MV-PRM: Process Reward Model for Multi-View Visual Reasoning

Anonymous ACL submission

Abstract

Multi-view visual reasoning requires MLLMs to align and integrate cross-view information effectively. However, existing models heavily rely on outcome supervision and implicit aggregation, causing spurious alignment and unrectified inference-time errors. To bridge this gap, we leverage test-time scaling as an effective intervention and propose **MV-PRM**, the first process reward model tailored for multi-view visual reasoning. To formulate effective process rewards for complex spatial interactions, we introduce a systematic methodology that structures chain-of-thought (CoT) into sequential functional units (*Perceive, Align, Integrate, Conclude*) and enforces multi-dimensional step-level supervision across logical correctness and cross-view alignment. To realize this methodology, we develop a fully automated pipeline that employs a generator MLLM to synthesize unit-structured trajectories and a teacher MLLM for automated multi-dimensional annotation, curating 54k step-level samples from AirCopBench. Extensive experiments demonstrate that as a Best-of-N (BoN) reranker, MV-PRM yields substantial gains across diverse base MLLMs, outperforming general-purpose MLLMs serving as verifiers and offering a scalable paradigm for robust multi-view visual reasoning.

1 Introduction

Multi-view visual reasoning—the ability to align and integrate complementary visual information across diverse viewpoints—remains a fundamental challenge for MLLMs (Zhu et al., 2025; Xu et al., 2025; Huang et al., 2024; Ji et al., 2025; Yang et al., 2025; Wang et al., 2026) in collaborative perception and embodied AI. Applications such as multi-UAV systems (Hong et al., 2024; Lin et al., 2025) and autonomous driving (Zhang et al., 2024; Wang et al., 2025a) require models to perform structured cross-view perception, alignment, and progressive integration to facilitate reliable decision-making. Despite significant advances in high-level reasoning and planning (Zheng et al., 2025; Chang et al., 2025; Zhao et al., 2026), vanilla MLLMs frequently suffer from logical fragmentation and inconsistent reasoning across disparate views.

We attribute the inadequacy of existing MLLMs in multi-view reasoning to two interconnected limitations: **implicit cross-view aggregation** and **scarce process-level supervision**. First, under outcome supervision, models rely on black-box feature fusion rather than verified spatial associations. This precipitates *spurious alignment*—arriving at correct final answers through flawed intermediate perception or inconsistent integration, leaving inference-time errors unrectified. Second, the combinatorial complexity of multi-view scenarios creates an expansive reasoning path space, where delayed outcome feedback is too sparse to penalize early logical collapses, causing models to drift during complex inference. While test-time scaling via PRMs offers a promising solution, existing PRMs (She et al., 2025; Zhang et al., 2025b; Sun et al., 2025) are tailored for textual or mathematical domains, leaving them blind to the unique spatial constraints of multi-view visual reasoning.

To bridge this gap, **we propose MV-PRM, the first process reward model (PRM) for multi-view visual reasoning**. To formulate effective multi-dimensional process rewards, we first decompose complex multi-view reasoning into four sequential functional units: *Perceive* extracts localized single-view information, *Align* maps spatial correspondences across different viewpoints, *Integrate* establishes global perception, and *Conclude* deduces the final answer. For each unit, we rigorously evaluate its logical correctness and the quality of cross-view alignment, thereby providing explicit verification to effectively mitigate spurious alignment during test-time scaling. Driven by these multi-dimensional reward objectives, **MV-PRM** utilizes a multi-task dual-head architecture

optimized jointly for logical correctness and cross-view alignment quality. This joint optimization enables both heads to mutually regularize through shared representations, delivering dense feedback that suppresses early spatial drift and unlocks efficient test-time scaling.

To realize this paradigm, gathering step-level process supervision—especially trajectories tailored to our specialized functional units—via manual annotation is prohibitively labor-intensive and unscalable. To circumvent this bottleneck, **we develop a fully automated pipeline to construct the MV-PRM Dataset**—a pioneering asset comprising approximately 54k step-level samples derived from AirCopBench (Zha et al., 2026). This pipeline automates data curation by employing a generator MLLM to synthesize structured four-unit trajectories, which are subsequently evaluated by a powerful teacher MLLM to assign multi-dimensional, high-fidelity process rewards. Leveraging this fine-grained data, **MV-PRM** is efficiently trained via Low-Rank Adaptation (LoRA) (Hu et al., 2021), enabling it to provide reliable step-level evaluation for multi-view visual reasoning.

