IGOA: ONLINE ALIGNMENT FOR UNIFIED-CANVAS REASONING VIA INFORMATION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have excelled at processing multiple images as sequential inputs. However, their performance drops when textual instructions and one or more visual instances are jointly rasterized within one canvas, which we define as the Picture-in-Picture (PiP) Gap. We introduce a PiP evaluation framework with a flagship challenge and a paired diagnostic: PiP-All places the query and indexed visual instances in one image, whereas PiP-Visual preserves image order and source pixels at the benchmark-rendering stage, omits the query from the canvas, and supplies it only as language tokens. On three representative models, 96.1%–99.2% of the total gap remains under PiP-Visual, showing that query rasterization alone cannot explain the degradation. We measure this gap end-to-end through each model's released visual processor and inference pipeline. PiP further groups questions by their dominant requirement: subregion referencing or cross-region compositional reasoning. Treating this as a learnable gap, we propose Information-Geometric Online Alignment (IGOA) for multi-context RLVR. IGOA coordinates Original-Interleaved and Unified-Canvas contexts using a scalable diagonal active-Fisher proxy and coordinate-wise log-volume expansion over accumulated Fisher memory. These scores drive a mirror-descent update that prioritizes non-redundant context contributions. Experiments across 13 models show a consistent PiP bottleneck, while IGOA improves PiP performance across four backbone groups and also improves performance on five standard multimodal benchmarks. *Code is available at https://anonymous.4open.science/r/IGOA_PiP-EA76.*

1 INTRODUCTION

MLLMs are increasingly evaluated on reasoning over multiple images, which are commonly supplied as separate sequence elements with explicit boundaries (Meng et al., 2025b; Wang et al., 2025a; Xu et al., 2026). In diagrams, documents, and densely interleaved scenes (Du et al., 2026; Luo et al., 2025; Jiang et al., 2025; Feng et al., 2025), however, text and multiple visual instances can share one canvas. Such inputs require reference resolution and evidence integration within a common visual coordinate system. Whether models preserve task behavior when otherwise equivalent multimodal content is reorganized in this way remains underexplored.

This question is distinct from evaluating a harder task or a new knowledge domain. A paired reorganization test can retain the query, answer, source images, instance order, and decoding protocol while changing whether instance boundaries are supplied by the multimodal sequence or must be recovered within a shared canvas. The resulting performance difference therefore provides a controlled axis for studying presentation sensitivity in multimodal reasoning.

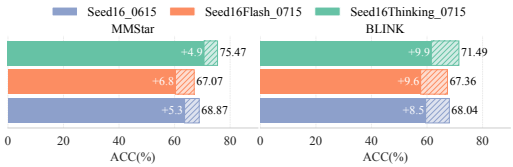
RLVR provides a mechanism for improving multimodal reasoning with verifiable feedback (Zheng et al., 2025; Bai et al., 2025; Kimi Team et al., 2025), but standard training mixtures do not account for the changing usefulness of separate-sequence and unified-canvas organizations during optimization. Our preliminary study in Section 2.1.1 reveals that even leading models exhibit a nearly 10% accuracy gap when source-image content and pixels are preserved by the benchmark renderer before model-native preprocessing, as shown in Figure 1b. This highlights a systematic presentation sensitivity when semantic instances occupy the same visual coordinate space. We therefore ask:

How can we diagnose and optimize the presentation sensitivity caused by unifying semantically equivalent multimodal inputs on one canvas?

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.



(a) Two input paradigms: *Left*: Text and images are provided as independent modalities. *Right*: PiP-All renders text and indexed images within one canvas.



(b) The hatched segment indicates the accuracy gap relative to the original input setting.

Figure 1: Pilot study overview and results. (a): Original-Interleaved and PiP-All inputs for the same semantic query. (b): Accuracy of Doubao-Seed models on multi-instance queries.

We first construct a PiP benchmark with a flagship challenge and a paired diagnostic. PiP-All rasterizes the query, indices, and visual instances within one canvas, directly testing fully rasterized multimodal-scene reasoning. PiP-Visual preserves image ordering and source pixels at the rendering stage but supplies the query only as language tokens, probing the contribution of recovering instructions from pixels. Each paired format is evaluated through the same model-native preprocessing and decoding pipeline. We assign each question one dominant capability label: Sub-region Referencing (SR) for local reference resolution, or Cross-region Compositional Reasoning (CR) for integration across regions. Leading models suffer drops exceeding 10%, and most of the gap remains under PiP-Visual on three representative models.

We propose **I**nformation-**G**eometric **O**nline **A**lignment (IGOA) for multi-context RLVR. DoReMi and DoGE estimate pretraining mixtures with proxy models (Xie et al., 2023; Fan et al., 2023); IGOA coordinates contexts online in the RLVR policy. It computes active-Fisher log-volume gain from learnable-frontier samples relative to accumulated Fisher memory, prioritizing policy-sensitive directions not covered by prior updates. Mirror descent converts these gains into smooth weights across Original-Interleaved and Unified-Canvas contexts.

Experimental results demonstrate that IGOA consistently outperforms strong token- and sequence-level RLVR baselines (Yu et al., 2025; Zheng et al., 2025) on the PiP benchmark and five standard multimodal benchmarks. The gains across architectures and scales show that the improvements are not confined to the PiP-All evaluation condition. In summary, our main contributions are:

- We introduce PiP-All as a fully rasterized challenge, together with the paired PiP-Visual query-modality diagnostic, and further define mutually exclusive dominant SR and CR labels to enable fine-grained analysis of instance-level failures.
- We propose IGOA, which applies active-Fisher log-volume gain to multi-context RLVR coordination through accumulated Fisher memory and mirror-descent updates, with a diagonal approximation for scalable MLLM training.
- Paired query-modality and renderer-scale controls characterize unified-canvas sensitivity beyond query rasterization and at multiple raw sub-image scales. Paired post-training evaluation and an allocation-matched static control test Original retention and the benefit of online coordination, while IGOA also improves five standard multimodal benchmarks.

2 PiP Benchmark

Existing VQA benchmarks typically decompose multi-instance problems into independent image-text pairs, allowing MLLMs to rely on explicit sequential indexing. However, it remains unclear whether MLLMs can maintain consistency when heterogeneous instances are unified within a shared visual reference frame. To investigate this, we remap MMStar (Chen et al., 2024) and BLINK (Fu et al., 2024) inputs into PiP-All canvases, as shown in Figure 1a, while preserving source-image content and pixels at the rendering stage, and evaluate the Doubao-Seed series (ByteDance Seed, 2025) with its released inference pipeline. Figure 1b shows a substantial discrepancy under unchanged task semantics, reaching 9.9% on BLINK. This motivates PiP as a diagnostic tool for high-density multi-instance scenarios.

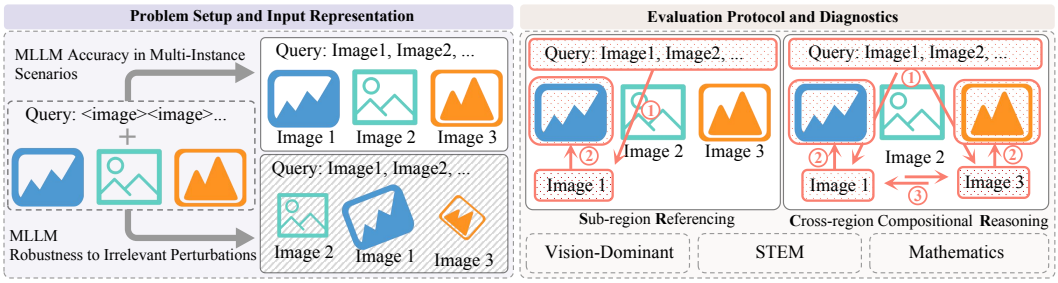


Figure 2: Overview of the PiP benchmark. *Left*: Original samples are transformed into the fully rasterized PiP-All canvas. PiP-Visual preserves image ordering and source pixels at the rendering stage while omitting the query from the canvas and supplying it only as language tokens. *Right*: PiP evaluates sub-region referencing and cross-region compositional reasoning across domains.

2.1 Benchmark Design and Protocol

2.1.1 Problem Setup and Input Representation

As illustrated in Figure 2, PiP-All, denoted x_i^A, places the query and one or more indexed visual instances within one coordinate space. This includes single-image samples, which use the same text–image–marker rendering with one visual instance. PiP-Visual, denoted x_i^V, preserves image ordering and source pixels at the rendering stage but supplies the query only as language tokens. The Original-Interleaved input is denoted x_i^O.

For model m, let P_m denote its released multimodal preprocessing pipeline and M_m the subsequent inference model, giving the end-to-end system $F_m = M_m \circ P_m$. With the same P_m and decoding configuration for every paired format, the PiP Gap is $\Delta_m^{\text{PiP}} = \text{ACC}(F_m(x^O)) - \text{ACC}(F_m(x^A))$. This estimand measures presentation sensitivity of the released system, including the native visual allocation induced by each representation, and supports a common behavioral test across open and proprietary systems without replacing their model-specific processors.

We additionally report the query-token-controlled residual $\text{ACC}(F_m(x^O)) - \text{ACC}(F_m(x^V))$ and the query-modality intervention delta $\text{ACC}(F_m(x^V)) - \text{ACC}(F_m(x^A))$. Question content, answers, renderer-level source-image pixels, instance order, model-native preprocessing, and decoding settings are fixed within each paired comparison. Accordingly, PiP-All serves as the presentation stress test, while PiP-Visual and scale controls probe the contributions of query modality and raw rendering scale. Details are provided in Appendix A.1. This paired design changes presentation rather than task identity. Original-Interleaved exposes image boundaries through the input sequence, while PiP-All requires recovering the same indices, query, and evidence from a common pixel coordinate system.

2.1.2 Data Sources and Question Annotation

We consolidate 8,515 deduplicated candidates from seven VQA datasets for annotation and behavior-based selection. Raw source counts and filtering details are provided in Appendix A.2. We organize the benchmark into three model-agnostic reporting categories: Vision-Dominant, STEM, and Mathematics. Because every canvas contains at least one referenced visual instance, three annotators make a forced choice of one dominant label: SR for resolving the referenced local region, or CR for integrating evidence across visual regions. GPT5_0807, Gemini-2.5-pro-03-25, and DoubaoSeed16Thinking_0715 provide the labels, which are aggregated by majority vote.

2.1.3 Evaluation Protocol and Diagnostics

Beyond standard accuracy ACC, let $z_i \in \text{SR}, \text{CR}$ denote the mutually exclusive dominant label assigned to each benchmark sample. We compute:

$$\begin{aligned} \text{SRA} &= \mathbb{E}\big[\mathbf{1}\{M(x_i^A) = a_i^*\} \mid z_i = \text{SR}\big], \\ \text{CRA} &= \mathbb{E}\big[\mathbf{1}\{M(x_i^A) = a_i^*\} \mid z_i = \text{CR}\big], \end{aligned} \tag{1}$$

where a_i^* is the ground truth. Because SR and CR form a partition, every reporting population satisfies $\mathrm{ACC} = \pi_{\mathrm{SR}}\mathrm{SRA} + (1 - \pi_{\mathrm{SR}})\mathrm{CRA}$, where $\pi_{\mathrm{SR}} = \Pr(z_i = \mathrm{SR})$ is computed on that same population. Unless stated otherwise, PiP metrics use the flagship PiP-All setting. To assess sensitivity to visual variations, we introduce Perturbation Repeatability (PR), defined as prediction consistency under controlled visual transformations:

$$\mathrm{PR} = \mathbb{E}\left[\mathbf{1}\left\{M(x_i^{\mathrm{pert}}) = M(x_i^A)\right\}\right], \tag{2}$$

where x_i^{pert} is a perturbed PiP-All input. We additionally report perturbed-input accuracy $\mathrm{PACC} = \mathbb{E}[\mathbf{1}\{M(x_i^{\mathrm{pert}}) = a_i^*\}]$, which distinguishes stable correctness from stable error. PR is therefore interpreted jointly with ACC and PACC.

Finally, we use a behavior-based discriminative score D_i to select informative failures for diagnostic stress testing rather than average-case estimation. The score combines input-format sensitivity Δ_i, cross-model disagreement H_i, and perturbation instability S_i, as detailed in Appendix A.3. Random PiP and Full PiP controls retain 5.07–9.10 point gaps on strong closed-source models, showing that filtering amplifies rather than creates the phenomenon, as analyzed in Appendix A.4.5. A zero-task probe further shows high coarse-region IoU despite lower task accuracy, with details in Appendix A.3.3. Together, these controls show that query rasterization, raw scale, and coarse boundary localization do not individually explain the measured gap, supporting a residual presentation-related bottleneck under the evaluated preprocessing pipelines.

3 METHODS

To improve unified-canvas reasoning without training a PiP-only policy, we propose Information-Geometric Online Alignment (IGOA). Rather than treating visual contexts as a static mixture, IGOA formulates multi-context RLVR as online trajectory coordination: context-induced policy updates are weighted by their non-redundant active-Fisher contribution in a fixed parameterization.

