuracy (Acc.) and Average Token Consumption (Tok.) for each model. The best results in each metric are highlighted in **bold**, and the second best are underlined.

Benchmark \ Model	MMMU		MMMU-Pro				MathVerse		MathVision		HallusionBench		VisNumBench		Avg.	
	Acc.	Tok.	10 Options		4 Options		Vision		Avg.		Acc.	Tok.	Acc.	Tok.	Acc.	Tok.
	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.
Qwen2.5-VL-7B Series																
Vanilla Instruct	49.6	397.9	25.3	447.2	35.0	427.4	32.7	415.3	31.0	430.0	46.8	411.0	26.7	619.9	65.4	139.4
Prompt-Only	40.1	263.3	25.4	339.7	40.1	311.6	24.3	175.5	29.9	275.6	28.2	409.0	15.3	527.6	5.5	21.1
Standard SFT	50.5	<u>333.9</u>	20.5	<u>353.8</u>	32.4	<u>354.8</u>	19.6	318.3	24.2	342.3	40.5	350.6	22.4	482.2	67.3	105.2
Standard GRPO	52.5	339.3	21.9	359.0	32.9	364.3	21.9	<u>300.4</u>	25.6	<u>341.2</u>	40.9	<u>358.2</u>	22.2	<u>501.9</u>	67.8	<u>103.2</u>
Vision-SR1	<u>53.7</u>	693.1	42.8	820.6	51.9	797.5	38.9	781.3	44.5	799.8	<u>48.9</u>	659.7	27.5	797.4	67.2	432.4
Ada-DeVL (SFT)	52.6	369.0	39.0	368.9	50.2	387.9	37.3	357.9	42.2	371.6	<u>49.9</u>	459.1	<u>26.8</u>	607.8	<u>68.7</u>	128.4
Ada-DeVL (SFT+GRPO)	54.0	418.4	<u>39.3</u>	377.6	<u>50.6</u>	430.1	<u>38.0</u>	387.2	<u>42.6</u>	398.3	48.2	453.7	26.4	559.2	69.6	172.8
Qwen3-VL-8B Series																
Vanilla Instruct	62.2	2993.2	49.2	3802.6	53.4	3538.4	47.5	3212.1	50.0	3517.7	65.5	2514.6	49.0	6974.4	<u>75.0</u>	367.3
Prompt-Only	55.8	1411.1	42.3	2002.2	51.7	1770.9	41.0	1689.3	45.0	1820.8	46.6	2393.4	23.3	4005.8	73.7	448.5
Standard SFT	57.7	340.4	24.8	368.5	36.6	374.2	25.9	363.0	29.1	368.6	48.8	337.7	31.6	485.4	72.2	99.1
Standard GRPO	63.9	<u>496.6</u>	28.1	<u>546.1</u>	39.4	<u>554.7</u>	36.5	<u>537.6</u>	34.7	<u>546.1</u>	53.8	<u>598.8</u>	36.8	<u>836.6</u>	73.1	<u>99.5</u>
Qwen3-VL-8B-Thinking	68.0	5112.8	56.2	5482.1	<u>63.0</u>	5228.0	53.6	5174.1	57.6	5294.7	73.3	4144.0	54.7	8857.0	76.1	1221.6
Ada-DeVL (SFT)	63.0	1456.7	49.9	1715.0	57.3	1698.5	46.9	1968.9	51.4	1794.1	59.6	1364.6	46.0	3775.3	71.6	338.6
Ada-DeVL (SFT+GRPO)	<u>64.0</u>	1657.2	<u>53.1</u>	1929.4	63.1	1842.4	<u>52.1</u>	1748.3	<u>56.1</u>	1840.0	<u>66.0</u>	1317.6	<u>52.4</u>	3810.2	73.3	197.2

typically triggered by excessively long reasoning chains on visually straightforward tasks.

Scalability on Qwen3-VL-8B. The Qwen3-VL-8B baseline naturally produces extremely long responses (averaging over 3249.4 tokens, and up to 3802.6 on MMMU-Pro). Applying our framework directly curtails this unchecked verbosity. The SFT+GRPO trained model improves the MMMU accuracy from 62.2% to **64.0%** (and up to 63.1% on MMMU-Pro 4-Options, a **+9.7%** improvement), while aggressively halving the token consumption (e.g., from 2993.2 to **1657.2** on MMMU). By adaptively selecting concise pathways (A_1 or A_2) for simpler queries and reserving dense reasoning (A_3 , A_4) only for complex ones, Ada-DeVL effectively prevents the 8B model from falling into hallucinatory over-deliberation, proving that our decoupled taxonomy scales seamlessly to more capable foundation models. Notably, compared to the officially trained Qwen3-VL-8B-Thinking model, which relies on extensive high-quality data and extremely long reasoning chains (averaging 4889.0 tokens), Ada-DeVL achieves competitive accuracy across most benchmarks while reducing token consumption by over 68.1%. This highlights the significant efficiency advantage of our adaptive decoupling strategy, proving that competitive performance can be achieved without blindly extending reasoning lengths.

3.3 Analysis

Analysis of Reward Curves During Training. To verify the effectiveness and stability of our composite reward design during the GRPO stage, we

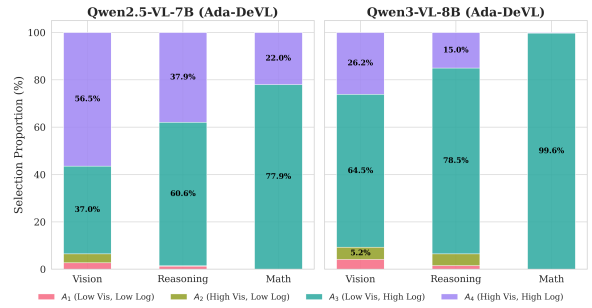


Figure 3: Reasoning pathway selection proportions across Vision, Reasoning, and Math domains for Qwen2.5-VL (left) and Qwen3-VL (right). As tasks become more mathematically intensive, the 8B model heavily converges to A_3 , demonstrating adaptive capability-dependent routing.

tracked the evolution of the Total Reward (R_{total}) alongside its individual sub-components throughout the training process. As illustrated in Figure 4, the training dynamics for Qwen2.5-VL and Qwen3-VL exhibit consistent and desirable behaviors.

Firstly, the primary performance metric—represented by the sum of the accuracy and format rewards ($R_{\text{ans}} + R_{\text{fmt}}$)—shows a steady upward trend, directly driving the overall Total Reward higher. This confirms that the policy successfully learns to generate correct answers while strictly adhering to the decoupled structural formats. For instance, in the Qwen3-VL model, this core reward progressively climbs from an initial average of ~ 0.59 to over ~ 0.63 , demonstrating effective optimization.