Our contributions are summarized as follows:

- We propose **MV-PRM**, the first process reward model tailored for multi-view visual reasoning. It incorporates a multi-dimensional process supervision mechanism that evaluates logical correctness and cross-view alignment quality, successfully unlocking reliable test-time scaling.

- We design a fully automated data construction pipeline to operationalize our systematic methodology. This pipeline employs a generator MLLM to synthesize reasoning trajectories structured into sequential functional units (*Perceive*, *Align*, *Integrate* and *Conclude*) and a teacher MLLM to generate parallel multi-dimensional annotations, curating a high-quality dataset of 54k step-level samples derived from AirCopBench.

- We conduct comprehensive experiments demonstrating that, as a BoN reranker, MV-PRM consistently yields substantial accuracy gains across diverse base MLLMs (notably boosting Qwen3-VL-8B-Instruct by **9.02%**). Furthermore, it significantly outperforms state-of-the-art general-purpose verifiers, proving the robust efficacy of our framework for test-time scaling.

2 Related Work

2.1 Multi-View Visual Reasoning

Multi-view visual reasoning requires MLLMs to synchronize and integrate information across diverse viewpoints for holistic scene comprehension. Recent benchmarks like MMIU (Meng et al., 2024), All-Angles (Yeh et al., 2026), and AirCopBench have exposed significant deficiencies in current MLLMs, particularly in spatial inference and cross-view object binding—tasks that demand far greater synergistic awareness than general multi-image understanding. To counter these challenges, recent efforts have enhanced the spatial comprehension of MLLMs via architectural innovations and large-scale dataset curation. LLaVA-3D (Zhu et al., 2024) leverages 3D-aware representations to facilitate object localization, while VLM4D (Zhou et al., 2025) integrates spatiotemporal cues to support complex tasks like trajectory planning in autonomous driving. Despite their progress in feature representation, these methods primarily rely on outcome-based supervision, treating multi-view integration as an unobservable "black box." Consequently, they remain prone to internal reasoning inconsistencies and hallucinations, as they lack an explicit cross-view alignment mechanism.

2.2 Process Reward Models

Reward models are essential for aligning MLLMs with supervisory signals. Early paradigms primarily relied on Outcome Reward Models (ORMs) (Cobbe et al., 2021; Wang et al., 2024a; Zhang et al., 2025a), which supervise only the final answer and often overlook flawed intermediate reasoning. To provide finer-grained supervision, PRMs (Lightman et al., 2023; Wang et al., 2025c,d; Cao et al., 2025; Li et al., 2026) evaluate intermediate reasoning steps, effectively capturing errors during the reasoning process. PRM800K (Lightman et al., 2023) is the first open-source process supervision dataset, containing step-level labels annotated by human experts on mathematical reasoning tasks. While manual annotation provides high-fidelity supervision signals, it is resource-intensive and limited in scale. To address this issue, Math-Shepherd (Wang et al., 2024b) introduced a Monte Carlo estimation framework for automatically constructing process supervision datasets without human