3.1 CROSS-CONTEXT PERCEPTION AND JOINT OPTIMIZATION

We construct two semantically equivalent yet visually complementary context spaces. The Original-Interleaved Domain $\mathcal{D}_M$ consists of interleaved image-text sequences with explicit instance boundaries. The Unified-Canvas Domain $\mathcal{D}_S$ uses PiP-All and requires grounding rasterized instructions and binding indexed instances within one coordinate system. Departing from fixed-ratio mixing, IGOA learns a simplex-valued coordination variable $\lambda_{d,t}$ that evolves online while avoiding abrupt changes in the training mixture. Here, t indexes controller updates, and the selected weights are held fixed for the intervening optimizer steps. At controller state t, the surrogate objective is:

$$\mathcal{J}(\theta_t) = \sum_{d \in \{M,S\}} \lambda_{d,t} \mathbb{E}_{q \sim \mathcal{D}_d, \{o_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \Psi(o_i, \theta_t) \right], \tag{3}$$

where G denotes the rollout group size, and Ψ represents the Group-Relative Advantage operator at a specific granularity, either token-level or sequence-level.

The two domains contain paired task semantics but induce different context-dependent policy updates. A fixed 1:1 mixture assumes their usefulness remains constant throughout RLVR, whereas $\lambda_{d,t}$ allows their contribution to evolve with the policy. This is the operational meaning of alignment in IGOA: coordinating training across the two organizations so that gains in unified-canvas reasoning do not come from optimizing an isolated PiP-only policy. The same coordination rule applies to either feedback granularity through Ψ.

3.2 INFORMATION-GEOMETRIC MANIFOLD MEASURE

To quantify the substantive contribution of each context space, IGOA identifies information located on the current learning frontier. We define the Active Fisher Information Matrix (AFIM) $F_{d,t}^{\mathrm{active}}$, which measures parameter sensitivity induced by samples that are challenging yet learnable under the current policy. For response o, we use the length-normalized score $\bar{s}_\theta(o|q) =$

$\nabla_\theta[L_o^{-1}\sum_{\ell=1}^{L_o}\log\pi_\theta(o_\ell|q,o_{<\ell})]$, preventing response length from scaling the statistic. AFIM retains the Fisher score outer-product geometry while conditioning it on the non-degenerate RLVR frontier.

$$F_{d,t}^{\text{active}} = \mathbb{E}_{q\sim\mathcal{B}_{d,t},o\sim\pi_{\theta_{\text{old}}}}\left[\mathbf{1}[0<c(q)<G]\bar{s}_{\theta_{\text{old}}}(o|q)\bar{s}_{\theta_{\text{old}}}(o|q)^T\right], \tag{4}$$

where $c(q)$ is the number of responses with positive verifiable correctness reward in the group. The correctness signal defines the active frontier, while the policy objective retains the combined correctness-and-format reward described in Appendix B.4. The indicator $\mathbf{1}[0<c(q)<G]$ restricts AFIM to challenging yet learnable samples. We estimate its diagonal with one Rademacher–Hutchinson probe (Hutchinson, 1989), obtained by a single reverse-mode VJP rather than per-response backward passes. Independent signs eliminate cross-response terms in expectation, making the diagonal estimator conditionally unbiased. Appendix B.3 gives the construction, proof, and bias induced by the nonlinear gain. Building upon this, we introduce the Incremental Information Gain $\Delta\mathcal{I}_{d,t}$ to measure the non-redundant contribution of d relative to accumulated Fisher memory:

$$\Delta\mathcal{I}_{d,t} = \log\det(A_t + F_{d,t}^{\text{active}} + \epsilon I) - \log\det(A_t + \epsilon I), \tag{5}$$

where $A_{t+1} = \alpha A_t + \sum_d \lambda_{d,t} F_{d,t}^{\text{active}}$ denotes the accumulated Fisher-memory update. This log-volume expansion discounts update directions already represented in memory and prioritizes context trajectories with under-covered active policy sensitivity. Appendix B.9 examines the contribution of the active frontier condition and reports sensitivity to rollout group size. Appendix B.1 derives the log-volume decomposition of $\Delta\mathcal{I}_{d,t}$ and its diagonal specialization.

Under the diagonal implementation used in training, the gain has the coordinate-wise form

$$\Delta\mathcal{I}_{d,t} = \sum_k \log\left(1 + \frac{\widehat{f}_{d,t,k}}{a_{t,k}+\epsilon}\right). \tag{6}$$

The ratio discounts coordinates already covered by memory, distinguishing marginal information from raw Fisher magnitude.

The active condition is important in RLVR because groups with uniformly incorrect or uniformly correct responses provide degenerate correctness-based credit assignment. Restricting the estimate to $0<c(q)<G$ focuses the geometric score on the current learnable frontier. Accumulated memory then distinguishes frontier sensitivity that expands under-covered parameter directions from sensitivity that repeats directions already emphasized by earlier updates. Thus, the coordination signal captures marginal policy information rather than reward uncertainty or raw gradient magnitude alone.

3.3 Online Coordination via Mirror Descent

To stabilize coordination under non-stationary RLVR training, we formulate the evolution of λ_t as an online decision-making problem over the simplex. IGOA employs Online Mirror Descent (OMD) so that the coordination variable can respond to newly observed information while remaining close to its previous state. Given the incremental information scores $\Delta\mathcal{I}_t$, we update λ_{t+1} by solving:

$$\lambda_{t+1} = \arg\max_{\lambda\in\Delta}\langle\lambda,\Delta\mathcal{I}_t\rangle - \frac{1}{\eta_t}\sum_d \lambda_d\log\frac{\lambda_d}{\lambda_{d,t}}. \tag{7}$$

The first term favors context trajectories with larger non-redundant information gain, while the KL term prevents abrupt changes in the coordination state. This variational problem yields the following exponentiated-gradient update:

$$\lambda_{d,t+1} = \frac{\lambda_{d,t}\exp(\eta_t\Delta\mathcal{I}_{d,t})}{\sum_{d'}\lambda_{d',t}\exp(\eta_t\Delta\mathcal{I}_{d',t})}. \tag{8}$$

Unlike a memoryless softmax rule, this update preserves the previous coordination state λ_t, allowing IGOA to evolve smoothly under changing information signals. We use an adaptive step size scaled by historical information fluctuations, with details and derivation provided in Appendix B.2. The resulting weights coordinate context-induced optimization paths by marginal information gain without requiring a fixed mixture schedule.

Table 1: Main evaluation results under PiP-All. Overall columns report Original-Interleaved accuracy, PiP-All accuracy, and Perturbation Repeatability (PR). Domain columns directly report the SR/CR decomposition within Mathematics, STEM, and Vision-Dominant questions. † marks models whose outputs contributed to behavior-based selection.

Model	Overall			Mathematics		STEM		Vision-Dominant	
	Ori ACC	ACC	PR	SRA	CRA	SRA	CRA	SRA	CRA
GPT5_0807†	63.91	49.33	16.76	44.50	51.13	51.06	49.81	48.65	49.38
Claude-Sonnet-4.6	73.26	53.96	40.64	50.67	42.62	50.00	34.78	45.12	64.29
Gemini-2.5-pro-03-25†	<u>73.72</u>	58.40	38.57	55.50	53.07	**72.77**	62.26	47.10	61.41
DoubaoSeed16Thinking_0715†	70.02	55.33	29.73	59.16	55.99	57.02	60.38	48.26	51.87
Kimi-K2.5	**82.20**	**64.16**	**54.85**	<u>60.56</u>	**77.36**	61.54	<u>68.42</u>	49.41	63.79
MiMo-VL-7B-SFT-2508	52.67	35.73	21.72	39.27	27.18	36.17	37.74	41.31	35.27
MiMo-VL-7B-RL-2508	55.93	39.73	38.18	39.79	31.72	47.66	36.23	47.49	37.76
Qwen3-VL-30B-A3B-Instruct	57.33	46.57	29.50	46.60	46.60	47.23	40.00	47.88	51.45
Qwen3-VL-30B-A3B-Thinking	63.60	52.20	35.84	48.69	57.61	52.77	42.26	52.90	57.68
Qwen3-VL-235B-A22B-Instruct	67.13	57.67	<u>44.33</u>	53.40	60.84	60.00	59.25	<u>54.05</u>	56.85
Qwen3-VL-235B-A22B-Thinking	73.70	62.33	41.86	52.88	70.23	58.30	62.64	**60.23**	**65.56**
Qwen3.5-35B-A3B	72.47	60.53	36.01	58.12	67.96	62.13	64.91	48.65	59.34
Qwen3.5-122B-A10B	73.27	<u>63.87</u>	40.92	**63.35**	<u>71.52</u>	<u>65.96</u>	**69.06**	47.10	<u>64.73</u>

3.4 Computational Efficiency

Computing the full active Fisher matrix is infeasible for large-scale MLLM training. IGOA maintains coordinate-wise second-order statistics, reducing $\mathcal{O}(P^2)$ matrix operations to $\mathcal{O}(P)$ vector operations. One Rademacher VJP is computed per context every 16 optimizer steps, with λ fixed between controller updates, and the temporary context statistic is released after updating the gain and memory. Appendix B.3 details the estimator, implementation, and its monotonicity and diminishing-return properties. As detailed in Appendix B.9.3, the diagonal AFIM state requires 1.86 GiB per worker for Qwen3-VL-8B and 3.49 GiB for Qwen3-VL-30B-A3B, without additional rollout generation.

4 Experiments

4.1 PiP Benchmark

4.1.1 Experiment Setup

Evaluated Models. We evaluate 13 checkpoints from GPT5 (OpenAI, 2025), Claude-Sonnet-4.6 (Anthropic, 2026), Gemini-2.5-Pro (Google DeepMind, 2025), DoubaoSeed16Thinking (ByteDance Seed, 2025), Kimi-K2.5 (Kimi Team, 2026), MiMo-VL (Xiaomi LLM-Core Team et al., 2025), Qwen3-VL (Bai et al., 2025), and Qwen3.5 (Qwen Team, 2026). Table 1 lists model sizes and tuning variants.

Implementation Details. The PiP benchmark consists of 1.5K highly discriminative samples selected from 8,515 deduplicated candidates drawn from seven open-source single- and multi-image datasets. To ensure evaluation accuracy and consistency, we replace traditional string matching with semantic equivalence assessments leveraging GPT5_0807, with the temperature fixed to 0. Detailed source statistics and evaluation prompts are provided in Appendix A.2 and Appendix A.3.2. Main results use PiP-All. Appendix A.3.4 reports the paired PiP-Visual query-modality diagnostic and renderer-scale controls. Supplying the query as language tokens changes accuracy by at most 0.60 points on three representative models. We verify judge-choice stability by re-evaluating the same fixed answer model on 500 randomly sampled PiP examples with Gemini-2.5-Pro, Doubao-Seed, and five human experts. Detailed results are reported in Appendix A.4.4.

4.1.2 Main Results

We systematically evaluate the performance of leading MLLMs on the PiP benchmark in Table 1. By analyzing quantitative metrics across multiple dimensions, we draw the following key findings.

Table 2: Main results on PiP-All and five standard multimodal benchmarks. PiP columns report accuracy and Perturbation Repeatability (PR). Domain-conditioned SRA/CRA results are reported separately in Appendix B.8. **Bold** and underline denote the best and second-best results. AVG is the arithmetic mean across the five standard benchmarks.

Method	PiP ACC	PR	AVG	BLINK	MuirBench	MathVerse	MMMU_Pro	MMStar
Qwen2.5-VL-7B-Instruct	26.80	12.94	45.35	48.05	52.60	34.70	37.11	54.27
VL-Rethinker-Qwen2.5	30.53	20.99	51.18	53.00	57.46	42.46	41.67	61.33
NoisyRollout-Qwen2.5	31.60	22.14	51.25	52.89	57.80	43.12	40.17	62.26
DAPO_mix	31.93	25.45	50.73	51.74	54.13	42.71	41.67	63.40
GSPO_mix	32.40	26.14	50.63	53.58	57.34	42.06	39.19	61.00
IGOA_token	35.79	**32.17**	54.23	55.63	58.91	45.94	44.17	**66.52**
IGOA_seq	**37.15**	31.69	**54.52**	**56.16**	**59.47**	**46.35**	**44.36**	66.24
MiMo-VL-7B-SFT-2508	35.73	21.72	59.08	56.78	69.13	55.25	46.53	67.73
VL-Rethinker-MiMo	40.10	31.96	61.66	58.02	72.72	57.79	50.23	69.52
NoisyRollout-MiMo	40.73	33.51	61.94	59.02	72.37	58.77	50.88	68.68
DAPO_mix	42.46	37.35	62.44	58.68	74.46	58.88	50.63	69.53
GSPO_mix	42.73	36.32	62.44	58.68	74.46	58.88	50.64	69.53
IGOA_token	47.22	39.69	**66.41**	**62.08**	**77.26**	**64.37**	**55.21**	73.15
IGOA_seq	**47.26**	**42.93**	65.90	62.03	76.99	62.66	53.64	**74.20**
Qwen3-VL-8B-Instruct	43.27	35.86	61.90	64.08	70.20	51.87	53.23	70.13
DAPO_mix	46.87	45.83	64.19	65.08	71.53	55.61	56.01	72.73
GSPO_mix	49.80	43.60	65.59	67.63	73.50	55.91	56.87	74.06
IGOA_token	**54.79**	**51.77**	68.34	**70.20**	**76.48**	59.92	58.69	76.40
IGOA_seq	54.13	50.73	**69.01**	68.60	76.43	**61.99**	**60.55**	**77.47**
Qwen3-VL-30B-A3B-Instruct	46.57	29.50	63.38	60.36	71.00	56.42	58.20	70.93
DAPO_mix	47.64	35.97	65.42	63.37	73.61	58.44	59.42	72.26
GSPO_mix	48.26	29.39	64.44	60.31	71.88	58.19	59.57	72.26
IGOA_token	53.54	**47.89**	**70.15**	66.04	**78.47**	63.69	**65.12**	77.45
IGOA_seq	**53.94**	40.71	69.67	**66.31**	76.46	**64.87**	62.93	**77.79**

Performance degradation under unified presentation. Empirical results reveal a substantial *PiP Gap* across all evaluated models when transitioning from Original-Interleaved inputs to PiP-All. Kimi-K2.5 obtains the highest PiP-All accuracy at 64.16%, yet drops by 18.04 points from its 82.20% Original-Interleaved accuracy; Claude-Sonnet-4.6 similarly drops by 19.30 points. The gap therefore remains pronounced in recent frontier systems.