Crucially, the auxiliary constraint rewards behave exactly as intended. The Tag Misalignment

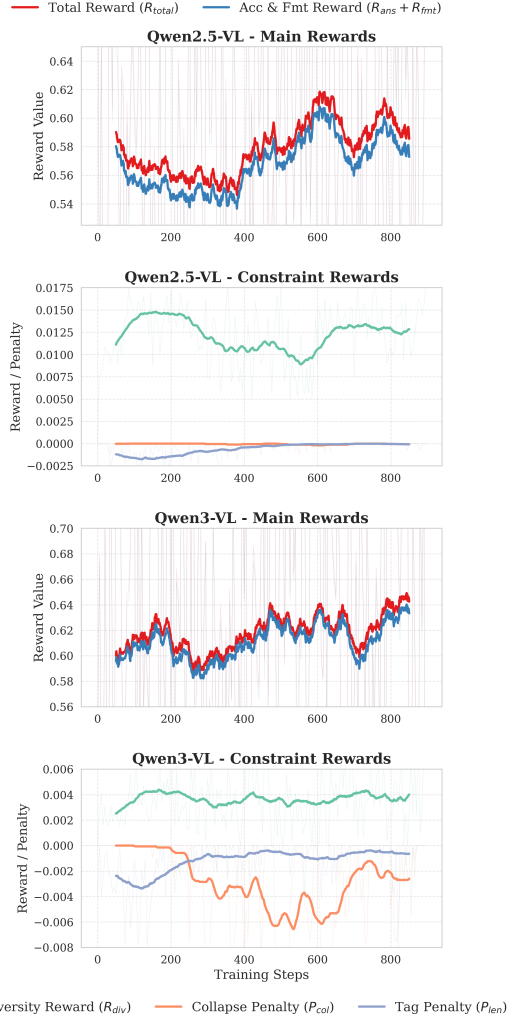


Figure 4: Evolution of the Total Reward and its sub-components during the GRPO alignment stage. The steady increase in accuracy and format rewards, coupled with the decay of the tag misalignment penalty, demonstrates stable and effective optimization.

Penalty (P_{tag}), which prevents the model from exploiting reasoning shortcuts (i.e., padding low-complexity paths with overly dense analysis), starts with a noticeable negative magnitude. However, as training progresses, this penalty rapidly decays toward zero for both models. This indicates that the models quickly learn to align their actual computational expenditure with the chosen tag. Simultaneously, the Collapse Penalty (P_{col}) remains minimal, and the Diversity Reward (R_{div}) maintains a positive presence. This synergy proves that the policy continuously explores the visuo-reasoning taxonomy without degenerating into mode collapse, ensuring that the acquired adaptive routing behavior is both robust and structurally sound.

Table 2: Ablation Study on SFT Strategies. Comparing standard cross-model distillation against our Self-Distillation strategy on the Qwen2.5-VL-3B student model.

Model \ Benchmark	MMMU		MMMU-Pro Vision		Avg.	
	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.
Qwen2.5-VL-3B-Instruct	42.5	520.0	26.6	509.3	34.6	514.7
+SFT ¹	<u>47.9</u>	<u>431.1</u>	<u>28.2</u>	<u>454.4</u>	<u>38.1</u>	<u>442.8</u>
+Self Distillation SFT	49.4	433.2	30.5	409.3	40.0	421.3

Analysis of Pathway Selection Proportions. To examine how Ada-DeVL allocates computational resources, we aggregated pathway selection proportions across three cognitive dimensions: *Vision* (MMMU-Pro-Vision, HallusionBench, VisNum-Bench), *Reasoning* (MMMU, MMMU-Pro options, MathVerse), and *Math* (MathVision, MathVerse). Figure 3 visualizes these distributions.

The results reveal a task-dependent routing continuum. As cognitive demand shifts from *Vision* to *Math*, reliance on explicit visual analysis (A_4) decreases, while pure logical deduction (A_3) surges. Specifically, *Vision* benchmarks induce the highest A_4 usage (56.5% for 7B, 26.2% for 8B) to prevent visual hallucinations. Conversely, *Math* tasks eliminate intuitive pathways (A_1, A_2), allocating compute exclusively to deep thinking (A_3, A_4).

Crucially, routing preferences reflect intrinsic model capabilities. Across all dimensions, Qwen3-VL-8B shifts the computational burden from A_4 to A_3 compared to the 7B model. This is most profound on *Math* tasks: while the 7B model uses A_4 for 26.1% of queries to parse complex charts, the 8B model routes 99.6% to A_3 . This indicates that the 8B model, with its stronger vision encoder, implicitly recognizes it requires less external visual scaffolding, optimally bypassing redundant visual parsing to dedicate compute solely to logical deduction. This demonstrates the core principle of Ada-DeVL: adapting reasoning intensity strictly to the model’s own capability boundaries.

4 Ablation Study

4.1 The Necessity of Self-Distillation

To verify the advantage of Self-Distillation over standard cross-model distillation, we conducted an ablation using Qwen2.5-VL-3B as the student. We compared fine-tuning on difficulty-aligned trajectories generated by a stronger Qwen2.5-VL-7B teacher (Cross-Distillation) against fine-tuning

¹Data generated using Qwen2.5-VL-7B-Instruct.

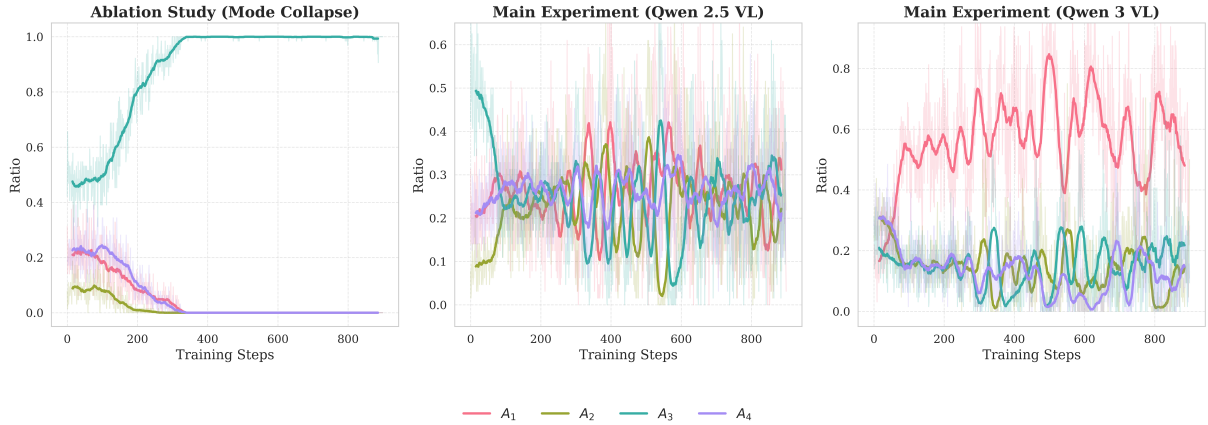


Figure 5: Training step vs. pathway selection probability. **Left:** The ablation experiment without diversity reward and collapse penalty exhibits severe mode collapse. **Middle & Right:** Main experiments on Qwen2.5-VL and Qwen3-VL maintain balanced, adaptive routing distributions.