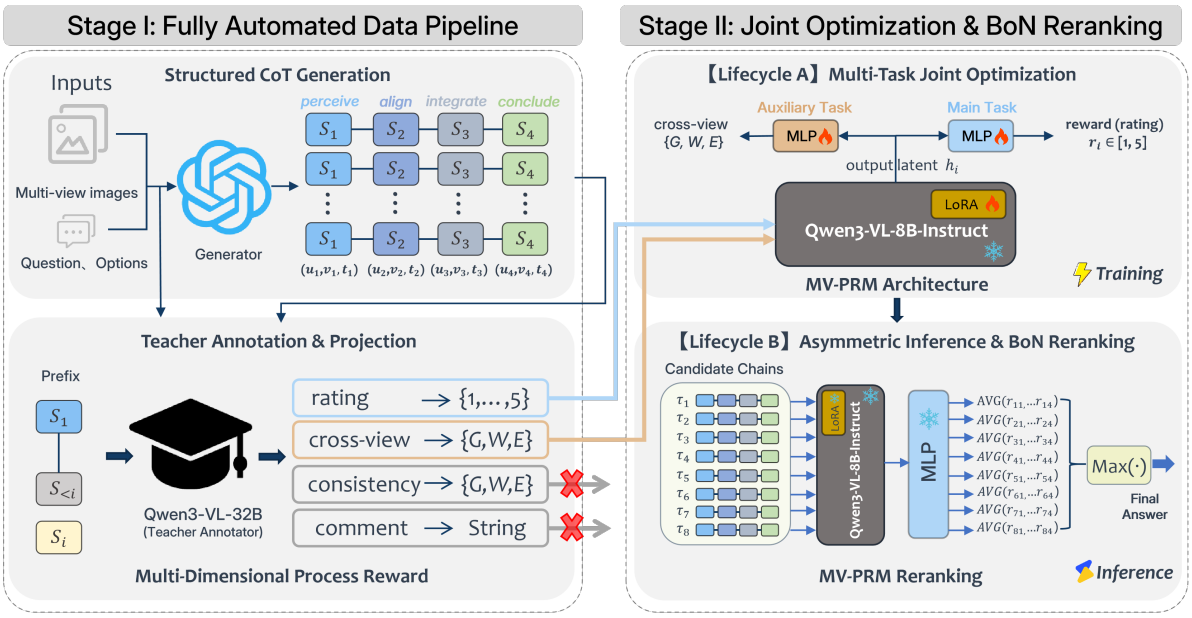

Figure 1: Overall framework of MV-PRM. Stage I automatically constructs a multi-dimensional process supervision dataset. Stage II performs multi-task joint optimization followed by asymmetric BoN reranking for final inference.

intervention. Recent research has extended process supervision to the multimodal domain. VisualPRM (Wang et al., 2025d) pioneered this shift with VisualPRM400K, followed by advanced variants like Athena-PRM (Wang et al., 2025b) and DreamPRM (Cao et al., 2025). However, existing multimodal PRMs are predominantly restricted to single-image inputs, with limited exploration of multi-view scenarios that involve cross-view reasoning.

3 Method

As illustrated in Figure 1, our framework systematically unfolds across two sequential stages: (1) **Stage I: Fully Automated Data Pipeline**, which structures raw inputs into sequential functional reasoning units and employs a powerful teacher MLLM to generate multi-dimensional annotations, which are selectively projected into targeted supervisory signals; and (2) **Stage II: Joint Optimization & BoN Reranking**, which trains the MV-PRM architecture via multi-task learning to simultaneously align reasoning logic and cross-view alignment. During inference, we adopt an asymmetric strategy, discarding the auxiliary alignment head while retaining the primary reward head to execute BoN reranking.

3.1 Problem Formulation

Multi-View Visual Reasoning Input Space. Let $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$ denote a set of n synchronized images capturing a scene from distinct viewpoints. Given a natural language question q and a finite set of candidate options $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, the objective is to select the correct option $a^* \in \mathcal{A}$ by leveraging cross-view visual cues.

Process-Level Trajectory. The MLLM solves the query by generating a CoT reasoning trajectory, formulated as an ordered sequence of intermediate logical steps $\tau = (S_1, S_2, \ldots, S_T)$. The auto-regressive joint probability of generating τ is defined as:

$$P(\tau \mid \mathcal{I}, q) = \prod_{i=1}^{T} P(S_i \mid S_{<i}, \mathcal{I}, q) \qquad (1)$$

The final prediction $a \in \mathcal{A}$ is derived via a deterministic extraction function $a = \text{Extract}(\tau)$.