Perturbation consistency and the effectiveness of RL. Most models exhibit PR scores between 20% and 40%, revealing strong prediction sensitivity to layout and appearance variations. Since PR measures consistency rather than correctness, we interpret it jointly with PiP-All ACC and PACC. Domain-conditioned SRA and CRA separately diagnose which dominant capability is associated with errors. Scaling improves baseline capability, while RL-tuned variants generally show better consistency. Appendix A.4 analyses the accuracy gap Δ, cross-model inconsistency H, and prediction instability S_i. Appendix A.5 provides supplemental results.

4.2 IGOA

4.2.1 Datasets and Implementation Details

To enhance fine-grained reasoning under complex visual scenarios, we construct a training dataset based on MMPR-Tiny[1]. The dataset is derived from the online reinforcement learning stage of InternVL3.5 (Wang et al., 2025d), with the goal of strengthening the model's image understanding and reasoning capabilities. During data preparation, we incorporate all 4K multi-image samples and additionally randomly sample 6K single-image samples to maintain distributional balance, forming an Original-Interleaved Domain $\mathcal{D}_M$ of 10K samples. For IGOA, we apply PiP-All rendering to each sample in $\mathcal{D}_M$ while preserving semantic consistency, thereby constructing the corresponding Unified-Canvas Domain $\mathcal{D}_S$. Unless otherwise stated, we freeze the vision encoder and update the language model and multimodal projector during RLVR fine-tuning. AFIM and the accumulated memory A_t are computed only over trainable parameters. Training hyperparameter configurations and resource consumption are reported in Appendix B.5.

[1] https://huggingface.co/datasets/OpenGVLab/MMPR-Tiny

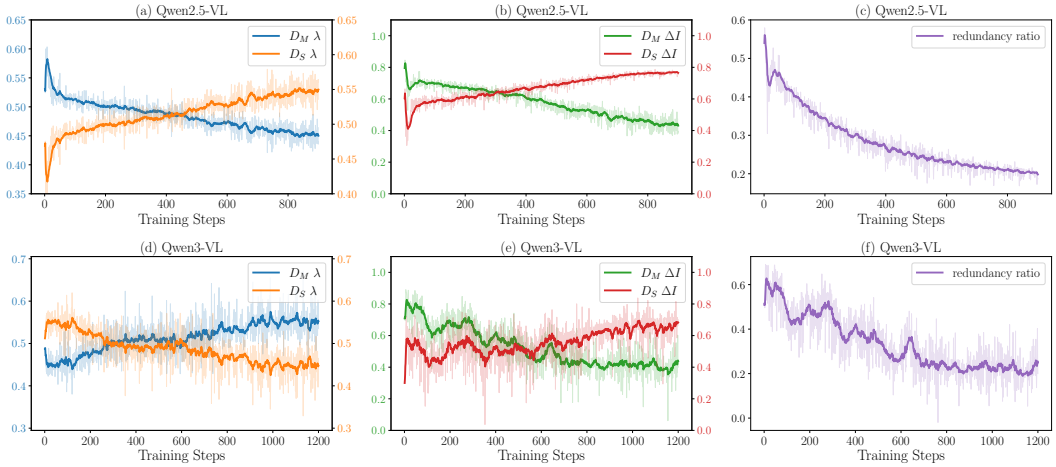


Figure 3: Training dynamics of IGOA: coordination variables $\lambda_{d,t}$ (a, d), non-redundant information gain $\Delta\mathcal{I}_d$ (b, e), and diagonal-Fisher overlap $\mathcal{R}_t$ across contexts (c, f), defined in Appendix B.3. The two coordination weights are complementary and should be read against their matching axes. *Qwen2.5-VL* and *Qwen3-VL* denote Qwen2.5-VL-Instruct and Qwen3-VL-30B-A3B-Instruct.

4.2.2 Baselines

We evaluate IGOA across various model backbones and feedback granularities. Specifically, we compare IGOA_token against the representative token-level method DAPO, and IGOA_seq against the sequence-level baseline GSPO. All DAPO, GSPO, and IGOA comparisons use the same 20K paired-context dataset, equal Original-Interleaved and Unified-Canvas rollout counts before weighting, the same filter $0 < c(q) < G$, and matched rollout and optimization budgets. DAPO_mix and GSPO_mix are fixed-mix baselines, whereas $\lambda_{d,t}$ changes the objective weight rather than query exposure or rollout count. To separate online coordination from its average allocation, Static-$\bar{\lambda}$ is retrained from the same initialization with λ_d fixed to IGOA's time average $\bar{\lambda}_d = T^{-1}\sum_{t=1}^{T}\lambda_{d,t}$, while matching data, per-context rollouts, optimizer steps, rewards, and random seed. For the 7B backbones, our evaluation includes two approaches: VL-Rethinker (Wang et al., 2025b), which employs iterative reflection for self-correction, and NoisyRollout (Liu et al., 2025a), which enhances reasoning robustness in complex spatial layouts via noise injection. To investigate scaling laws, we conduct comparative analyses between IGOA and its baselines using the Qwen3-VL series.

4.2.3 Main Results

Table 2 shows that IGOA consistently outperforms the granularity-matched baselines across all eight comparisons, improving PiP-All ACC, PR, and five-benchmark AVG by 5.22, 7.19, and 4.04 points on average, respectively. The gains hold across backbones, model scales, and both surrogate granularities. Appendix B.8 provides the domain-conditioned decomposition. Table 9 further shows that IGOA preserves Original-Interleaved performance while substantially narrowing the Original-to-PiP gap. More importantly, compared with Static-$\bar{\lambda}$, which matches IGOA's average context allocation, IGOA improves PiP ACC by 7.48 points and General AVG by 3.96 points. This allocation-matched control indicates that the gains arise from online trajectory coordination rather than a favorable fixed mixture. Controlled ablations in Appendix B.9 further verify the contributions of dynamic coordination, accumulated Fisher memory, and active-frontier filtering under matched data and per-context rollout exposure. Appendix B.6 reports matched three-seed uncertainty, while consistent gains on the five standard benchmarks indicate that the improvement is not confined to PiP-All.

4.2.4 Analysis of Online Trajectory Coordination

Figure 3 tracks non-redundant gain $\Delta\mathcal{I}_d$, coordination weights $\lambda_{d,t}$, and diagonal-Fisher overlap $\mathcal{R}_t$. The learned allocation is backbone-dependent. On Qwen2.5-VL, λ_S rises as the marginal gain of $\mathcal{D}_S$ increasingly exceeds that of $\mathcal{D}_M$. On Qwen3-VL-30B-A3B, λ_S instead decreases from

roughly 0.55 to 0.45 while λ_M increases. In both cases, the controller responds to the realized history rather than imposing a fixed curriculum direction. Accumulated Fisher memory discounts directions covered by earlier updates, while OMD converts marginal gain into smooth weight changes rather than a fixed mixture or memoryless difficulty rule. Panels (c, f) additionally show the overlap between context-specific diagonal Fisher vectors. Controlled ablations in Appendix B.9 evaluate the downstream contribution of the signal, memory, and active frontier.

5 Related Work

Multimodal Reasoning Benchmarks. MLLM benchmarks span perception, expert knowledge, compositional reasoning, and mathematical or diagram understanding (Li et al., 2024; Yue et al., 2024; Zhang et al., 2024b; Liu et al., 2024; Xu et al., 2026; Jiang et al., 2025; Qiao et al., 2025). General-purpose suites establish capability breadth, but typically present language instructions and visual inputs in distinct modality slots. Multi-image benchmarks such as BLINK and MuirBench introduce comparison, correspondence, and evidence integration across visual instances while retaining explicit image boundaries in the input sequence (Fu et al., 2024; Wang et al., 2025a). CollagePrompt concatenates images to study recognition efficiency and spatial arrangement effects (Xu et al., 2024), while document, GUI, and map benchmarks evaluate reasoning over naturally dense visual structures (Du et al., 2026; Luo et al., 2025; Feng et al., 2025). PiP studies a complementary question: whether an MLLM preserves behavior when the same query, answer, source images, and instance order are reorganized from an interleaved sequence into one canvas. Its paired design makes presentation sensitivity directly measurable. PiP-All captures the complete rasterized setting, PiP-Visual separates query modality, and SR/CR labels expose local-reference and cross-region requirements. The benchmark thereby links controlled evaluation of representation consistency with post-training on the same paired contexts.

RLVR and Information-Geometric Optimization for MLLMs. RLVR uses verifiable outcomes to improve mathematical, language, and multimodal reasoning (Shao et al., 2024; Wang et al., 2025c; Zhang et al., 2025). Multimodal extensions apply this paradigm to segmentation, fine-grained grounding, and visual reasoning (Liu et al., 2025b; Meng et al., 2025a; Shen et al., 2025); DAPO and GSPO provide representative token- and sequence-level optimization surrogates (Yu et al., 2025; Zheng et al., 2025). A related line of work adapts training mixtures: DoReMi uses proxy-model excess loss, while DoGE estimates generalization effects across domains (Xie et al., 2023; Fan et al., 2023). IGOA instead coordinates paired presentation contexts online as their contributions evolve with the RLVR trajectory. Its construction connects three established ideas. Fisher information measures local parameter sensitivity and defines the geometry underlying natural-gradient optimization (Amari et al., 2019). Log-determinant objectives quantify information volume and reward complementary directions (Krause et al., 2008). Mirror descent supplies multiplicative updates on the probability simplex (Nemirovski et al., 2009; Shalev-Shwartz, 2011). IGOA combines active-frontier Fisher estimates, accumulated diagonal memory, and marginal log-volume gain in one controller. The same coordination rule applies to token- and sequence-level RLVR while preserving optimization history through its mirror-descent state.

6 Conclusion

We study whether MLLMs preserve behavior when semantically equivalent interleaved inputs are reorganized on a unified canvas. PiP turns this question into a paired evaluation across 13 checkpoints, with PiP-All measuring the complete rasterized setting and PiP-Visual separating the contribution of query modality. The persistent gap across model families establishes presentation organization as a distinct axis of multimodal reasoning performance. We further introduce IGOA, which coordinates Original-Interleaved and Unified-Canvas RLVR contexts through active-Fisher marginal log-volume gain, accumulated memory, and mirror-descent updates. Across eight granularity-matched comparisons, IGOA improves PiP ACC, PR, and five-benchmark AVG by 5.22, 7.19, and 4.04 points on average. On Qwen3-VL-8B, it reaches 54.79% PiP ACC versus 47.31% for the allocation-matched Static-$\bar{\lambda}$ control. Consistent gains across four backbone groups and both RLVR granularities support information-geometric online alignment for unified-canvas reasoning.

AI USE STATEMENT

We used generative AI tools to assist with dataset annotation, model-output evaluation, and manuscript refinement. Specifically, we used GPT-5, Gemini-2.5-Pro, and Doubao-Seed as independent annotators for the dominant reasoning-requirement labels, with the final labels determined by majority vote, and we additionally used GPT-5 as a semantic-equivalence judge for model responses. We also used generative AI for language refinement and to improve the clarity of technical and mathematical exposition. We developed and verified the research methodology, experimental design, scientific claims, and interpretation of results ourselves. We reviewed all AI-assisted annotations and evaluations for correctness and consistency, and further examined evaluation robustness using independent model judges and human experts. We take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We provide an anonymous implementation of PiP and IGOA at `https://anonymous.4open.science/r/IGOA_PiP-EA76`. Detailed benchmark construction and evaluation protocols are provided in Appendices A.1–A.3, while the IGOA pseudocode, training configuration, and implementation details are given in Appendices B.4–B.5. We additionally report multi-seed robustness, allocation-matched controls, and ablation studies in Appendices B.6–B.9. All training and evaluation data are constructed from publicly available datasets, with deduplication and train–evaluation overlap checks explicitly documented.

REFERENCES

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher information and natural gradient learning in random deep networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 694–702. PMLR, 2019.

Anthropic. Introducing claude sonnet 4.6, 2026. URL `https://www.anthropic.com/news/claude-sonnet-4-6`.

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

ByteDance Seed. Doubao-seed-1.6. `https://seed.bytedance.com/en/seed1_6`, 2025.

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

Yong Du, Yuchen Yan, Fei Tang, Zhengxi Lu, Chang Zong, Weiming Lu, Shengpei Jiang, and Yongliang Shen. Test-time reinforcement learning for gui grounding via region consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30593–30601, 2026.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation. *arXiv preprint arXiv:2310.15393*, 2023.

Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Google DeepMind. Gemini 2.5 Pro, 2025. URL https://deepmind.google/technologies/gemini/. Large multimodal language model.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806.