on the 3B model’s own self-generated data (Self-Distillation). As shown in Table 2, the Self-Distillation variant achieves **49.4%** accuracy on MMMU and **30.5%** on MMMU-Pro Vision, outperforming Cross-Distillation by **+1.5** and **+2.3** points. The average accuracy improves from 38.1% to **40.0%**, while the average token consumption drops from 442.8 to **421.3** tokens. These results confirm that trajectories from a stronger teacher, although objectively higher in quality, exceed the intrinsic capability boundaries of the student model, introducing severe exposure bias and causing the 3B model to hallucinate logic it cannot intrinsically support. In contrast, Self-Distillation anchors the formatting and reasoning depth to the model’s actual cognitive level, delivering a better-calibrated perception of problem difficulty and yielding both higher accuracy and better efficiency.

4.2 Impact of Diversity Reward and Collapse Penalty

To validate the critical role of the diversity reward (R_{div}) and collapse penalty (P_{col}) in maintaining adaptive routing behavior during GRPO alignment, we conducted an ablation experiment by removing these two components. As shown in Figure 5 (left), without explicit diversity constraints, the policy rapidly suffers from severe mode collapse. Specifically, as training progresses, the probability of selecting a single reasoning pathway (A_3) disproportionately surges, eventually dominating the distribution by exceeding an 80% selection rate. Conversely, the selection frequencies for the other three pathways plummet to near zero. This indicates that without structural regularization, the pol-

icy exploits a static, monolithic shortcut, entirely abandoning adaptive behavior.

In contrast, as shown in the middle and right panels of Figure 5, our main experiments on both Qwen2.5-VL and Qwen3-VL models successfully maintain a stable and balanced probability distribution across all four modes throughout the entire training process. By integrating the sliding-window diversity reward and collapse penalty, the Ada-DeVL framework effectively forces the model to continuously explore the full 2×2 taxonomy. This ensures that the model dynamically selects the most appropriate visuo-reasoning pathway tailored to the specific query complexity, successfully preventing the degeneration into a uniform reasoning pattern. Notably, the high proportion of the simplest A_1 pathway in Qwen3-VL (right panel) is not mode collapse, but a capability-driven shift: its stronger intrinsic capabilities allow it to correctly answer more queries using only concise reasoning, thereby naturally maximizing efficiency.

4.3 The Necessity of Dual-Dimension Filtering

To demonstrate the necessity of the dual-dimensional filtering mechanism, we conducted an ablation study using Qwen2.5-VL-7B-Instruct. We compared models trained on unfiltered SFT data versus our filtered SFT data, ensuring that the data volume and all other training settings were completely identical. As shown in Table 3, the model trained on filtered data consistently outperforms the one trained on unfiltered data across the MMMU-Pro benchmark. The dual-dimensional filtering effectively prevents the model from learning "right answers for wrong reasons" (e.g., hallucinated per-

ception or flawed logic), thereby improving overall accuracy.

Table 3: Ablation Study on the Dual-Dimension Filtering strategy. Evaluated on the Qwen2.5-VL-7B-Instruct model across MMMU-Pro subsets.

Setting	10 Options	4 Options	Vision	Average
Unfiltered SFT Data	37.9	49.6	36.7	41.4
Filtered SFT Data (Ours)	39.0	50.2	37.3	42.2

5 Related Work

5.1 Adaptive Test-Time Scaling in MLLMs

Inspired by text-only reasoning breakthroughs (Guo et al., 2025; Muennighoff et al., 2025), scaling test-time computation has expanded to MLLMs (Lin et al., 2025a). Recent frameworks optimize long-horizon trajectories via decoupled visual-language reasoning (Xia et al., 2025; Li et al., 2026), interleaved text-image traces (Gu et al., 2026), or search-based methods (Zhang et al., 2025; Wang et al., 2025a). However, their monolithic execution paradigms apply exhaustive reasoning uniformly, inducing computational redundancy and “overthinking” on simpler tasks—where excessive parsing introduces noise and hallucinations (Sui et al., 2025; Shoby et al., 2026). While adaptive mechanisms like budget-aware routing, dynamic reasoning depth adaptation, or perception decoupling attempt to mitigate this (Gou et al., 2026; Lin et al., 2025b; Zhao et al., 2025; Jia et al., 2026), they lack dynamic allocation of both visual and logical resources. In contrast, **Ada-DeVL** introduces a 2×2 visuo-reasoning taxonomy to adaptively align computational intensity with query difficulty, preventing overthinking while preserving maximal cognitive capacity.

5.2 Self-Distillation and RL-based Post-training

To mitigate *exposure bias* from student-teacher capability gaps, recent paradigms favor on-policy self-distillation, leveraging a model’s own successful generations or privileged feedback to refine reasoning without external teachers (Ranzato et al., 2016; Zhao et al., 2026; Zelikman et al., 2022; Gulcehre et al., 2023; Hübotter et al., 2026). Furthermore, to alleviate the training and inference inefficiencies caused by excessively long rationales during distillation, chunk-wise and skip-thinking

approaches have been proposed to accelerate reasoning (Chen et al., 2025). Building on this, our **Self-Distillation SFT** applies difficulty-aware filtering to ensure trajectories align with the model’s intrinsic visual-logical capacity. For policy alignment, GRPO (Shao et al., 2024) efficiently replaces PPO (Schulman et al., 2017). However, unconstrained RL often suffers from mode collapse and reward hacking (Wu et al., 2025; Gao et al., 2023). We introduce structure-aware reinforcement regularization to preserve routing diversity and strictly enforce our visuo-reasoning taxonomy.