Process Reward Model Objective. Outcome-based supervision exclusively maximizes the marginal probability $P(a^* \mid \mathcal{I}, q)$, leaving the model vulnerable to spurious alignment where a correct option masks an invalid latent cross-view alignment state $\mathbf{z}_{\text{align}}$:

$$P(\mathbf{z}_{\text{align}} = \text{invalid} \mid a = a^*, \mathcal{I}, q) \gg 0 \qquad (2)$$

To mitigate this issue, our objective is to learn a process reward model $\mathcal{R}_\phi$ that assigns a step-level correctness score r_i:

$$r_i = \mathcal{R}_\phi(S_i \mid S_{<i}, \mathcal{I}, q) \qquad (3)$$

While $\mathcal{R}_\phi$ outputs a scalar score for efficient inference, its optimization is regularized by a multi-task objective that jointly constrains reasoning logic and cross-view alignment.

3.2 Functional Unit Partitioning and Multi-Dimensional Process Reward Design

To mitigate spurious alignment, we propose a comprehensive process supervision mechanism. It deconstructs the reasoning trajectory into four functional units: *Perceive*, *Align*, *Integrate*, and *Conclude*. Grounded in these structured units, we design a multi-dimensional process reward to simultaneously evaluate intermediate reasoning units across two critical dimensions: fine-grained logical correctness and cross-view alignment quality.

3.2.1 Functional Unit Partitioning of Trajectories

Instead of allowing the base MLLM to freely generate an unconstrained, variable-length CoT, we constrain it to natively generate the trajectory τ as a fixed sequence of exactly four sequential steps: $\tau = (S_1, S_2, S_3, S_4)$. Each step S_i ($i \in \{1, 2, 3, 4\}$) is formalized as a structural triplet:

$$S_i = (u_i, v_i, t_i) \tag{4}$$

where $u_i \in \{\mathcal{U}_{\text{perceive}}, \mathcal{U}_{\text{align}}, \mathcal{U}_{\text{integrate}}, \mathcal{U}_{\text{conclude}}\}$ denotes the functional unit identifier, $v_i \subseteq \{1, \ldots, n\}$ is the active view subset, and t_i is the generated reasoning text. The four sequential phases are defined as follows:

- **Individual View Perception** ($u_1 = \mathcal{U}_{\text{perceive}}$): The model inspects the viewpoints in v_1 independently, extracting localized semantic and spatial information within each view.
- **Cross-View Alignment** ($u_2 = \mathcal{U}_{\text{align}}$): The model cross-references multiple views ($|v_2| \geq 2$) to perform fine-grained spatial registration and establish entity correspondences.
- **Information Integration** ($u_3 = \mathcal{U}_{\text{integrate}}$): The model fuses complementary visual cues across v_3 to resolve perspective-dependent spatial conflicts, thereby constructing a coherent global spatial perception.
- **Final Conclusion** ($u_4 = \mathcal{U}_{\text{conclude}}$): Operating on the integrated global representation, the model reasons about the question q to deduce the final answer within t_4.

3.2.2 Multi-Dimensional Process Reward

Grounded in the functional unit partitioning established in Sec. 3.2.1, we transition from evaluating generic reasoning text to auditing structured reasoning units. Unlike conventional scalar PRMs, we characterize intermediate states through a multi-dimensional supervision space. Specifically, conditioned on the historical context $S_{<i}$, each reasoning step S_i is evaluated against the multi-dimensional schema defined in Table 1.

Table 1: The structured multi-dimensional evaluation and annotation schema.

Dimension	Value Range	Description
`rating`	$\{1, 2, 3, 4, 5\}$	Overall correctness and quality of the reasoning step.
`cross-view`	$\{G, W, E\}^\dagger$	Quality of cross-view alignment and multi-view binding.
`consistency`	$\{G, W, E\}^\dagger$	Logical consistency with respect to the historical context $S_{<i}$.
`comment`	String	Natural language explanation serving as a semantic regularizer.

† G: GOOD (Precise), W: WEAK (Slight Drift), E: ERROR (Hallucinated/Misaligned)

Dimension Semantics. As detailed in Table 1, instead of a crude binary label, `rating` leverages a 5-tier ordinal score to capture subtle deductive nuances (Wang et al., 2023). To explicitly combat spurious alignment, `cross-view` monitors spatial-geometric alignment across viewpoints, utilizing penalizing tokens (`W/E`) to capture spatial drifts or severe hallucinations. Finally, `consistency` safeguards historical coherence against $S_{<i}$ to prevent late-stage trajectory collapse.