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

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2024.

Kimi Team. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025a.

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.

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

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025a.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 38405–38453, 2025b.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

OpenAI. GPT-5. https://openai.com, 2025. Large language model.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025. doi: 10.18653/v1/2025.acl-long.983.

Qwen Team. Qwen3.5. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-35B-A3B.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011. doi: 10.1561/2200000018.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025a.

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*, 2025b.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-08-20.

Xiaomi LLM-Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, et al. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, 2023.

Siyu Xu, Yunke Wang, Daochang Liu, Bo Du, and Chang Xu. Collageprompt: A benchmark for budget-friendly visual recognition with gpt-4v. *arXiv preprint arXiv:2403.11468*, 2024.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

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

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 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, 2024a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A PiP Benchmark **13**

A.1 PiP Rendering and Perturbation Details 13

A.2 Multi-Source Data and Attribute Labels Details 14

A.3 Evaluation Protocol and Diagnostics Details 15

A.4 Additional Analysis of the PiP Benchmark 18

A.5 Main Results and Diagnostic Analysis 20

A.6 PiP Benchmark Example 21

B Information-Geometric Online Alignment **23**

B.1 Log-Volume Interpretation of Incremental Information Gain 23

B.2 Online Mirror Descent Update 24

B.3 Diagonal Approximation and Scalability 25

B.4 Pseudocode for IGOA 26

B.5 Implementation Details 26

B.6 Training-Seed Robustness 27

B.7 Paired Alignment and Allocation-Matched Control 28

B.8 Main Results on PiP Benchmark 28

B.9 Ablation Study 29

A PiP Benchmark

A.1 PiP Rendering and Perturbation Details

Each PiP input is generated from an original VQA sample through deterministic composition without changing the task content or ground-truth answer a^*. We distinguish the fully rasterized PiP-All setting from the paired PiP-Visual query-modality diagnostic.

Table 3: Raw source pool before cross-benchmark deduplication and validity filtering (9,115 samples in total).

Type	Benchmark	Link	Samples
Single-Image	MathVerse Zhang et al. (2024a)	https://huggingface.co/datasets/AI4Math/MathVerse	788
	MMStar Chen et al. (2024)	https://huggingface.co/datasets/Lin-Chen/MMStar	1500
	RealworldQA xAI (2024)	https://huggingface.co/datasets/xai-org/RealworldQA	765
Multi-Image	BLINK Fu et al. (2024)	https://huggingface.co/datasets/BLINK-Benchmark/BLINK	726
	MMMU_Pro Yue et al. (2025)	https://huggingface.co/datasets/MMMU/MMMU_Pro	138
	MuirBench Wang et al. (2025a)	https://huggingface.co/datasets/MUIRBENCH/MUIRBENCH	2598
	ReMI Kazemi et al. (2024)	https://huggingface.co/datasets/mehrankazemi/ReMI	2600
		Total	**9,115**

Given an image sequence $I_1, \ldots, I_n, n \geq 1$ and question q, PiP-All x^A renders both q and the indexed images into one canvas. Each sub-image receives an *Image i* label preserving its original order, and images are arranged using a regular grid or near-square layout. For $n > 1$, the column count C is dynamically selected from $1, n-1$, and $R = \lceil n/C \rceil$ in row-major order. For $n = 1$, we set $C = R = 1$. Row heights and column widths equal the maxima of their assigned images, with fixed spacing (pad_x=20, pad_y=10).

Fixed pixel-level spacing is maintained between adjacent sub-images to prevent visual overlap or boundary ambiguity. The question text is rendered in Microsoft YaHei (at 50pt) and embedded in a dedicated region at the top of the canvas. Any *Image* markers appearing in the question are replaced with indexed markers *Image i* and rendered in the same font family at 50 pt or 20 pt depending on the layout scale, with a deterministic fallback to *Arial* or *DejaVu Sans* to ensure consistent appearance across environments. No other prompts or visual cues are introduced. PiP-Visual x^V preserves image ordering and source pixels at the rendering stage, omits q from the canvas, and supplies the same q only as language tokens, so the model need not recover its instruction from pixels. Thus, $\mathrm{ACC}(F_m(x^O)) - \mathrm{ACC}(F_m(x^V))$ is the query-token-controlled residual, while $\mathrm{ACC}(F_m(x^V)) - \mathrm{ACC}(F_m(x^A))$ is the query-modality intervention delta. Together with the scale controls, these paired differences isolate major construction factors while preserving task semantics.

Each checkpoint is evaluated through its released native multimodal processor and inference pipeline. For a given checkpoint, Original-Interleaved, PiP-All, PiP-Visual, and all perturbation variants use the same processor configuration, decoding settings, and answer judge. Consequently, the PiP Gap is an end-to-end system property under the checkpoint's intended interface.

For the PR repeatability stress test, we generate PiP-All variants using (i) bounded rotations sampled from $[-45°, 45°]$; (ii) isotropic scaling in $[0.5, 1.5]$; (iii) translations up to ± 30 pixels; and (iv) brightness/contrast changes within $\pm 20\%$. We additionally vary global layout and background using solid colors, Gaussian noise (offset ± 20), and structural stripes (10–40px). The question, source images, and reference answer remain fixed, while the transformations intentionally vary visual difficulty. Fixed seeds produce identical variants across models. We generate three variants per sample and use PR only as the clean-to-perturbed prediction-agreement diagnostic in Equation 2.

A.2 Multi-Source Data and Attribute Labels Details

The seven source datasets contribute 9,115 raw samples, as detailed in Table 3. Removing cross-benchmark overlap and invalid records during source consolidation yields 8,515 candidates. Behavior-based selection then produces the final 1.5K diagnostic benchmark. Deduplication covers normalized question–answer content and source images across datasets, and the retained pool is checked against MMPR-Tiny before evaluation. The sources span mathematics, STEM, commonsense reasoning, and vision-centric tasks, with both single-image and multi-image inputs. All retained samples preserve their question content and ground-truth answers while changing only the input presentation.

We define a region relative to the unified canvas. Every sample contains the rendered question, at least one indexed marker, and one or more visual instances. Single-image samples use the same construction with *Image 1*. Labeling is therefore a forced choice $z_i \in \mathrm{SR}, \mathrm{CR}$ rather than two independent binary decisions. SR denotes resolving the referenced local region that contains the evidence, whereas CR denotes integrating evidence across two or more visual regions, either within

or across indexed instances, through comparison, aggregation, exclusion, conditional reasoning, or consistency checking. CR takes precedence when both operations are present.

GPT5_0807, Gemini-2.5-pro-03-25, and DoubaoSeed16Thinking_0715 independently assign the dominant label, and majority vote produces exactly one final label for every sample. Each annotator receives the rendered unified canvas together with the question. Annotation uses question semantics and indexed visual structure only, without model difficulty or performance signals. SR/CR supports stratified diagnosis only. Overall ACC, PACC, and PR are independent of these labels. The following prompt implements this forced-choice protocol:

Prompt Template for Annotation Model

You are an assistant that assigns one dominant reasoning requirement to a visual question rendered with indexed visual instances. Output the result in **strict JSON format**.

Evaluation Criteria:

1. **SR—Sub-region Referencing**: choose SR when answering primarily requires identifying or resolving the referenced local region that contains the relevant evidence.

2. **CR—Cross-region Compositional Reasoning**: choose CR when answering requires integrating evidence from two or more visual regions, either within or across indexed instances, including comparison, aggregation, exclusion, conditional reasoning, or consistency checking.

3. Choose **exactly one** label. If both region resolution and cross-region integration are involved, choose CR because it captures the dominant compositional requirement.

Output Requirements:

1. Output **only** a single JSON object containing the fields `question` and `dominant_requirement`.

2. `dominant_requirement` must be exactly `"SR"` or `"CR"`.

3. Do **not** include any explanations, commentary, or additional text.

Example Input:

Visual input: `<image>`

Question: What is the shape in the top-left corner?

Example Output:

{"question":"What is the shape in the top-left corner?","dominant_requirement":"SR"}

Now process the following question:

Visual input: `<image>`

Question: {question}

We release the three individual labels and their majority-voted final label with the benchmark. Before behavior-based selection, the deduplicated candidate pool contains 49.1% SR and 50.9% CR examples. The final 1,500-example benchmark contains 686 SR and 814 CR examples (45.7% and 54.3%, respectively). The three annotators reach unanimous agreement on 81.13% of examples, with Fleiss' $\kappa = 0.747$. The SR/CR partition is used only for stratified diagnosis. The headline ACC, PACC, and PR results do not depend on these labels.

A.3 Evaluation Protocol and Diagnostics Details

A.3.1 Three Core Discriminative Dimensions.

We use the predictive outputs of GPT5_0807, Gemini-2.5-pro-03-25, and DoubaoSeed16Thinking_0715 to compute three complementary dimensions of diagnostic value: input sensitivity, cross-model disagreement, and perturbation instability.

Accuracy Gap. To quantify presentation sensitivity, let $\mathcal{M}$ denote the selection models. For sample i, the Original-to-PiP accuracy gap is:

$$\Delta_i = \mathbb{E}_{M\sim\mathcal{M}}\left[M(x_i^O)\right] - \mathbb{E}_{M\sim\mathcal{M}}\left[M(x_i^A)\right]. \tag{9}$$

Here x_i^O and x_i^A are the Original-Interleaved and PiP-All inputs. A larger $|\Delta_i|$ indicates higher input-format sensitivity and thus stronger discriminability.

Cross-Model Prediction Inconsistency. This dimension captures the degree of predictive disagreement among MLLMs, reflecting epistemic uncertainty. Let $\mathcal{A}_i$ be the set of all candidate answers produced by the MLLMs, and define the empirical answer distribution as:

$$p_i(a) = \frac{1}{|\mathcal{M}|}\sum_{M\in\mathcal{M}} \mathbf{1}M(x_i) = a, \quad a \in \mathcal{A}_i. \tag{10}$$

The entropy is then formulated as: $H(x_i) = -\sum p_i(a)\log p_i(a)$. High-entropy samples indicate a lack of consensus among models and typically correspond to more challenging reasoning tasks, making them suitable for evaluating generalization.

Prediction Instability. Instability assesses prediction sensitivity to the controlled visual stress transformations. For each sample i, we generate K perturbed variants $\{x_i^{\text{pert},j}\}_{j=1}^K$ from PiP-All x_i^A. The sample-level instability score S_i is:

$$S_i = \frac{1}{|\mathcal{M}|\cdot K}\sum_{M\in\mathcal{M}}\sum_{j=1}^{K} \mathbf{1}M(x_i^{\text{pert},j}) \neq M(x_i^A). \tag{11}$$

Higher S_i indicates less stable predictions under the controlled visual variations.

Composite Discriminative Score. We apply min–max normalization to the three dimensions, yielding $\tilde{\Delta}_i$, $\tilde{H}_i$, and $\tilde{S}_i$. The final composite discriminative score is:

$$D_i = \tilde{\Delta}_i + \tilde{H}_i + \tilde{S}_i. \tag{12}$$

This score is entirely data-driven, ensuring transparency and reproducibility. A higher D_i indicates stronger discriminative potential across all three dimensions.

A.3.2 Answer Judgement via LLM-based Evaluator

Given that scenarios such as open-ended responses, paraphrased expressions, and unit conversions cannot be reliably judged through strict string matching, we adopt GPT5_0807 as a unified judge model to perform a binary evaluation of semantic equivalence between the model outputs and the reference answers. To maximize reproducibility, we employ greedy decoding by setting the temperature to 0 and fixing the random seed to 0. The specific prompt template is as follows:

Prompt Template for Judge Model

Role

You are an AI judging assistant responsible for deciding whether two supplied answer strings are semantically equivalent.

Task

Given a ground-truth answer and a model-generated output, evaluate whether the output matches the ground-truth answer.

Input

1. **Correct Answer**: gt_answer
2. **Model's Response**: response

Judging Criteria

1. **Core Judgment**:
1.1. Focuses on whether the model's response is fundamentally correct compared to the correct answer.
2. **Format Insensitivity**
2.1. Case Insensitivity: Ignores capitalization differences in responses (e.g., "apple" and "Apple").
2.2. **Synonyms/Expressions**: Treat synonymous expressions with the same core meaning as equivalent.
3. **Multiple Choice Question Handling**
3.1. If both strings provide option letters, compare the option letters directly.
3.2. If both strings provide answer content, compare their semantic meaning under the format-insensitivity rules above.
3.3. Use only the two supplied strings. Do not infer an option-letter/content correspondence that is not explicitly present in them.
4. **Chain-of-Thought (CoT) Answer Processing**
4.1. The model's answer typically includes a detailed chain of thought steps.
4.2. You need to accurately extract the model's final conclusion or answer from these thought steps.
4.3. The evaluation will be based on whether the model's final answer meets the above criteria. The detail of its thought process or the way it is expressed is not the focus of this evaluation.
5. **Focus on the Final Conclusion**
Even if there are minor flaws in the intermediate steps of the CoT, if the final conclusion is essentially consistent with the correct answer, it should be considered correct.
Output format
For correctly answered questions, output <result>True</result>; for incorrect answers, output <result>False </result>. Your final answer should be enclosed in <result></result>.

A.3.3 Zero-Task Localization Probe

Table 4: Comparison between standard reasoning accuracy and zero-task localization performance.