Conclusion

We introduced Ada-DeVL, a framework that optimizes inference-time scaling for Multimodal Large Language Models. By decoupling visual perception from logical reasoning, we proposed a 2×2 taxonomy that dynamically tailors computational expenditure to query complexity. Trained via Self-Distillation SFT and GRPO with structure-aware rewards, Ada-DeVL effectively mitigates token waste and overthinking. Experiments on Qwen2.5-VL and Qwen3-VL demonstrate that our approach achieves highly competitive accuracy while drastically reducing inference costs, paving an efficient path for scalable multimodal reasoning.

Limitations

While Ada-DeVL successfully decouples visual and logical reasoning intensity, our taxonomy is currently quantized into a discrete 2×2 action space. In highly nuanced scenarios, the boundary between “Low” and “High” complexity is continuous rather than categorical. Furthermore, the dual-dimension filtering in our SFT phase relies on a closed-source evaluator, introducing an external dependency during the data curation process. Future work will explore continuous resource allocation paradigms and fully self-supervised reward modeling to eliminate external evaluator dependency.

References

- Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. 2026. *Phi-4-reasoning-vision-15b technical report*. Preprint, arXiv:2603.03975.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang,

- Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Xiaoshu Chen, Sihang Zhou, Ke Liang, Xiaoyu Sun, and Xinwang Liu. 2025. [Skip-thinking: Chunk-wise chain-of-thought distillation enable smaller language models to reason better and faster](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12142–12157, Suzhou, China. Association for Computational Linguistics.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. [Scaling laws for reward model overoptimization](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.
- Soumya Suvra Ghosal, Youngeun Kim, Zhuowei Li, Ritwick Chaudhry, Linghan Xu, Hongjing Zhang, Jakub Zablocki, Yifan Xing, and Qin Zhang. 2026. [Visref: Visual refocusing while thinking improves test-time scaling in multi-modal large reasoning models](#). *Preprint*, arXiv:2603.00207.
- Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. 2026. [Reasoning-aligned perception decoupling for scalable multi-modal reasoning](#). *Preprint*, arXiv:2506.04559.
- Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. 2026. [Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning](#). In *The Fourteenth International Conference on Learning Representations*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. [Reinforced self-training \(rest\) for language modeling](#). *Preprint*, arXiv:2308.08998.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2026. [Vision-r1: Incentivizing reasoning capability in multimodal large language models](#). In *The Fourteenth International Conference on Learning Representations*.
- Jonas Hübötter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Büning, Carlos Guestrin, and Andreas Krause. 2026. [Reinforcement learning via self-distillation](#). *Preprint*, arXiv:2601.20802.
- Hongrui Jia, Chaoya Jiang, Shikun Zhang, and Wei Ye. 2026. Mitigating visual context degradation in large multimodal models: A training-free decoupled agentic framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 9574–9585.
- Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. 2026. [Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization](#). In *The Fourteenth International Conference on Learning Representations*, pages 1–20.
- Yujie Lin, Ante Wang, Moya Chen, Jingyao Liu, Hao Liu, Jinsong Su, and Xinyan Xiao. 2025a. [Investigating inference-time scaling for chain of multi-modal thought: A preliminary study](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15654–15667, Vienna, Austria. Association for Computational Linguistics.
- Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025b. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China. Association for Computational Linguistics.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. [Sequence level training with recurrent neural networks](#). *Preprint*, arXiv:1511.06732.

- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Abin Shoby, Ta Duc Huy, Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Anton van den Hengel, Phi Le Nguyen, Johan W. Verjans, and Vu Minh Hieu Phan. 2026. [Overthinking causes hallucination: Tracing confounder propagation in vision language models](#). *Preprint*, arXiv:2603.07619.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. [Stop overthinking: A survey on efficient reasoning for large language models](#). *Transactions on Machine Learning Research*.
- Qwen Team. 2026. [Qwen3.5-omni technical report](#). *Preprint*, arXiv:2604.15804.
- V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, Wenmeng Yu, Weihang Wang, Wei Li, Shuaiqi Duan, Sheng Yang, Ruiliang Lv, Mingdao Liu, and 78 others. 2026. [Glm-5v-turbo: Toward a native foundation model for multimodal agents](#). *Preprint*, arXiv:2604.26752.
- Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. [Measuring multimodal mathematical reasoning with math-vision dataset](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Xiyao Wang, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025a. [Scaling inference-time search with vision value model for improved visual comprehension](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1184.
- Zhichao Wang, Cheng Wan, and Dong Nie. 2025b. [Review of inference-time scaling strategies: Reasoning, search and rag](#). *Preprint*, arXiv:2510.10787.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. 2025. [Visnumbench: Evaluating number sense of multimodal large language models](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3830–3840.
- Siye Wu, Jian Xie, Yikai Zhang, Aili Chen, Kai Zhang, Yu Su, and Yanghua Xiao. 2025. [ARM: Adaptive reasoning model](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. [Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning](#). *Preprint*, arXiv:2505.14677.
- 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, and 3 others. 2024. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. [MMMU-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [STar: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems*.
- Congzhi Zhang, Jiawei Peng, Zhenglin Wang, Yilong Lai, Haowen Sun, Heng Chang, Fei Ma, and Weijiang Yu. 2025. [VReST: Enhancing reasoning in large vision-language models through tree search and self-reward mechanism](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3922–3941, Vienna, Austria. Association for Computational Linguistics.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. [Mathverse: Does your multi-modal llm truly see thenbsp;diagrams innbsp;visual math problems?](#) page 169–186, Berlin, Heidelberg. Springer-Verlag.
- Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. [Self-distilled reasoner: On-policy self-distillation for large language models](#). *Preprint*, arXiv:2601.18734.
- Zhengyi Zhao, Shubo Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, Binyang Li, Kam-Fai Wong, and Xian Wu. 2025.

T²: An adaptive test-time scaling strategy for contextual question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3731–3756, Suzhou, China. Association for Computational Linguistics.

A Training and Experimental Details

In this section, we provide the detailed hyperparameters and hardware configurations used throughout the two-stage training pipeline. All model training and inference procedures were implemented using the `ms-swift` framework.

A.1 Supervised Fine-Tuning (SFT)

In the first stage, Self-Distillation Supervised Fine-Tuning, we applied full-parameter fine-tuning to both the Qwen2.5-VL-7B and Qwen3-VL-8B models to ensure that the models fully internalize the decoupled 2×2 formatting syntax. For both models, we used a learning rate of 1×10^{-5} , trained for 3 epochs on a dataset comprising approximately 4,000 high-quality, difficulty-aligned trajectories constructed from the original Vision-SR1-47K dataset. The maximum sequence length was capped at 8192 for the 7B model and 16384 for the 8B model to accommodate varying lengths of visual and logical reasoning traces.