Lifecycle Dynamics. Crucially, this schema is not static but dynamically manifests in three distinct morphological forms to balance supervision density with operational efficiency:

- **Teacher Labeling:** The teacher model annotates the full schema (`rating`, `cross-view`, `consistency`, `comment`) to construct a dense multi-view supervision landscape.
- **Student Training:** During multi-task joint optimization, MV-PRM ($\mathcal{R}_\phi$) targets a projected subspace, optimizing *only* the `rating` and `cross-view` dimensions for efficiency (see Sec. 3.3.3).

- **Test-Time Inference:** To eliminate asymmetric computational overhead, $\mathcal{R}_\phi$ deploys only the specialized `rating` head to score intermediate steps during BoN selection (Sec. 3.4.2).

3.3 Automated Data Construction Pipeline for Multi-view Visual Reasoning

Although the multi-dimensional schema proposed in Sec. 3.2.2 provides process-level feedback, manually annotating such fine-grained signals across multiple viewpoints is prohibitively labor-intensive and inherently unscalable. Therefore, we develop a fully automated pipeline to construct this process-level supervision dataset via *structured CoT generation* and *automatic step-level annotation*.

3.3.1 Structured CoT Generation

We employ InternVL3.5-38B (Wang et al., 2025e) to generate structured reasoning trajectories for the AirCopBench training set. To obtain reasoning chains containing structured functional units, we utilize carefully designed few-shot prompts to guide the generator to externalize its latent reasoning steps. Each step is strictly formatted into the (u_i, v_i, t_i) triplet and delimited by specialized tokens to facilitate automated parsing. To ensure data quality, an automated validation pipeline is implemented: any generated trajectory that violates the structured triplet constraints or exhibits degeneration (*e.g.*, severe repetitive token hallucination) is dynamically discarded and regenerated. This structured formulation externalizes the black-box multi-view fusion into an observable reasoning chain, thereby enabling subsequent multi-task optimization to localize and penalize step-level reasoning errors. Detailed prompts are listed in Appendix A.

3.3.2 Automatic Step-level Annotation

To obtain high-fidelity and scalable supervision signals for each reasoning unit, we utilize Qwen3-VL-32B (Bai et al., 2025) as a powerful teacher model to execute rule-based evaluation. For each generated step S_i, the teacher assesses its reasoning quality by considering the multi-modal context, formalized as the mapping: $\mathcal{M}_{\text{teacher}}(\mathcal{I}, q, S_{<i}, S_i) = \mathbf{L}_i$, where $\mathbf{L}_i$ represents the complete four-dimensional structured label set defined in Table 1. This pipeline yields approximately **54k** high-fidelity, step-level training instances.

3.3.3 Target Space Projection for Distillation

Crucially, while the dense text `comment` and historical `consistency` serve as vital cognitive scaffolds to ensure the teacher model outputs unbiased judgments, treating them as direct optimization targets during PRM training would introduce unnecessary noise and computational overhead. Therefore, we strategically project this four-dimensional annotation space into a distilled, dual-objective space for MV-PRM training, focusing exclusively on the `rating` and `cross-view` dimensions.

3.4 MV-PRM Training and Inference

Following the distillation strategy defined in Sec. 3.3.3, we instantiate the dual-objective training framework by architecting a specialized multi-task model. MV-PRM leverages Qwen3-VL-8B-Instruct (Bai et al., 2025) as its architectural backbone, parameter-efficiently adapted via LoRA. To enable multi-task joint optimization, we append two independent MLP heads to the latent representations of step-delimiting tokens: a regression head for continuous *quality rating* and an auxiliary classification head for discrete *cross-view alignment*.