	Ori ACC	ACC	IoU	IoU>0.5
GPT5_0807	63.91	49.33	0.91	95.4
Gemini-2.5-pro-03-25	73.72	58.40	0.89	94.2
DoubaoSeed16_0615	62.00	45.47	0.87	93.8
Qwen3-VL-30B-A3B-Instruct	57.33	46.57	0.84	92.7

To disentangle fundamental spatial localization from complex logical integration, we implement a zero-task localization probe. Unlike PiP-All reasoning tasks, the model is prompted solely to output the bounding box coordinates $[x_{min}, y_{min}, x_{max}, y_{max}]$ for the sub-images. Localization accuracy is quantified using IoU, with an $IoU > 0.5$ serving as the threshold for success.

As illustrated in Table 4, despite a sharp decline in PiP VQA accuracy, coarse region localization remains strong, with average zero-shot IoU scores exceeding 0.85 and localization success rates above 90%. These results show that models recover coarse sub-image boundaries despite the PiP-All accuracy drop. Together with PiP-Visual and scale controls, the probe shows that coarse boundary localization alone does not account for the gap, suggesting that the remaining errors arise when the end-to-end system must use referenced content within the unified presentation.

Table 5: Paired controls for PiP construction. PiP-All rasterizes the query and images in one canvas. PiP-Visual preserves image ordering and source pixels, omits the query from the canvas, and supplies it only as language tokens. Small/Big ACC scale each sub-image by $0.5\times$/$1.5\times$. Retained Gap is $(\mathrm{ACC}_O - \mathrm{ACC}_V)/(\mathrm{ACC}_O - \mathrm{ACC}_A)$.

Model	Ori ACC	PiP-All	PiP-Visual	Small ACC	Big ACC	Retained Gap
DoubaoSeed1.6	62.00	45.47	46.02	46.51	46.94	96.7%
GPT5	63.91	49.33	49.45	49.13	49.88	99.2%
Gemini-2.5-Pro	73.72	58.40	59.00	59.93	59.21	96.1%

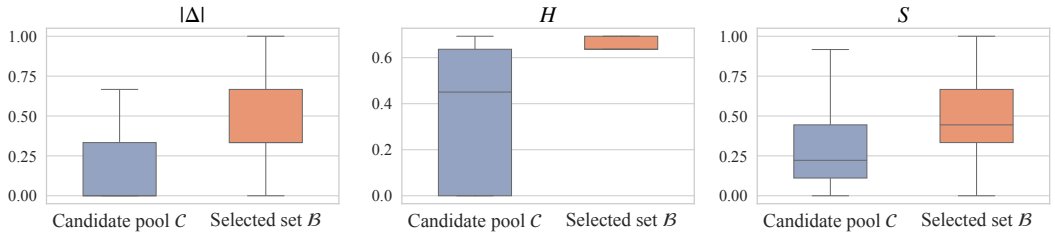


Figure 4: Distributional comparison of discriminative signals. The box plots illustrate the distributions of the candidate sample pool $\mathcal{C}$ and the selected PiP benchmark$\mathcal{B}$ across three signals: $|\Delta|$, H, and S.

A.3.4 Robustness Analysis of PiP Construction

To probe query-modality and scale effects, we compare PiP-All with three variants:

- **PiP-Visual** preserves image ordering and source pixels at the rendering stage, omits the query from the canvas, and supplies it only as language tokens.
- **Small-ACC** reduces each sub-image by 50%.
- **Big-ACC** enlarges each sub-image by 50% to measure sensitivity to increased renderer-side visual scale.

As shown in Table 5, supplying query tokens improves PiP-All by only 0.55, 0.12, and 0.60 points for DoubaoSeed1.6, GPT5, and Gemini-2.5-Pro. Defining $\rho_{\mathrm{ret}} = (\mathrm{ACC}_O - \mathrm{ACC}_V)/(\mathrm{ACC}_O - \mathrm{ACC}_A)$, PiP-Visual retains 96.7%, 99.2%, and 96.1% of the total gap. Small-ACC and Big-ACC differ from PiP-All by at most about 1.5 points. These interventions show that query rasterization and raw rendering scale do not individually explain the measured gap and support a residual presentation-related bottleneck under the evaluated preprocessing pipelines. We retain PiP-All as the complete rasterized stress test and PiP-Visual as its paired query-modality diagnostic.

A.4 Additional Analysis of the PiP Benchmark

This subsection provides supplementary statistics for the PiP benchmark.

A.4.1 Comparison between Candidate Pool and Selected Set

We conduct a statistical analysis of three discriminative signals, namely accuracy gap $|\Delta|$, cross-model prediction inconsistency H, and prediction instability S, to compare the distributional differences between the candidate sample pool $\mathcal{C}$ and the final PiP benchmark $\mathcal{B}$.

As illustrated in Figure 4, the PiP benchmark exhibits systematic distributional shifts relative to the candidate sample pool across all three discriminative signals. Specifically, both the median and interquartile range (IQR) of accuracy gap $|\Delta|$ increase substantially, indicating that the selected samples are more sensitive to variations in input formulation. The shift in cross-model prediction inconsistency H is particularly pronounced, with samples densely concentrated in high-entropy regions, suggesting that MLLMs are less likely to reach consensus on these cases. Furthermore, the overall upward shift in prediction instability S implies that these samples are more prone to predictive

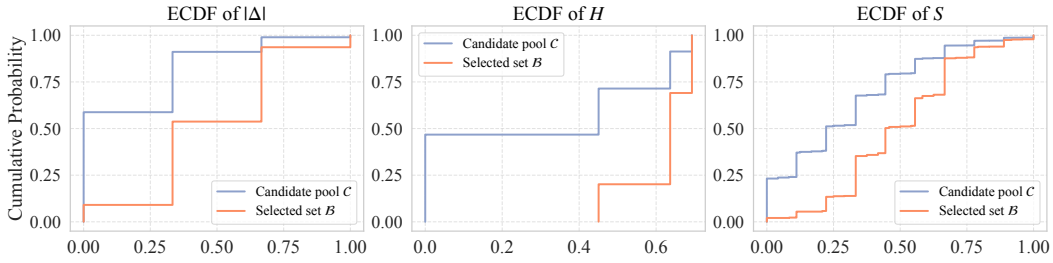


Figure 5: Empirical cumulative distribution function (ECDF) comparison of discriminative signals. The three subplots present the ECDFs of $|\Delta|$, H, and S over the candidate sample pool $\mathcal{C}$ and the selected PiP benchmark $\mathcal{B}$, respectively.

fluctuations under controlled visual stress transformations. Collectively, these findings show that filtering enriches the benchmark for the intended diagnostic signals.

A.4.2 Distributional Shift of Discriminative Signals

To further examine the distributional characteristics of discriminative signals in the PiP benchmark, we plot the empirical cumulative distribution functions (ECDFs) of three metrics for both the candidate pool $\mathcal{C}$ and the final set $\mathcal{B}$.

As shown in Figure 5, the PiP benchmark exhibits a consistent rightward shift across all three metrics. For accuracy gap $|\Delta|$, the candidate-pool ECDF rises sharply at low values, whereas the filtered set has markedly lower cumulative probability in this regime. The shift is also pronounced for cross-model disagreement H, with selection concentrating higher-entropy cases. For prediction instability S, the filtered samples remain right-shifted across the interval, especially in the moderate-variation regime. The ECDFs therefore show broad distributional enrichment rather than a change driven only by a few extreme cases, confirming that the selected benchmark concentrates the intended diagnostic signals.

A.4.3 Correlation among Discriminative Signals

To quantify redundancy among the three discriminative signals, we compute pairwise Spearman rank correlation coefficients over the candidate pool $\mathcal{C}$. By capturing monotonic dependencies, this metric serves as an effective proxy for signal overlap. To ensure objectivity, this analysis is conducted prior to sample selection, preventing distributional shifts induced by the selection process from distorting the underlying correlation structure among signals. Figure 6 illustrates the rank correlation between the three categories of discriminative signals. Accuracy-gap magnitude $|\Delta|$ and cross-model inconsistency H have a moderate-to-strong association ($\rho = 0.671$), whereas their correlations with prediction instability S are low. Thus, $|\Delta|$ and H share a meaningful difficulty component but are not interchangeable, while S contributes a more distinct signal. The composite score retains all three to cover presentation sensitivity, disagreement, and perturbation instability rather than assuming that they are mutually independent.

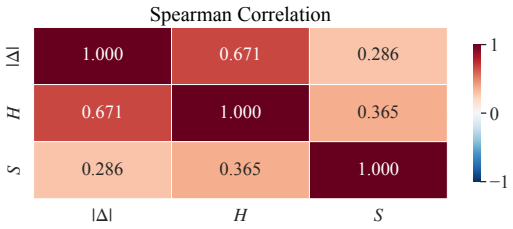


Figure 6: Spearman correlation matrix of discriminative signals on the candidate sample pool $\mathcal{C}$. The heatmap visualizes the pairwise rank correlations among $|\Delta|$, H, and S.

A.4.4 Robustness to Judge Model

To assess the robustness of our evaluation protocol, we re-evaluate Qwen3-VL-30B-A3B-Instruct on 500 randomly sampled examples from the PiP benchmark using three independent semantic-equivalence judges—GPT5_0807, Gemini-2.5-Pro, and Doubao-Seed—and five human experts. All

Table 6: Judge-choice stability for the domain-level accuracy of the fixed answer model Qwen3-VL-30B-A3B-Instruct on 500 PiP examples. All judges score the same predictions using the same semantic-equivalence criterion.

Judge	Mathematics ACC	STEM ACC	Vision-Dominant ACC
GPT5_0807	46.18	44.92	48.02
Gemini-2.5-pro-03-25	46.34	46.05	47.56
DoubaoSeed16_0615	46.48	46.59	46.65
Human Expert	46.92	44.96	48.11

Table 7: Sampling sensitivity of the PiP benchmark. Filtered PiP denotes the behavior-selected diagnostic set, Random PiP denotes stratified random subsets with the same size, and Full PiP denotes the full candidate pool. We report original-input accuracy, PiP accuracy, and the corresponding PiP Gap. Degradation on Random and Full PiP tests whether the gap exists before behavior-based filtering.

Model	Filtered PiP			Random PiP			Full PiP		
	Ori	PiP	Gap	Ori	PiP	Gap	Ori	PiP	Gap
DoubaoSeed1.6	62.00	45.47	16.53	69.20±0.82	60.10±0.92	9.10±0.61	70.87	63.45	7.42
GPT5	63.91	49.33	14.58	73.30±0.66	65.25±0.86	8.05±0.64	74.12	66.92	7.20
Gemini-2.5-Pro	73.72	58.40	15.32	76.05±0.59	70.85±0.73	5.20±0.46	76.40	71.33	5.07

evaluators apply the same answer-equivalence criteria. The human row reports the average across experts. Table 6 reports domain-level ACC because judge choice affects correctness scoring, whereas the SR/CR labels are fixed before evaluation and PR is computed directly from each clean/perturbed prediction pair. As shown in Table 6, domain-level ACC varies by at most 1.67 points across the model judges and human evaluation. With the answer model fixed, this experiment directly measures judge-choice stability and shows that PiP accuracy is comparatively stable across the evaluated model and human judges.

A.4.5 Sensitivity to Behavior-Based Selection

To assess how behavior-based selection changes measured difficulty, we evaluate the PiP Gap under three sampling settings: Filtered PiP, Random PiP, and Full PiP. Filtered PiP denotes the diagnostic set selected by the composite behavior score described in Section 2.1.3. Random PiP is constructed by stratified random sampling from the same candidate pool, with the same size as Filtered PiP and matched proportions over source datasets, question categories, and SR/CR types. Full PiP contains all candidate samples before behavior-based filtering. For Random PiP, we use five fixed stratified sampling seeds and report the mean and sample standard deviation.

Table 7 shows that the PiP Gap consistently appears across all sampling protocols. Filtered PiP yields the largest drop, which is expected because it is designed as a diagnostic stress test that concentrates samples with high input-format sensitivity, cross-model disagreement, and perturbation instability. However, clear degradation also remains on Random PiP and Full PiP, where samples are not selected by the behavior score. Thus, the PiP Gap occurs before behavior-based selection, while filtering concentrates naturally occurring cases with strong presentation sensitivity. Filtered PiP serves as the diagnostic stress test rather than an average-case estimate, whereas Random and Full PiP characterize the broader candidate distribution. Because the models in this control contributed to the selection score, their Filtered-PiP values are interpreted as in-sample diagnostics. Generality is assessed separately on the unmarked held-out checkpoints in Table 1.