In our dual-dimension filtering mechanism for SFT data construction, we employed a closed-source evaluator model. Evaluating reasoning trajectories rigorously requires a model with exceptionally strong perceptual reasoning and reflective judgment capabilities. Most open-source models currently lack the necessary capacity for such rigorous evaluation. Therefore, we opted to rely on the closed-source model to ensure data quality.

A.2 Group Relative Policy Optimization (GRPO)

The second stage, GRPO-based Alignment, demands significantly higher computational resources due to the necessity of generating multiple trajectories (group size) per query and calculating reference model KL divergences. Consequently, this stage was executed on a cluster of 8 Ascend 910b NPUs (64GB each). We utilized the Vision-SR1-47K dataset for this stage. The policy models were trained with a learning rate of 1×10^{-6} . We generated a group size of $G = 8$ per query. The maximum model length was set to 6144 for Qwen2.5-VL-7B (trained for 300 steps, 99.4 NPU hours) and 16384 for Qwen3-VL-8B (trained for

900 steps, 386.4 NPU hours). To enforce format adherence and prevent mode collapse, the composite reward parameters were strictly controlled: the Diversity Bonus λ and Collapse Penalty η were both set to 0.02, with the minimum and maximum format selection ratio thresholds γ_{low} and γ_{high} constrained to 0.3 and 0.7 respectively within a sliding window of size $w = 64$. For the non-linear tag-misalignment penalty, we set the maximum penalty C_{max} to 0.05, the length threshold $L_{\text{threshold}}$ to 350 characters, and the scaling coefficient α to $\frac{1}{175}$. A repetition penalty of 1.05 was applied during the rollouts to prevent generative loops.

B Benchmark Details

To comprehensively evaluate the proposed Ada-DeVL framework, we select a diverse set of multi-modal benchmarks that demand varying degrees of visual perception and logical reasoning:

- **Comprehensive MLLM Benchmarks:** MMMU (Yue et al., 2024) and MMMU-Pro (Yue et al., 2025) (evaluating multi-discipline and vision-centric reasoning with 10-option, 4-option and vision settings).
- **Mathematical & Geometric Reasoning:** MathVerse (Zhang et al., 2024) and MathVision (Wang et al., 2024), which heavily rely on rigorous logical deduction interrelated with visual charts and diagrams.
- **Visual Hallucination & Perception:** HalluBench (Guan et al., 2024) and VisNumBench (Weng et al., 2025), which focus on fine-grained visual perception and numerical text extraction, explicitly testing the model’s resistance to visual errors.

C Case Study

In this section, we present qualitative examples demonstrating the adaptive routing and efficiency of our Ada-DeVL framework.

C.1 Case 1: Mitigating Overthinking (Efficiency)

As illustrated in Figure 6, both models answer correctly, but Vision-SR1 suffers from overthinking, generating excessive tokens for a straightforward visual question. Ada-DeVL dynamically selects the A_1 pathway to provide a highly efficient response.

C.2 Case 2: The Four Visuo-Reasoning Pathways of Ada-DeVL

We demonstrate how Ada-DeVL dynamically routes different queries to the four decoupled pathways based on their distinct visual and logical complexities (see Figure 7, Figure 8, and Figure 9).

D Prompts Used in Inference and Training

In this section, we detail the exact prompt templates utilized across the different stages of the Ada-DeVL framework. These prompts are crucial for defining the structural syntax of the decoupled visuo-reasoning pathways and guiding the model’s adaptive routing behavior.

Specifically, the prompts are categorized as follows:

- **Pathway-Specific Generation Prompts:** Used during the Progressive Complexity Probing (PCP) phase of the Self-Distillation SFT data construction. These prompts force the base model to generate reasoning trajectories strictly adhering to one of the four predefined formats (Figure 10, Figure 11, Figure 12, Figure 13).
- **Ada-DeVL Core Prompt:** The unified system prompt employed during the GRPO alignment stage and final inference (Figure 14). It explicitly defines the 2×2 taxonomy and instructs the model to autonomously select the most appropriate reasoning and visual analysis depth based on the query’s complexity.
- **SFT Data Filter Prompts:** The evaluation templates used by the closed-source evaluator model to score the self-generated trajectories. These prompts independently assess the quality of concise reasoning (Figure 15), detailed logical deduction (Figure 16), and comprehensive visual parsing (Figure 17), ensuring only high-quality, hallucination-reduced data is retained for SFT.
- **Error Analysis Prompt:** The template used by the closed-source model to autonomously categorize the failure cases of the vanilla baselines into Visual Perception Errors, Logical Reasoning Errors, or Other Errors. (Figure 18).

Case 1: Simple Visual Question

User Query:



What is the common term for the yellow area surrounding the site of an infection?

A. I don't know and I don't want to guess B. Corona C. Border D. Halo E. Toxin zone

Vision-SR1 (Correct but Verbose):

<description>

The image shows a leaf from a plant with several large, irregularly shaped lesions. The lesions appear as patches of yellowish or light green tissue surrounding darker, necrotic areas. The yellow region forms a distinct ring-like margin around the brownish central lesion, creating a halo-like appearance around the site of infection.

</description><think>

To determine the common term for the yellow area surrounding the site of an infection on a plant leaf, let's examine the image closely. The leaf shows distinct lesions with a yellow halo around them. This halo is a ring-like area of yellowing or discoloration surrounding the affected region. Now, let's consider the options provided:

A. I don't know and I don't want to guess - This option is not useful as it doesn't provide any insight into what the term might be.

B. Corona - A corona typically refers to a halo of light or a crown-like structure, which doesn't fit the context of plant pathology.

C. Border - This could potentially refer to the boundary or edge of the lesion, but it doesn't specifically describe the yellow area.

D. Halo - This term is commonly used in plant pathology to describe the ring of yellow or healthy tissue surrounding a diseased area.

E. Toxin zone - This refers to an area where toxins are present, which doesn't specifically describe the yellowing around the lesion.

Given the visual evidence from the image and the context of plant pathology, the yellow area surrounding the lesion is best described as a "halo." This matches option D, which is commonly used in scientific and everyday language related to plant diseases.

</think> D

Ada-DeVL-7B (Correct and Concise, A_1 Pathway):

<simple_think>

The yellow area around the infected spot is called a halo, which is a common term used in plant pathology to describe this phenomenon.