3.4.1 Multi-Task Joint Optimization

To address both reasoning quality and cross-view alignment, MV-PRM is trained via a joint objective consisting of two components: **(i) Main Task (Quality Rating)**, where we optimize the continuous rating output $r_i \in [1, 5]$ via the regression head using a mean squared error (MSE) loss ($\mathcal{L}_{\text{rating}}$); and **(ii) Auxiliary Task (Cross-view Classification)**, where we optimize the discrete alignment labels (*Good*, *Weak*, or *Error*) via the auxiliary classification head using a cross-entropy loss ($\mathcal{L}_{\text{cross-view}}$).

This joint optimization strategy is formalized via a weighted multi-task loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rating}} + \lambda \mathcal{L}_{\text{cross-view}} \tag{5}$$

where λ denotes the balancing hyperparameter. Specifically, the auxiliary `cross-view` classification task forces the model to penalize subtle view inconsistencies that a pure `rating` objective might overlook.

3.4.2 Asymmetric Inference and Chain-Level Scoring

Inspired by the perspective supervision paradigm (Wang et al., 2022; Yang et al., 2023), we employ an asymmetric inference strategy that discards the auxiliary classification head. During training, the shared LoRA parameters ensure that cross-view geometric constraints are implicitly crystallized

Table 2: Comprehensive performance comparison on AirCopBench across various models and question types. Bold numbers indicate the best results across all models within each row. The small red numbers in parentheses show the improvement (+) or decrease (-) brought by +MV-PRM (BoN@8).

Question Type	InternVL3.5-8B		InternVL3.5-38B		Qwen3-VL-8B-Ins		Qwen3-VL-32B-Ins		Ovis2.5-9B	
	Base	+MV-PRM	Base	+MV-PRM	Base	+MV-PRM	Base	+MV-PRM	Base	+MV-PRM
Avg	53.64	61.30 (↑7.66)	53.73	61.49 (↑7.76)	51.21	60.23 (↑9.02)	58.39	**66.34** (↑7.95)	59.07	64.69 (↑5.62)
Obj. Matching	57.14	68.57 (↑11.43)	65.71	**74.29** (↑8.58)	66.67	72.38 (↑5.71)	67.62	72.38 (↑4.76)	66.67	72.38 (↑5.71)
Obj. Grounding	59.82	60.71 (↑0.89)	58.93	65.18 (↑6.25)	51.79	53.57 (↑1.78)	63.39	**72.32** (↑8.93)	57.14	60.71 (↑3.57)
Obj. Recognition	52.38	63.10 (↑10.72)	61.90	72.62 (↑10.72)	57.14	71.43 (↑14.29)	59.52	72.62 (↑13.10)	67.86	**75.00** (↑7.14)
Causal Assess.	70.24	76.19 (↑5.95)	66.67	77.38 (↑10.71)	59.52	72.62 (↑13.10)	71.43	**84.52** (↑13.09)	76.19	79.76 (↑3.57)
Quality Assess.	38.10	**47.62** (↑9.52)	30.95	28.57 (↓2.38)	28.57	28.57 (—)	33.33	36.90 (↑3.57)	36.90	39.29 (↑2.39)
Usability Assess.	56.10	**58.54** (↑2.44)	48.78	56.10 (↑7.32)	32.93	50.00 (↑17.07)	39.02	51.22 (↑12.20)	36.59	54.88 (↑18.29)
Obj. Counting	34.57	34.57 (—)	27.16	29.63 (↑2.47)	35.80	38.27 (↑2.47)	24.69	**48.15** (↑23.46)	34.57	43.21 (↑8.64)
What to Collab.	62.96	**81.48** (↑18.52)	70.37	**81.48** (↑11.11)	75.31	77.78 (↑2.47)	76.54	**81.48** (↑4.94)	72.84	75.31 (↑2.47)
Why to Collab.	62.96	64.20 (↑1.24)	65.43	75.31 (↑9.88)	61.73	71.60 (↑9.87)	72.84	74.07 (↑1.23)	75.31	**79.01** (↑3.70)
Who to Collab.	50.62	76.54 (↑25.92)	53.09	66.67 (↑13.58)	62.96	86.42 (↑23.46)	80.25	87.65 (↑7.40)	77.78	**90.12** (↑12.34)
Scene Desc.	63.16	57.89 (↓5.27)	52.63	61.40 (↑8.77)	50.88	59.65 (↑8.77)	64.91	**66.67** (↑1.76)	61.40	61.40 (—)
Scene Comp.	51.16	58.14 (↑6.98)	55.81	**60.47** (↑4.66)	44.19	58.14 (↑13.95)	55.81	58.14 (↑2.33)	53.49	48.84 (↓4.65)
Obs. Posture	22.22	**40.74** (↑18.52)	29.63	37.04 (↑7.41)	14.81	33.33 (↑18.52)	33.33	33.33 (—)	37.04	37.04 (—)
When to Collab.	34.48	34.48 (—)	27.59	37.93 (↑10.34)	27.59	31.03 (↑3.44)	48.28	48.28 (—)	48.32	**55.17** (↑6.85)