A.5 Main Results and Diagnostic Analysis

We evaluate 13 model checkpoints, as summarized in Figure 7. Our analysis yields three primary findings:

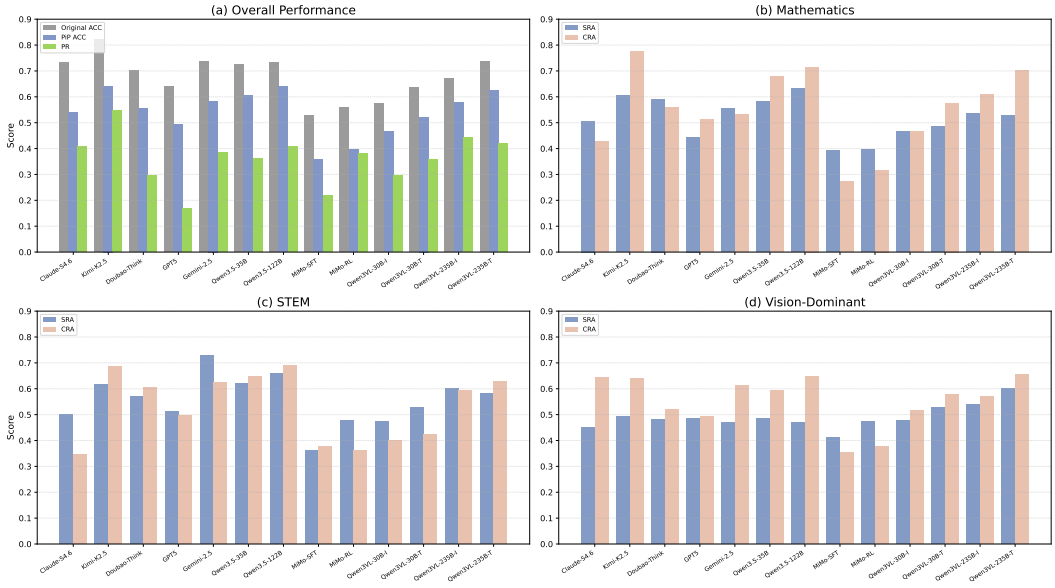


Figure 7: Main PiP-All results for the 13 evaluated checkpoints. (a) Original-Interleaved ACC, PiP-All ACC, and PR. (b)–(d) SRA and CRA across Mathematics, STEM, and Vision-Dominant questions.

Performance degradation under PiP-All. Figure 7(a) shows a consistent drop from Original-Interleaved to PiP-All accuracy, averaging 13.8 percentage points. Claude-Sonnet-4.6 and Kimi-K2.5 drop by 19.30 and 18.04 points, respectively. Together with the PiP-Visual control, this indicates that current models remain sensitive to dense unified-canvas presentation beyond query rasterization alone.

Domain-specific reasoning bottlenecks. Figures 7(b)–7(d) compare mathematics, STEM, and vision tasks. Kimi-K2.5 and the Qwen3.5 models are strongest on mathematical and STEM composition, while the relative SRA/CRA profiles vary across model families and domains. These differences show that local reference resolution and cross-region composition form distinct bottlenecks rather than one uniformly dominant failure stage.

Low perturbation consistency. PR ranges from 16.76% for GPT5_0807 to 54.85% for Kimi-K2.5 and remains substantially below Original-Interleaved accuracy for most models. RL- and thinking-tuned models such as MiMo-VL-RL and Qwen3-Thinking generally improve PR and CRA, partially mitigating this instability.

A.6 PiP Benchmark Example

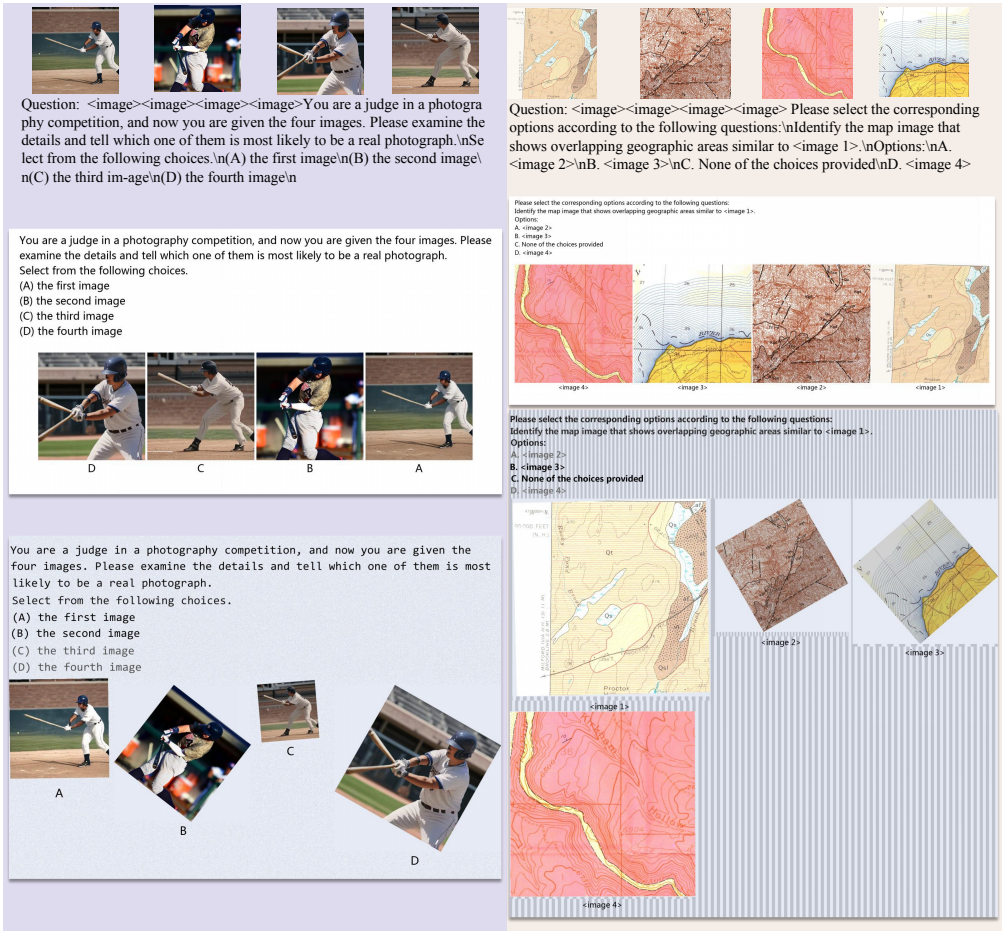


Figure 8: PiP Benchmark Example

B Information-Geometric Online Alignment

B.1 Log-Volume Interpretation of Incremental Information Gain

We provide a concise interpretation of the incremental information gain used in IGOA. Let $A \succeq 0$ denote the accumulated Fisher memory and let $F \succeq 0$ denote the active Fisher information contributed by a context at the current step. With a damping coefficient $\epsilon > 0$, IGOA defines

$$\Delta\mathcal{I}(F; A) = \log\det(A + F + \epsilon I) - \log\det(A + \epsilon I). \tag{13}$$

The following proposition shows that this quantity measures the log-volume expansion of the current active information after normalizing by the accumulated information geometry.

Proposition A.1. For any $A \succeq 0$, $F \succeq 0$, and $\epsilon > 0$,

$$\Delta\mathcal{I}(F; A) = \log\det\left(I + (A + \epsilon I)^{-\frac{1}{2}} F (A + \epsilon I)^{-\frac{1}{2}}\right). \tag{14}$$

Equivalently, if $\{\mu_k\}_{k=1}^{p}$ are the eigenvalues of

$$\widetilde{F} = (A + \epsilon I)^{-\frac{1}{2}} F (A + \epsilon I)^{-\frac{1}{2}}, \tag{15}$$

then

$$\Delta\mathcal{I}(F; A) = \sum_{k=1}^{p} \log(1 + \mu_k). \tag{16}$$

Proof. Since $A + \epsilon I$ is positive definite, we can write

$$A + F + \epsilon I = (A + \epsilon I)^{\frac{1}{2}} \left[I + (A + \epsilon I)^{-\frac{1}{2}} F (A + \epsilon I)^{-\frac{1}{2}}\right] (A + \epsilon I)^{\frac{1}{2}}. \tag{17}$$

Taking determinants gives

$$\det(A + F + \epsilon I) = \det(A + \epsilon I)\det\left(I + (A + \epsilon I)^{-\frac{1}{2}} F (A + \epsilon I)^{-\frac{1}{2}}\right). \tag{18}$$

Taking logarithms and subtracting $\log\det(A + \epsilon I)$ yields Eq. equation 14. Eq. equation 16 follows from the eigenvalue decomposition of the positive semi-definite matrix $\widetilde{F}$. □

Interpretation. Proposition A.1 shows that $\Delta\mathcal{I}(F; A)$ is not merely the magnitude of the current Fisher information. Instead, it measures log-volume expansion after normalizing the active Fisher matrix by accumulated Fisher memory $A + \epsilon I$. In these normalized coordinates, information aligned with directions already strongly represented in A contributes less relative gain, whereas information along weakly covered directions produces larger expansion. Therefore, the score favors context trajectories that introduce non-redundant policy-relevant information rather than repeatedly reinforcing directions already captured by past updates.

Diagonal approximation. For scalable MLLM training, we use a diagonal approximation. Let

$$A = \mathrm{diag}(a_1, \ldots, a_p), \quad F = \mathrm{diag}(f_1, \ldots, f_p), \tag{19}$$

where $a_k \geq 0$ and $f_k \geq 0$. Then Eq. equation 13 becomes

$$\Delta\mathcal{I}(F; A) = \sum_{k=1}^{p} \log\left(1 + \frac{f_k}{a_k + \epsilon}\right). \tag{20}$$

This form preserves the same relative-gain interpretation in a coordinate-wise manner. Unlike a trace score $\sum_k f_k$, which only measures the instantaneous Fisher magnitude, Eq. equation 20 evaluates each active direction relative to the accumulated memory a_k. As a result, IGOA can distinguish newly informative directions from directions that are large but redundant with previous updates.

B.2 Online Mirror Descent Update

In this section, we derive the online update of the trajectory coordination variable used in IGOA. Let $\lambda_t \in \Delta$, where $\Delta = \{\lambda \in \mathbb{R}_+^D : \sum_{d=1}^D \lambda_d = 1\}$ denotes the probability simplex over context spaces. We initialize $\lambda_{d,0} = 1/D$ for all d, which guarantees $\lambda_{d,t} > 0$ throughout training and makes the KL divergence well-defined.

Given the incremental information gain vector $\Delta\mathcal{I}_t = [\Delta\mathcal{I}_{1,t}, \ldots, \Delta\mathcal{I}_{D,t}]$, IGOA updates λ_{t+1} using the mirror-ascent form of online mirror descent with the negative entropy mirror map:

$$\lambda_{t+1} = \arg\max_{\lambda \in \Delta} \left\{ \langle \lambda, \Delta\mathcal{I}_t \rangle - \frac{1}{\eta_t} D_{\mathrm{KL}}(\lambda \| \lambda_t) \right\}, \tag{21}$$

where

$$D_{\mathrm{KL}}(\lambda \| \lambda_t) = \sum_{d=1}^D \lambda_d \log \frac{\lambda_d}{\lambda_{d,t}}. \tag{22}$$

The linear term favors context trajectories with larger non-redundant information gain, while the KL term constrains the coordination variable to remain close to its previous state.

Closed-form solution. Introducing a Lagrange multiplier μ for the simplex constraint, the Lagrangian is:

$$\mathcal{L}(\lambda, \mu) = \sum_{d=1}^D \lambda_d \Delta\mathcal{I}_{d,t} - \frac{1}{\eta_t} \sum_{d=1}^D \lambda_d \log \frac{\lambda_d}{\lambda_{d,t}} + \mu \left(\sum_{d=1}^D \lambda_d - 1 \right). \tag{23}$$

Taking the derivative with respect to λ_d and setting it to zero gives:

$$\Delta\mathcal{I}_{d,t} - \frac{1}{\eta_t} \left(\log \frac{\lambda_d}{\lambda_{d,t}} + 1 \right) + \mu = 0. \tag{24}$$

Therefore,

$$\lambda_d = \lambda_{d,t} \exp\left(\eta_t \Delta\mathcal{I}_{d,t}\right) \exp\left(\eta_t \mu - 1\right). \tag{25}$$

The factor $\exp(\eta_t \mu - 1)$ is shared by all context spaces and is determined by the normalization constraint. Hence,

$$\lambda_{d,t+1} = \frac{\lambda_{d,t} \exp\left(\eta_t \Delta\mathcal{I}_{d,t}\right)}{\sum_{d'=1}^D \lambda_{d',t} \exp\left(\eta_t \Delta\mathcal{I}_{d',t}\right)}. \tag{26}$$

This update is an exponentiated-gradient recursion on the simplex. Unlike a memoryless softmax rule, it preserves the previous coordination state through the multiplicative prior $\lambda_{d,t}$ and therefore induces a smooth online trajectory over context spaces.

Adaptive step size. We use the following adaptive schedule:

$$\eta_t = \frac{\sqrt{2 \log D}}{\sqrt{\sum_{\tau=0}^t \|\Delta\mathcal{I}_\tau\|_\infty^2 + \epsilon}}. \tag{27}$$

This schedule reduces the update magnitude when the cumulative scale of the information-gain signal becomes large, preventing overly aggressive changes in the coordination variable.

Standard OMD property. Let $g_t = \Delta\mathcal{I}_t$ and assume $\|g_t\|_\infty < \infty$. For any fixed comparator $\lambda^\star \in \Delta$, the entropy-regularized OMD update satisfies the standard regret bound:

$$\sum_{t=0}^{T-1} \langle \lambda^\star - \lambda_t, g_t \rangle \leq \mathcal{O}\left(\sqrt{\log D \sum_{t=0}^{T-1} \|g_t\|_\infty^2} \right). \tag{28}$$

Thus, the cumulative information gain achieved by the online coordination sequence approaches that of the best fixed coordination vector in hindsight up to a sublinear regret term.

Implication for IGOA. The OMD update provides a principled online mechanism for evolving the context coordination weights. It balances the previous coordination state, the current non-redundant information gain, and the adaptive step size. The bound concerns the realized constructed information-gain sequence. Task-level effectiveness is evaluated empirically through the controlled coordination ablations.