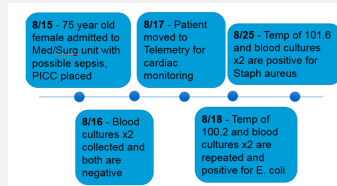
</simple_think>

<answer>D</answer>

Figure 6: Comparison between Vision-SR1 and Ada-DeVL on a visually straightforward question. Ada-DeVL selects the A_1 pathway, avoiding the overthinking and excessive token consumption exhibited by Vision-SR1.

A₁: Low Visual, Low Logic

Query:



What BSI organisms are reported?

A. E. coli B. Staph aureus C. E. coli and Staph aureus

Ada-DeVL:

<simple_think>

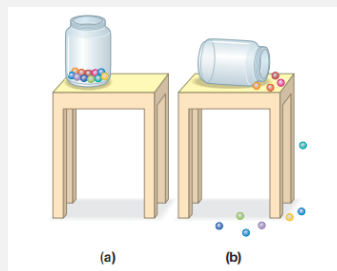
The blood cultures on 8/25 were positive for Staph aureus, and the cultures on 8/18 were positive for E. coli. Both organisms are reported.

</simple_think>

<answer>C</answer>

A₂: High Visual, Low Logic

Query:



Look at the two pictures below. Which picture appears to have more order?

A. a B. b C. same of the two

Ada-DeVL:

<image_analysis>

Visual objects: Two tables, two jars, colorful candies. Text content: "(a)", "(b)". Spatial layout: Picture (a) shows a jar upright on a table with candies contained inside; picture (b) shows a jar lying on its side on a table with candies scattered around it. Scene & context: The images depict two scenarios comparing orderliness of candy storage. Actions/relationships: In (a), candies are contained within the jar; in (b), candies are spilled out and scattered. Diagrams/charts/symbols: No diagrams or charts present. Domain knowledge: Order refers to arrangement and containment.

</image_analysis>

<simple_think>

The candies in picture (a) are contained within the jar, indicating order, while in picture (b), the candies are scattered, indicating disorder.

</simple_think>

<answer>A</answer>

Figure 7: Examples of Ada-DeVL dynamically routing queries to the A₁ (Low Visual, Low Logic) and A₂ (High Visual, Low Logic) pathways based on their inherent simplicity.

*A*₃: Low Visual, High Logic

Query:

<i>Isotope</i>	<i>Mass (amu)</i>	<i>Percent Abundance</i>
Mg-24	24.0	79.0%
Mg-25	25.0	10.0%
Mg-26	26.0	11.0%

There are three common isotopes of naturally occurring magnesium as indicated in the table below. <image 1> Using the information above, calculate the average atomic mass of magnesium. Write your answer number only, in the unit of amu.

Ada-DeVL:

<think>

To calculate the average atomic mass of magnesium, we need to use the formula for the weighted average:

[Average Atomic Mass = (Mass of Isotope 1 × Abundance of Isotope 1) + (Mass of Isotope 2 × Abundance of Isotope 2) + (Mass of Isotope 3 × Abundance of Isotope 3)]

Given data:

- Mg-24: 24.0 amu, 79.0% abundance
- Mg-25: 25.0 amu, 10.0% abundance
- Mg-26: 26.0 amu, 11.0% abundance

Step 1: Convert percentages to decimals for calculation.

- 79.0% = 0.79
- 10.0% = 0.10
- 11.0% = 0.11

Step 2: Multiply each isotope's mass by its abundance.

- (24.0 × 0.79 = 18.96)
- (25.0 × 0.10 = 2.50)
- (26.0 × 0.11 = 2.86)

Step 3: Sum these products to get the average atomic mass.

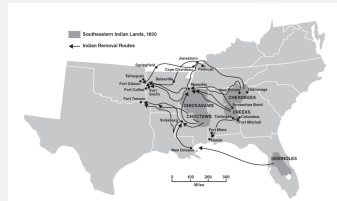
[18.96 + 2.50 + 2.86 = 24.32]

Therefore, the average atomic mass of magnesium is 24.32 amu.

</think>

<answer>24.32</answer>

Figure 8: Example of Ada-DeVL routing a computationally intensive query to the *A*₃ (Low Visual, High Logic) pathway for detailed logical deduction without requiring exhaustive visual parsing.

Query:

Which event in Native American history is LEAST similar to the migration depicted in the map above? A. The Trail of Tears B. King Philip's War C. Creek War of 1836 D. The Battle of Little Big Horn

Ada-DeVL:

<image_analysis>

Visual objects: Map of the southeastern United States, labeled Indian Removal Routes, shaded areas indicating Southeastern Indian Lands in 1830, labeled Native American tribes (Cherokee, Chickasaw, Choctaw, Creek, Seminole), and marked locations such as forts, cities, and rivers. Text content: "Southeastern Indian Lands, 1830", "Indian Removal Routes", tribe names, place names, scale bar "0 100 200 300 Miles". Spatial layout: The map centers on the southeastern U.S., with routes originating from tribal territories and converging toward westward destinations like Oklahoma and Arkansas. Scene & context: Depicts forced relocation of Native American tribes during the Indian Removal era under U.S. government policy. Actions/relationships: Shows movement of tribes away from their ancestral lands to designated territories west of the Mississippi River. Diagrams/charts/symbols: Arrows indicate direction of migration; shaded regions denote original lands; labels identify tribes and key locations. Domain knowledge: Relates to U.S. history, specifically the Trail of Tears and broader Indian Removal policies.

</image_analysis><think>

The map illustrates the forced migration of Native American tribes (Cherokee, Chickasaw, Choctaw, Creek, Seminole) from their ancestral lands in the southeastern U.S. to territories west of the Mississippi River, primarily during the 1830s. This is known historically as the Indian Removal Era, culminating in events like the Trail of Tears.

Option A (Trail of Tears): Directly matches the map's depiction — it was the forced removal of the Cherokee Nation, one of the tribes shown.

Option B (King Philip's War): Occurred in New England (Massachusetts) in 1675–1676; involved Native tribes resisting English colonization. It predates the Indian Removal Era by over 150 years and involved warfare, not forced relocation via government policy.

Option C (Creek War of 1836): Involved conflict between U.S. forces and the Creek Nation over land rights and removal; occurred near the time period depicted on the map and relates to removal pressures.

Option D (Battle of Little Big Horn): Took place in Montana Territory in 1876; involved Lakota, Northern Cheyenne, and Arapaho tribes against U.S. Army forces. While also involving Native resistance, it occurred much later than the removal era shown on the map.

Comparing all options, King Philip's War is least similar because it was a colonial-era conflict focused on territorial defense and sovereignty, not a government-mandated forced migration like those depicted on the map.