Note: *Obj.*, *Assess.*, *Collab.*, *Desc.*, *Comp.*, and *Obs.* denote Object, Assessment, Collaborate, Description, Comparison, and Observing, respectively.

into the backbone representations. Consequently, the remaining regression head inherently evaluates reasoning quality through a multi-view lens, eliminating heuristic tuning during Best-of-N (BoN) reranking and maximizing efficiency.

For a candidate chain $\tau = \{S_i\}_{i=1}^{T}$, the chain score $R(\tau)$ is computed by averaging the step-level rewards r_i derived from the `rating` head:

$$R(\tau) = \frac{1}{T}\sum_{i=1}^{T} r_i \tag{6}$$

Finally, the candidate chain with the highest score $R(\tau)$ is selected as the final answer during BoN reranking.

4 Experiments

4.1 Experiment Settings

4.1.1 Datasets

We leverage the MV-PRM Dataset, introduced in Sec. 3.3, for model training. Comprising approximately 54k step-level annotations derived from the training split of AirCopBench, this dataset provides essential multi-task joint supervision to simultaneously guide step-level quality evaluation and cross-view alignment.

4.1.2 Evaluation Benchmarks

Our empirical evaluation is primarily conducted on the **AirCopBench** test set, which encompasses 1,031 multi-view VQA instances. These samples span 14 visual reasoning tasks, thereby facilitating a holistic assessment of multi-view collaborative perception.

4.1.3 Baselines

To thoroughly verify the efficacy of MV-PRM, we evaluate it against two primary categories of baselines: (1) **Base MLLMs** under direct inference (including the InternVL3.5, Qwen3-VL, and Ovis2.5 series), and (2) **Zero-shot MLLM Evaluators** serving as off-the-shelf process critics for ablation. Due to space constraints, the detailed configurations, model parameters, and specific variants of these baselines are elaborated in Appendix B.1.

4.1.4 Implementation Details

We train the MV-PRM framework on the MV-PRM dataset using a total batch size of 64, distributed across 8 NVIDIA A100 GPUs. The full training convergence dynamics are illustrated in Figure 2, and detailed hyperparameter configurations are provided in Appendix B.2.

4.2 Main Results

Table 2 reports a comprehensive performance comparison across 14 visual reasoning tasks on AirCopBench. The empirical results demonstrate that **+MV-PRM (BoN@8)** consistently elevates the end-to-end question-answering performance across di-

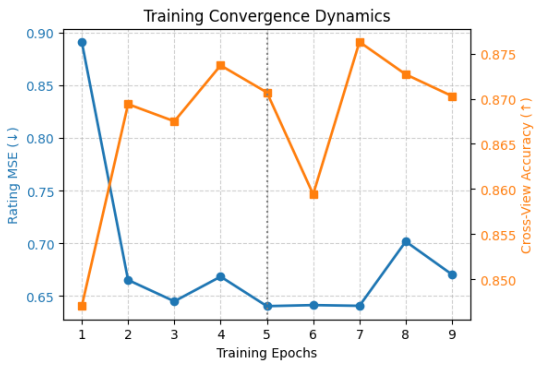

Figure 2: Training convergence dynamics. The primary Rating MSE and auxiliary Cross-View Accuracy are tracked per epoch. Training is terminated at Epoch 5 to secure the model checkpoint with the highest reasoning quality.

verse MLLM backbones, validating its effectiveness as a multi-view process reward model.