B.3 Diagonal Approximation and Scalability

This section shows how the information-geometric score in IGOA can be computed efficiently with a diagonal active Fisher approximation. The approximation is not intended to exactly recover the full Fisher geometry, since cross-parameter correlations are discarded. Instead, it provides a scalable coordinate-wise surrogate that preserves the monotonicity and diminishing-return structure needed for online context coordination.

The full active Fisher matrix is

$$F_{d,t}^{\mathrm{active}} = \mathbb{E}_{q\sim\mathcal{B}_{d,t},\, o\sim\pi_{\theta_{\mathrm{old}}}}\left[\mathbf{1}[0 < c(q) < G]\, \bar{s}_{\theta_{\mathrm{old}}}(o|q)\bar{s}_{\theta_{\mathrm{old}}}(o|q)^{\top}\right], \tag{29}$$

where

$$\bar{\ell}(o|q) = \frac{1}{L_o}\sum_{\ell=1}^{L_o}\log \pi_{\theta_{\mathrm{old}}}(o_\ell|q, o_{<\ell}), \qquad \bar{s}_{\theta_{\mathrm{old}}}(o|q) = \nabla_{\theta_{\mathrm{old}}}\bar{\ell}(o|q). \tag{30}$$

The mean log-likelihood makes the statistic invariant to a uniform replication of response tokens and prevents longer responses from receiving systematically larger Fisher magnitude. Directly forming the desired diagonal,

$$f_{d,t} = \frac{1}{N_{d,t}}\sum_{j\in\mathcal{A}_{d,t}} \bar{s}_j \odot \bar{s}_j, \tag{31}$$

would require materializing or repeatedly backpropagating $N_{d,t}$ per-response gradients. Squaring the ordinary batch gradient is not equivalent because it introduces cross-response terms. We instead draw independent signs with $\Pr(\xi_j = 1) = \Pr(\xi_j = -1) = 1/2$ and compute

$$v_{d,t} = \nabla_{\theta_{\mathrm{old}}}\left(\frac{1}{\sqrt{N_{d,t}}}\sum_{j\in\mathcal{A}_{d,t}} \xi_j\bar{\ell}_j\right) = \frac{1}{\sqrt{N_{d,t}}}\sum_{j\in\mathcal{A}_{d,t}} \xi_j\bar{s}_j, \tag{32}$$

$$\widehat{f}_{d,t} = v_{d,t}\odot v_{d,t}. \tag{33}$$

Because $\mathbb{E}[\xi_i\xi_j] = \mathbf{1}[i = j]$,

$$\mathbb{E}_\xi[\widehat{f}_{d,t} \mid \mathcal{A}_{d,t}] = \frac{1}{N_{d,t}}\sum_{j\in\mathcal{A}_{d,t}} \bar{s}_j \odot \bar{s}_j = f_{d,t}. \tag{34}$$

Thus, one reverse-mode VJP provides an unbiased stochastic estimate of the per-response squared-score average without `vmap`, repeated per-response backward passes, or an aggregate-gradient-square bias. This unbiasedness applies to $\widehat{f}_{d,t}$, not to its nonlinear log-gain. For fixed memory a_t, concavity gives

$$\mathbb{E}_\xi\left[\widehat{\Delta\mathcal{I}}_{d,t} \mid \mathcal{A}_{d,t}, a_t\right] \le \sum_k \log\left(1 + \frac{f_{d,t,k}}{a_{t,k} + \epsilon}\right). \tag{35}$$

IGOA therefore uses $\widehat{\Delta\mathcal{I}}_{d,t}$ as a realized stochastic coordination signal rather than an unbiased estimator of the exact log-volume gain. We use the same probe count for both contexts and draw one independent probe per context at each controller update. The EMA memory smooths probe noise over updates. We maintain the accumulated memory as

$$A_{t+1} \approx \mathrm{Diag}(a_{t+1}), \quad a_{t+1} = \alpha a_t + \sum_d \lambda_{d,t}\widehat{f}_{d,t}. \tag{36}$$

For the training-dynamics analysis in Figure 3, we report the cosine overlap between the two context-specific diagonal Fisher vectors,

$$\mathcal{R}_t = \frac{\langle \widehat{f}_{M,t}, \widehat{f}_{S,t} \rangle}{\|\widehat{f}_{M,t}\|_2 \|\widehat{f}_{S,t}\|_2 + \epsilon}. \tag{37}$$

This descriptive statistic lies in $[0, 1]$ because the entries are non-negative. A smaller value indicates less coordinate-wise overlap between the active sensitivities induced by the two context spaces.

Under this approximation, the incremental information gain becomes

$$\Delta\mathcal{I}_{d,t} = \log\det\left(A_t + F_{d,t}^{\text{active}} + \epsilon I\right) - \log\det\left(A_t + \epsilon I\right) \tag{38}$$

$$\approx \sum_{k=1}^{P} \log\left(1 + \frac{\widehat{f}_{d,t,k}}{a_{t,k} + \epsilon}\right). \tag{39}$$

Thus, each coordinate contributes according to its newly observed active Fisher magnitude normalized by the information already accumulated along that coordinate. Directions that have been repeatedly covered receive smaller marginal gain, while under-explored but policy-sensitive directions receive larger gain.

This surrogate preserves the desired structural properties. For each coordinate, let

$$g_k(f, a) = \log\left(1 + \frac{f}{a + \epsilon}\right), \quad f \geq 0,\ a \geq 0,\ \epsilon > 0. \tag{40}$$

Then

$$\frac{\partial g_k}{\partial f} = \frac{1}{a + \epsilon + f} > 0, \quad \frac{\partial g_k}{\partial a} = -\frac{f}{(a + \epsilon)(a + \epsilon + f)} \leq 0, \tag{41}$$

and

$$\frac{\partial^2 g_k}{\partial f^2} = -\frac{1}{(a + \epsilon + f)^2} < 0. \tag{42}$$

Therefore, the diagonal score increases with newly observed active Fisher information, decreases with accumulated memory, and exhibits diminishing returns with respect to additional Fisher mass. These properties are sufficient for IGOA's online coordination rule, which only requires a stable scalar score for comparing the non-redundant contributions of different context spaces.

The full Fisher representation requires $\mathcal{O}(P^2)$ storage and expensive log-determinant computation. In contrast, the diagonal approximation persistently maintains only a_t, requiring $\mathcal{O}(P)$ storage and $\mathcal{O}(P)$ vector operations per context space in addition to the VJP. All diagonal statistics are computed in each backbone's fixed native parameterization. The method uses their relative-gain and diminishing-return structure rather than claiming reparameterization invariance.

B.4 Pseudocode for IGOA

Pseudocode 1 summarizes the main training procedure of IGOA. The controller is updated every K optimizer steps, and the resulting coordination weights are used during the next interval.

B.5 Implementation Details

We optimize the model using AdamW with a learning rate of 1.0×10^{-6}, a weight decay of 0.01, and train for 30 epochs. For the reward formulation, we use answer correctness and format validity with $\alpha_r = \beta_r = 0.5$ to avoid task-specific reward engineering. The rollout group size is set to $G = 8$, with sampling temperature $T = 1.0$ and $top_p = 1.0$. For IGOA, the coordination vector is initialized uniformly, with $\lambda_{M,0} = \lambda_{S,0} = 0.5$. AFIM is estimated only from active rollout groups satisfying $0 < c(q) < G$. We use one independent Rademacher probe per context and update the controller every $K = 16$ optimizer steps. The coordination vector λ remains fixed between controller updates. Each probe differentiates a signed sum of mean response log-likelihoods through a single reverse-mode VJP. After each VJP, we square its local sharded gradients into $\widehat{f}_{d,t}$ and release them

Algorithm 1 IGOA: Information-Geometric Online Alignment

Require: Domains D_M, D_S; policies $\pi_\theta, \pi_{\text{old}}$; surrogate granularity g; group size G; controller interval K; decay α; regularizer ϵ; $\lambda_{M,0} = \lambda_{S,0} = 1/2$; $a_0 = \mathbf{0}$.

Set controller index $t \leftarrow 0$.
for each optimizer step $n = 1, 2, \ldots$ **do**
  for each context space $d \in \{M, S\}$ **do**
    Sample prompts, generate G responses per prompt, compute shaped policy rewards and verifiable correctness rewards, and retain groups satisfying $0 < c(q) < G$ under correctness.
    Compute group-relative advantages and collect the retained responses in $\mathcal{A}_{d,n}$.
  end for
  if $n \bmod K = 0$ **then**
    for each context space $d \in \{M, S\}$ **do**
      Draw independent $\xi_j \in \{-1, +1\}$ and compute one VJP $v_{d,t} \leftarrow \nabla_{\theta_{\text{old}}}\big(|\mathcal{A}_{d,n}|^{-1/2} \sum_j \xi_j \bar{\ell}_j\big)$.
      Set $\widehat{f}_{d,t} \leftarrow v_{d,t} \odot v_{d,t}$.
      Compute $\Delta\mathcal{I}_{d,t} \leftarrow \sum_{k=1}^{P} \log\big(1 + \widehat{f}_{d,t,k}/(a_{t,k} + \epsilon)\big)$.
    end for
    Set $\eta_t \leftarrow \sqrt{2\log 2}/\sqrt{\sum_{\tau=0}^{t} \|\Delta\mathcal{I}_\tau\|_\infty^2 + \epsilon}$.
    Set $\lambda_{d,t+1} \leftarrow \frac{\lambda_{d,t} \exp(\eta_t \Delta\mathcal{I}_{d,t})}{\sum_{d'} \lambda_{d',t} \exp(\eta_t \Delta\mathcal{I}_{d',t})}$.
    Set $a_{t+1} \leftarrow \alpha a_t + \sum_d \lambda_{d,t} \widehat{f}_{d,t}$.
  end if
  Optimize θ on retained prompts with the clipped RLVR surrogate at granularity g, weighted by λ_t.
  Update the rollout policy $\pi_{\text{old}} \leftarrow \pi_\theta$ after the optimizer step.
  if $n \bmod K = 0$ **then**
    Set $t \leftarrow t + 1$.
  end if
end for

before the next context, so the implementation neither materializes per-response gradients nor squares the ordinary batch gradient. The accumulated Fisher memory is updated with EMA decay $\alpha = 0.99$, and we set $\epsilon = 10^{-8}$. The Fisher state covers trainable parameters only. Explicit KL regularization is disabled across all experiments. Models with up to 8B parameters are trained on 16 NVIDIA GPUs, while the 30B-A3B model is trained on 32 NVIDIA GPUs. Before training, we apply the same exact-ID, normalized question–answer, and source-image overlap audit between MMPR-Tiny and every evaluation source. Because BLINK, MuirBench, MathVerse, MMMU-Pro, and MMStar are all contained in the audited PiP candidate pool, the five conventional evaluation sets are covered individually. Accordingly, no audited match remains in the retained training/evaluation split.

B.6 Training-Seed Robustness

We assess sensitivity to stochastic RLVR training using Qwen3-VL-8B-Instruct in the token-level setting. We run DAPO_mix and IGOA_token with three matched random seeds while keeping the training data, initialization checkpoint, number of optimizer steps, rollout budget, rollout group size, sampling configuration, reward design, and evaluation protocol fixed. Table 8 reports the mean and standard deviation across the three runs. In addition to the absolute scores, we report the seed-matched improvement of IGOA over DAPO. This distinguishes a consistent method-level gain from variation induced by a single RLVR trajectory.

Across the three matched seeds, IGOA_token improves PiP-All ACC from 46.73 ± 0.75 to 54.23 ± 0.54, corresponding to a seed-matched gain of 7.50 ± 0.63 points. Using the two-sided t interval $\bar{\Delta} \pm t_{0.975,2} s_\Delta / \sqrt{3}$, the 95% confidence intervals for the mean matched gains are $[5.93, 9.07]$ on PiP-All ACC, $[3.85, 6.79]$ on PR, and $[3.05, 4.73]$ on General AVG. Thus, the positive gains persist across the evaluated seeds rather than being driven by one favorable RLVR trajectory. The General

Table 8: Training-seed robustness on Qwen3-VL-8B-Instruct with the token-level RLVR surrogate. Results are mean ± sample standard deviation over three matched random seeds. Δ denotes the mean and standard deviation of the seed-matched IGOA_token improvements over DAPO_mix. All methods use identical data, rollout, optimization, reward, and evaluation budgets.

Method	PiP ACC	SRA	CRA	PR	General AVG
DAPO_mix	46.73 ± 0.75	49.35 ± 0.92	44.49 ± 0.58	45.63 ± 0.84	64.13 ± 0.42
IGOA_token	54.23 ± 0.54	57.90 ± 0.93	51.01 ± 0.39	50.95 ± 0.92	68.02 ± 0.32
Δ (IGOA $-$ DAPO)	$+7.50\pm0.63$	$+8.55\pm0.72$	$+6.52\pm0.69$	$+5.32\pm0.59$	$+3.89\pm0.34$

Table 9: Paired post-training alignment and allocation-matched control on Qwen3-VL-8B-Instruct with the token-level RLVR surrogate. Original and PiP-All are evaluated on the same examples. Gap denotes Original ACC minus PiP ACC. Static-$\bar{\lambda}$ fixes the Unified-Canvas objective weight to IGOA's training-time average under matched data, rollouts, optimizer steps, rewards, initialization, and random seed.