</think>

<answer>B</answer>

Figure 9: Example of Ada-DeVL routing a highly complex query to the A₄ (High Visual, High Logic) pathway, which necessitates both comprehensive visual analysis and detailed logical deduction.

A_1 (Low Visual, Low Logic) Format Prompt

You are an AI assistant that answers multiple-choice questions based on images. For every question, you MUST output your response in EXACTLY this format and NOTHING ELSE:

<simple_think>

Your concise reasoning here (1–5 sentences only)

</simple_think>

<answer>X</answer>

Rules:

- Inside `<simple\_think>`: Provide ONLY a brief 1–5 sentence analysis. Do NOT include detailed step-by-step reasoning! The analysis must be within five sentences.
- Inside `<answer>`: If the question is a multiple-choice question (with lettered options A, B, C, D, etc.), output ONLY a single uppercase letter corresponding to the correct choice. If the question is NOT a multiple-choice question, output ONLY a single number (integer or decimal). No spaces, punctuation, or additional characters permitted.
- NEVER add text before `<simple\_think>` or after `</answer>`.
- STRICTLY follow this XML-like tag structure. Any deviation will cause evaluation failure.

Figure 10: Prompt for A_1 pathway generation, used in SFT Data Construction.

A_2 (High Visual, Low Logic) Format Prompt

You are an AI assistant that answers multiple-choice questions based on images. For every question, you MUST output your response in EXACTLY this format and NOTHING ELSE:

<image_analysis>

Visual objects: [list key objects]. Text content: [quote visible text]. Spatial layout: [describe positions]. Scene & context: [setting/purpose/...]. Actions/relationships: [interactions/...]. Diagrams/charts/symbols: [describe if present]. Domain knowledge: [relevant expertise]

</image_analysis>

<simple_think>

Your concise reasoning here (1–5 sentences only)

</simple_think>

<answer>X</answer>

Rules:

- Inside `<image\_analysis>`: Systematically describe ALL seven aspects: (1) Visual objects, (2) Text content, (3) Spatial layout, (4) Scene & context, (5) Actions or relationships, (6) Diagrams/charts/symbolic elements, (7) Domain-relevant knowledge. Be factual and exhaustive.
- Inside `<simple\_think>`: Provide ONLY a brief 1–5 sentence analysis. Do NOT include detailed step-by-step reasoning! The analysis must be within five sentences.
- Inside `<answer>`: If the question is a multiple-choice question (with lettered options A, B, C, D, etc.), output ONLY a single uppercase letter corresponding to the correct choice. If the question is NOT a multiple-choice question, output ONLY a single number (integer or decimal). No spaces, punctuation, or additional characters permitted.
- NEVER omit, reorder, or modify these tags. Format must be exact: image_analysis → simple_think → answer. Any extra text will invalidate the response.

Figure 11: Prompt for A_2 pathway generation, used in SFT Data Construction.

A_3 (Low Visual, High Logic) Format Prompt

You are an AI assistant that answers multiple-choice questions based on images. For every question, you MUST output your response in EXACTLY this format and NOTHING ELSE:

<think>[Structured detailed CoT formal reasoning]</think><answer>X</answer>

Rules:

- Inside `<think>`: Provide a DETAILED, multi-step chain-of-thought (CoT) analysis with explicit reasoning steps.
- Inside `<answer>`: If the question is a multiple-choice question (with lettered options A, B, C, D, etc.), output ONLY a single uppercase letter corresponding to the correct choice. If the question is NOT a multiple-choice question, output ONLY a single number (integer or decimal). No spaces, punctuation, or additional characters permitted.
- NEVER add text before `<think>` or after `</answer>`.
- STRICTLY follow this XML-like tag structure. Any deviation will cause evaluation failure.

Figure 12: Prompt for A_3 pathway generation, used in SFT Data Construction.

A_4 (High Visual, High Logic) Format Prompt

You are an AI assistant that answers multiple-choice questions based on images. For every question, you MUST output your response in EXACTLY this format and NOTHING ELSE:

<image_analysis>

Visual objects: [list key objects]. Text content: [quote visible text]. Spatial layout: [describe positions]. Scene & context: [setting/purpose/...]. Actions/relationships: [interactions/...]. Diagrams/charts/symbols: [describe if present]. Domain knowledge: [relevant expertise]

</image_analysis>

<think>

[Structured detailed CoT formal reasoning]</think>

<answer>X</answer>

Rules:

- Inside `<image\_analysis>`: Systematically describe ALL seven aspects: (1) Visual objects, (2) Text content, (3) Spatial layout, (4) Scene & context, (5) Actions or relationships, (6) Diagrams/charts/symbolic elements, (7) Domain-relevant knowledge. Be factual and exhaustive.
- Inside `<think>`: Provide a DETAILED, multi-step chain-of-thought (CoT) analysis with explicit reasoning steps.
- Inside `<answer>`: If the question is a multiple-choice question (with lettered options A, B, C, D, etc.), output ONLY a single uppercase letter corresponding to the correct choice. If the question is NOT a multiple-choice question, output ONLY a single number (integer or decimal). No spaces, punctuation, or additional characters permitted.
- NEVER omit, reorder, or modify these tags. Format must be exact: image_analysis → think → answer. Any extra text will invalidate the response.

Figure 13: Prompt for A_4 pathway generation, used in SFT Data Construction.

Ada-DeVL Core Prompt

You are a multimodal AI assistant. For every user query, you must autonomously select and strictly adhere to exactly one of the following four response formats based on the difficulty of the problem and the image. Output only the tags and content—no extra text.

Response Formats:

1. Concise reasoning without visual analysis

`<simple_think>

Your concise reasoning here. No step-by-step details.

</simple_think>`

`<answer>Final answer</answer>`

2. Concise reasoning with visual analysis

`<image_analysis>

Cover all 7 aspects factually: (1) Visual objects, (2) Text content, (3) Spatial layout, (4) Scene & context, (5) Actions/relationships, (6) Diagrams/charts/symbols, (7) Domain knowledge.

</image_analysis>`

`<simple_think>

Your concise reasoning here. No step-by-step details.

</simple_think>`

`<answer>Final answer</answer>`

3. Detailed reasoning without visual analysis

`<think>

Structured detailed CoT formal reasoning.

</think>`

`<answer>Final answer</answer>`

4. Detailed reasoning with visual analysis

`<image_analysis>

Cover all 7 aspects factually: (1) Visual objects, (2) Text content, (3) Spatial layout, (4) Scene & context, (5) Actions/relationships, (6) Diagrams/charts/symbols, (7) Domain knowledge.

</image_analysis>`

`<think>

Structured detailed CoT formal reasoning.

</think>`

`<answer>Final answer</answer>`

Critical Rules:

- Inside `<image\_analysis>`: Systematically describe ALL seven aspects: (1) Visual objects, (2) Text content, (3) Spatial layout, (4) Scene & context, (5) Actions or relationships, (6) Diagrams/charts/symbolic elements, (7) Domain-relevant knowledge. Be factual and exhaustive.
- Inside `<simple\_think>`: Provide a brief analysis. Do NOT include detailed step-by-step reasoning! This analysis must be completed within just a few sentences.
- Inside `<think>`: Provide a DETAILED, multi-step chain-of-thought (CoT) analysis with explicit reasoning steps.
- Inside `<answer>`: Output a single uppercase letter (A, B, C, D, E...) or the answer text. Do not allow any extra spaces, punctuation marks or any other characters.
- Close all tags properly. Separate tags with `\\n`.
- Never output explanatory text outside tags.

Figure 14: Ada-DeVL core prompt used for SFT training, RL alignment, and final inference. This prompt enforces the dynamic selection among the four visuo-reasoning pathways based on query complexity.

SFT Data Filter Prompt(Simple Think)

You are an expert AI evaluator.

Your task is to evaluate the concise reasoning provided in the ``<simple_think>`` section of an AI's response.

Problem: problem

AI's simplified reasoning:

{simple_think_content}

Evaluate the reasoning based ONLY on the following content criteria (do not evaluate formatting or XML tags, as format standardization is already handled):

1. Conciseness: Is the analysis brief and direct (ideally 1-5 sentences) without overly detailed, step-by-step CoT reasoning?
2. Reasonableness: Is the logic sound, direct, and sufficient to reach the correct conclusion?
3. Factual/Mathematical Correctness: Are the core facts and deductions correct?

Provide a detailed evaluation, a reason for rejection (if any), and a final quality score from 0 to 100. (Deduct points if the reasoning is too verbose/detailed or if it contains logical/factual errors).

Format your output exactly as a JSON object:

```
{
  "evaluation": "detailed text",
  "reject_reason": "none or specific reason",
  "score": 85
}
```

Figure 15: Prompt for Evaluating Simple Thinking Quality in SFT Data Filtering.

SFT Data Filter Prompt(Think)

You are an expert AI evaluator.

Your task is to evaluate the logical reasoning provided in the ``<think>`` section of an AI's response.

Problem: problem

AI's reasoning:

{think_content}

Evaluate the reasoning based ONLY on the following content criteria (do not evaluate formatting or XML tags, as format standardization is already handled):

1. Detailed CoT Analysis: Does it provide a detailed, multi-step chain-of-thought (CoT) analysis with explicit reasoning steps?
2. Logical Coherence: Do the steps logically follow each other without missing logical leaps?
3. Factual/Mathematical Correctness: Are the facts, calculations, and logical deductions correct?

Provide a detailed evaluation, a reason for rejection (if any), and a final quality score from 0 to 100. (Deduct points if the reasoning lacks multi-step depth or contains logical flaws).

Format your output exactly as a JSON object:

```
{
  "evaluation": "detailed text",
  "reject_reason": "none or specific reason",
  "score": 85
}
```

Figure 16: Prompt for Evaluating Chain-of-Thought Quality in SFT Data Filtering.

SFT Data Filter Prompt(Image Analysis)

You are an expert AI evaluator with vision capabilities.

Your task is to evaluate the ``<image_analysis>`` section of an AI's response.

Problem:

{problem}

Image analysis provided by AI:

{image_analysis_content}

Evaluate the analysis based ONLY on the following content criteria (do not evaluate formatting or XML tags, as format standardization is already handled):

1. Aspect Completeness: Does it systematically and exhaustively describe ALL seven required aspects: (1) Visual objects, (2) Text content, (3) Spatial layout, (4) Scene & context, (5) Actions or relationships, (6) Diagrams/charts/symbolic elements, and (7) Domain-relevant knowledge?
2. Accuracy & Factuality: Does the description accurately reflect the provided image without hallucinating non-existent content?
3. Relevance: Are the extracted visual elements and domain knowledge helpful and relevant for solving the problem?

Provide a detailed evaluation, a reason for rejection (if any), and a final quality score from 0 to 100. (Deduct points if any of the 7 aspects are missing, incomplete, or hallucinated).

Format your output exactly as a JSON object:

```
{
  "evaluation": "detailed text",
  "reject_reason": "none or specific reason",
  "score": 85
}
```

Figure 17: Prompt for Evaluating Image Analysis Quality in SFT Data Filtering.

Error Analysis Prompt for Failure Cases

You are an expert evaluator analyzing why a multimodal model answered a question incorrectly. Here is the original question, the incorrect model output, and the ground truth answer.

Original Question:

{question}

Model Output:

{response}

Ground Truth:

{ground_truth}

Please carefully analyze the root cause of the error. Classify the error into exactly one of the following 3 main categories, and select the most appropriate strictly defined subcategory.

1. Vision-related errors

Subcategories:

- "Object/Attribute misidentification": Incorrectly recognized objects, colors, shapes, or spatial relationships in the image.
- "Text/Data extraction failure": Misread or failed to extract text, numbers, labels, or data points from charts/graphs.
- "Visual context oversight": Missed crucial visual details or failed to synthesize the overall visual scene.
- "Visual hallucination": Acted upon visual elements that do not actually exist in the image.

2. Logical reasoning errors

Subcategories:

- "Calculation/Mathematical error": Mistakes in arithmetic, equations, or quantitative computations.
- "Flawed deduction": Incorrect logical sequence, false implications, or wrong application of domain principles.
- "Textual condition oversight": Ignored, forgot, or misinterpreted constraints explicitly stated in the text prompt.
- "Knowledge retrieval failure": Misapplied or lacked external common sense or domain-specific knowledge required for the task.

3. Other errors

Subcategories:

- "Formatting/Output issue": The reasoning was entirely correct, but the final answer was presented in an unacceptable format.
- "Refusal": The model refused to provide an answer.

Output a JSON object with the following fields:

- "reasoning": A brief explanation of the error analysis.
- "main_category": Must be exactly one of "1. Vision-related errors", "2. Logical reasoning errors", or "3. Other errors".
- "sub_category": Must be exactly one of the strict subcategory strings defined above.

Figure 18: Prompt used by the closed-source model to autonomously categorize failure cases into Visual Perception Errors, Logical Reasoning Errors, or Other Errors.