Method	$\bar{\lambda}_S$	Original ACC	PiP ACC	Gap	PR	General AVG
Base checkpoint	–	56.73	43.27	13.46	35.86	61.90
DAPO_mix	0.50	58.40	46.87	11.53	45.83	64.19
Static-$\bar{\lambda}$	0.49	58.27	47.31	10.96	45.62	64.38
IGOA_token	0.49	59.13	54.79	4.34	51.77	68.34

AVG follows the same arithmetic mean over the five general multimodal benchmarks used in the main results.

B.7 Paired Alignment and Allocation-Matched Control

We directly evaluate post-training Original-Interleaved and PiP-All behavior on the same 1.5K paired examples using identical decoding and judging protocols. Static-$\bar{\lambda}$ is initialized from the same Qwen3-VL-8B-Instruct checkpoint as IGOA_token and uses IGOA's measured time-averaged context weight throughout training. All remaining training and evaluation settings are matched.

The Original and PiP-All columns jointly measure whether unified-canvas gains preserve behavior under the original presentation. Relative to DAPO_mix, IGOA_token raises Original ACC by 0.73 points and PiP ACC by 7.92 points, reducing the paired presentation gap by 7.19 points. Static-$\bar{\lambda}$ matches IGOA's average context allocation: despite nearly identical Original ACC, online IGOA improves PiP ACC, PR, and General AVG by 7.48, 6.15, and 3.96 points, respectively. Thus, the result is not explained by replacing the uniform mixture with IGOA's average allocation.

B.8 Main Results on PiP Benchmark

We present the main results on the PiP benchmark to evaluate the fine-grained reasoning and perception capabilities of MLLMs under diverse tasks. These experiments involve multiple backbone architectures including the Qwen series and MiMo VL to verify the generalizability of the IGOA framework.

As shown in Table 10, IGOA variants consistently outperform the DAPO and GSPO baselines across the evaluated backbones. On MiMo-VL, IGOA_token reaches 57.93% Mathematics SRA, compared with 47.64% for DAPO_mix and 48.54% for GSPO_mix. On Qwen3-VL-8B, it reaches 62.57% STEM SRA, compared with 54.89% for DAPO_mix and 54.04% for GSPO_mix. The corresponding CRA and PR results provide complementary evidence across model families and scales. The headline PiP-All ACC results are reported in Table 2.

Table 10: Domain-conditioned PiP-All diagnostics across MLLM architectures. Each domain reports SRA, CRA, and Perturbation Repeatability (PR). Headline PiP-All accuracy is reported in Table 2. **Bold** and underline denote the best and second-best results.

Method	Mathematics			STEM			Vision-Dominant		
	SRA	CRA	PR	SRA	CRA	PR	SRA	CRA	PR
Qwen2.5-VL-7B-Instruct	23.40	16.83	16.67	31.06	19.62	12.80	35.91	36.51	11.05
VL-Rethinker-Qwen2.5	37.17	20.71	31.85	29.36	**37.36**	19.64	36.29	31.95	18.71
NoisyRollout-Qwen2.5	31.41	17.80	24.35	31.49	33.96	14.02	37.45	34.02	24.02
DAPO_mix	32.98	18.12	23.53	34.89	27.17	**27.92**	44.40	37.76	24.76
GSPO_mix	33.51	21.68	31.30	30.21	35.09	27.44	40.93	35.27	22.51
IGOA_token	41.18	22.96	**39.73**	**38.37**	32.26	26.39	**46.84**	37.46	**32.15**
IGOA_seq	**41.31**	**23.54**	39.19	38.21	37.28	26.30	46.72	**39.87**	31.44
MiMo-VL-7B-SFT-2508	39.27	27.18	28.93	36.17	37.74	22.16	41.31	35.27	14.06
VL-Rethinker-MiMo	48.49	28.24	38.40	40.52	46.15	31.63	44.80	36.52	28.77
NoisyRollout-MiMo	49.90	29.72	41.81	44.14	41.36	33.09	45.99	37.88	29.39
DAPO_mix	47.64	32.36	46.60	44.68	44.15	36.49	45.56	43.98	33.04
GSPO_mix	48.54	34.35	46.60	46.81	43.77	33.19	45.56	43.90	32.01
IGOA_token	**57.93**	35.69	**48.58**	**52.88**	47.64	36.82	47.81	**46.92**	36.96
IGOA_seq	52.16	**35.84**	47.51	50.04	**52.94**	**43.31**	**49.30**	46.89	**40.24**
Qwen3-VL-8B-Instruct	40.84	32.36	37.64	51.91	42.26	36.32	48.65	46.06	34.60
DAPO_mix	45.03	40.45	42.65	54.89	45.66	**53.60**	47.88	48.96	42.15
GSPO_mix	50.26	43.69	41.13	54.04	47.55	45.85	52.51	52.70	43.35
IGOA_token	**53.99**	45.07	**51.94**	**62.57**	**53.24**	53.08	**58.82**	**57.66**	50.78
IGOA_seq	51.74	**47.87**	43.70	61.16	51.93	52.75	56.64	56.93	**53.06**
DAPO_mix	50.83	45.96	35.73	47.51	50.19	32.58	47.05	45.23	39.61
GSPO_mix	46.60	53.72	29.02	50.91	43.77	30.25	47.10	45.23	29.00
IGOA_token	51.91	57.48	36.77	**52.27**	**53.74**	**55.06**	52.11	**52.35**	**48.84**
IGOA_seq	**56.09**	**60.21**	**40.33**	51.10	52.92	40.20	**55.79**	46.07	41.26

Table 11: Component and coordination-signal ablation on Qwen3-VL-30B-A3B-Instruct using the token-level RLVR surrogate. All variants use the same training data and total rollout budget, equal per-context rollout counts before weighting, rollout group size $G = 8$, sampling configuration, reward design, and RLVR surrogate. Fixed 1:1 Mix keeps $\lambda_M = \lambda_S = 0.5$ throughout training. Best results are in bold and second-best results are underlined.

Method	Coordination Signal	PiP ACC	SRA	CRA	PR	AVG
Fixed 1:1 Mix	Fixed $\lambda_M = \lambda_S = 0.5$	47.95	48.32	47.58	25.29	63.21
Reward-Entropy Weighting	$\lambda_{d,t} \propto \exp(H_{d,t}/\tau)$	50.12	50.47	49.82	37.64	65.73
Fisher-Trace	$\mathrm{Tr}(F_{d,t}^{\mathrm{active}})$	51.36	51.82	50.97	43.85	66.47
w/o Memory	$\log\det(F_{d,t}^{\mathrm{active}} + \epsilon I)$	52.18	52.64	51.79	47.32	67.21
Full IGOA	$\Delta\mathcal{I}_{d,t}$	**53.14**	**53.92**	**52.48**	**50.71**	**68.06**

B.9 Ablation Study

B.9.1 Component and Coordination Signal Ablation

To isolate the coordination signal, we use the token-level setting as a representative controlled ablation and compare IGOA with four variants under the same data and total rollout budget, equal per-context rollout counts before weighting, group size $G = 8$, reward design, and RLVR surrogate. Fixed 1:1 Mix denotes a fixed equal-ratio baseline whose two context objectives retain $\lambda_M = \lambda_S = 0.5$ throughout training. Reward-Entropy Weighting assigns larger coordination mass to the context with higher empirical reward entropy, serving as a difficulty-driven dynamic weighting baseline. Fisher-Trace uses $\mathrm{Tr}(F_{d,t}^{\mathrm{active}})$, which captures active parameter sensitivity but does not account for whether the corresponding directions have already been covered. The w/o Memory variant removes the accumulated Fisher memory A_t and uses only current-step active Fisher statistics. Full IGOA

Table 12: Ablation on the active learning-frontier condition and rollout group size. All variants use the same training data and RLVR surrogate. Relative cost is normalized by the default setting $G = 8$. Best results are in bold and second-best results are underlined.

Setting	Active Filter	G	PiP ACC	SRA	CRA	PR	AVG	Rel. Cost
w/o Active Filter	✗	8	50.43	53.16	48.12	45.31	65.10	1.00×
Full IGOA	✓	4	<u>51.68</u>	**55.02**	<u>48.87</u>	<u>47.86</u>	<u>65.48</u>	0.56×
Full IGOA	✓	8	**53.14**	<u>53.92</u>	**52.48**	**50.71**	**68.06**	1.00×

Table 13: Theoretical persistent memory overhead of IGOA under full-parameter training. IGOA persistently maintains one diagonal accumulated Fisher memory A_t over all trainable parameters. Context-specific Rademacher VJPs are processed sequentially and released after updating the information-gain score and A_t. For MoE models, the overhead is computed with respect to the total trainable parameters rather than the activated parameters per token, since A_t is maintained for all trainable parameters. Per-worker memory is reported using the actual parameter-sharding settings in our training: 16 GPUs for Qwen3-VL-8B-Instruct and 32 GPUs for Qwen3-VL-30B-A3B-Instruct.

Model	Persistent State	Total FP32 Mem.	Total BF16 Mem.	FP32 / GPU, 16-way	FP32 / GPU, 32-way
Qwen3-VL-8B-Instruct	one A_t	29.80 GiB	14.90 GiB	1.86 GiB	-
Qwen3-VL-30B-A3B-Instruct	one A_t	111.76 GiB	55.88 GiB	-	3.49 GiB

denotes IGOA_token in this controlled table and uses the log-determinant incremental gain to measure marginal expansion relative to accumulated active Fisher information.

As shown in Table 11, dynamic coordination signals improve over fixed mixing, but remain weaker than Full IGOA. Reward-Entropy Weighting improves over Fixed 1:1 Mix, suggesting that domain difficulty is useful but insufficient. Fisher-Trace and w/o Memory further improve performance by using policy-sensitive Fisher signals, yet they still lag behind Full IGOA. All variants generate equal numbers of rollouts from the two contexts before objective weighting, so these differences cannot be attributed to additional PiP query exposure. Moreover, the ordering among the dynamic weighting rules shows that performance depends on the coordination signal rather than merely allowing nonuniform weights. In this controlled setting, accumulated memory provides additional gains beyond current-step active parameter sensitivity, consistent with discounting directions emphasized by earlier updates.

B.9.2 Active frontier and rollout group size

We further analyze the active learning-frontier condition and the rollout group size G. IGOA estimates active Fisher information only from rollout groups satisfying $0 < c(q) < G$, where $c(q)$ is the number of responses with positive verifiable correctness reward in the group. This condition excludes groups that are either already solved or completely failed, allowing the Fisher estimate to focus on samples with non-degenerate correctness signals. Table 12 shows that removing the active filter causes a clear performance drop, especially on PR, suggesting that Fisher estimates from all rollout groups are more easily dominated by redundant or unlearnable directions. For rollout size, $G = 4$ reduces computational cost but yields noisier frontier estimates and weaker performance than $G = 8$. We therefore use $G = 8$ as the default setting, which offers a better trade-off between estimation reliability and training efficiency.

B.9.3 Computational Efficiency Analysis

We analyze the computational overhead of IGOA under full-parameter RLVR training. A direct implementation of the active Fisher matrix is infeasible for MLLMs, since it requires $\mathcal{O}(P^2)$ storage and matrix operations, where P is the number of trainable parameters. IGOA avoids this cost by using a diagonal active Fisher approximation. Under this formulation, both the active Fisher estimate and the accumulated Fisher memory are represented as coordinate-wise second-order statistics, reducing the computation of the information-geometric score to vector-level operations.

IGOA persistently maintains only one diagonal accumulated Fisher memory A_t. At a controller update, each context requires one reverse-mode VJP. The implementation directly squares each VJP's sharded gradients to accumulate the scalar gain and memory update, then releases those gradients before processing the next context. A temporary pending accumulator of the same sharded size is used only during that controller update so both context gains are evaluated against the same A_t; it replaces A_t after both VJPs and is not checkpointed as an additional state. Therefore, IGOA does not store per-response gradients or persistent context-specific Fisher states. For a model with P trainable parameters, the persistent memory overhead is:

$$\mathrm{Mem}_{\mathrm{IGOA}} = Pb, \tag{43}$$

where b is the number of bytes per element. Under N-way parameter sharding, the per-worker overhead becomes:

$$\mathrm{Mem}_{\mathrm{IGOA}}^{\mathrm{worker}} \approx \frac{Pb}{N}. \tag{44}$$

Thus, IGOA replaces the infeasible $\mathcal{O}(P^2)$ Fisher representation with one shardable $\mathcal{O}(P)$ diagonal state. Table 13 shows that the persistent memory cost is non-negligible under full-parameter training, but remains linear and shardable. For Qwen3-VL-8B-Instruct, maintaining A_t in FP32 requires 29.80 GiB in total, corresponding to 1.86 GiB per worker under 16-way parameter sharding. For Qwen3-VL-30B-A3B-Instruct, the total FP32 overhead is 111.76 GiB, corresponding to 3.49 GiB per worker under 32-way sharding. The two VJPs are invoked once every 16 optimizer steps through the actor's standard forward/backward path. The remaining operations are element-wise accumulation, coordinate-wise log-volume scoring, and the EMA update. No inversion, eigendecomposition, dense log determinant, per-response backward loop, or additional rollout generation is required. Under the matched Qwen3-VL-8B-Instruct configuration, this implementation increases end-to-end step time by 6.8% over DAPO_mix across 100 post-warmup steps.