4.2.1 Performance Gains on Collaborative Reasoning

As demonstrated in Table 2, incorporating MV-PRM consistently improves the accuracy across all evaluated MLLM families, including the InternVL, Qwen, and Ovis series. Notably, Qwen3-VL-32B-Instruct with MV-PRM achieves the highest average accuracy of **66.34%**, yielding a **+7.95%** absolute improvement over its greedy baseline. Across the entire benchmark, the performance gains are particularly pronounced in collaboration-intensive tasks, such as *Who to Collaborate* (+7.40% for Qwen3-VL-32B-Instruct, +23.46% for Qwen3-VL-8B-Instruct) and *What to Collaborate* (+18.52% for InternVL3.5-8B, +11.11% for InternVL3.5-38B). This consistent improvement indicates that while vanilla MLLMs frequently generate inconsistent or conflicting reasoning chains in complex multi-view scenarios, MV-PRM effectively identifies and selects the optimal reasoning chain from the sampled candidates.

4.2.2 Scalability and Homologous Affinity Across Models

The empirical results highlight two key properties of our approach: *robust scalability* and *homologous affinity*. First, MV-PRM scales effectively across different model sizes and architectural families. It provides substantial gains across scales, ranging from compact models (+9.02% for Qwen3-VL-8B-Instruct) to large-scale backbones (+7.76% for InternVL3.5-38B) without performance saturation. Second, the results reveal a distinct *homologous affinity* within the Qwen3-VL family. Specifically, Qwen3-VL-8B-Instruct achieves the highest average improvement (**+9.02%**) across the benchmark, closely followed by Qwen3-VL-32B-Instruct (**+7.95%**). Because our MV-PRM is built upon the Qwen3-VL-8B-Instruct backbone, it shares a highly compatible representation space with other models in the Qwen3-VL family. This representation alignment enables the reward model to evaluate the candidate reasoning chains more accurately.

4.3 Ablation Study

We primarily investigate whether the performance gains of MV-PRM stem from our proposed multi-task joint optimization or simply from the inherent reasoning capabilities of MLLMs used as rerankers. As demonstrated in Table 3, we evaluate MV-PRM against two strong off-the-shelf baselines—Qwen3-VL-8B-Instruct and Qwen3-VL-32B-Instruct—serving as direct zero-shot rerankers on the same pool of $N = 8$ sampled candidates.

The empirical results indicate that simply employing a powerful off-the-shelf MLLM as a reranker yields marginal or even negative gains compared to the greedy baseline. Specifically, while the zero-shot Qwen3-VL-32B-Instruct occasionally improves performance in localized tasks (e.g., achieving 32.10% in *Object Counting*), its average performance frequently falls below the vanilla InternVL3.5-38B baseline (52.96% vs. 53.73%). In contrast, MV-PRM provides a substantial and consistent boost, raising the average accuracy to **61.49%** for the 38B generator and **61.30%** for the 8B variant.

This performance gap becomes particularly prominent in collaboration-intensive tasks such as *Who to Collaborate*, where zero-shot evaluators severely struggle to parse multi-view coordination logic. In this category, Qwen3-VL-32B-Instruct achieves only 44.44% on the 38B base, significantly underperforming the baseline's 53.09%. This negative optimization implies that without targeted adaptation, generic MLLMs lack the alignment awareness required for cross-view alignment. Conversely, **MV-PRM** effectively elevates the score to **66.67%** (+13.58%), demonstrating that our step-level process rewards successfully verify the multi-view alignment that general-purpose models lack. Consequently, these findings confirm that the performance gains of MV-PRM stem di-

Table 3: Performance comparison of various rerankers for BoN@8 on AirCopBench using two different base models (InternVL3.5-38B and 8B). Bold numbers indicate the best results within each base model group. The red small numbers in parentheses show the improvement (+) or decrease (-) relative to the respective Base model.

Question Type	Base Model	+Qwen3-VL-8B-Ins	+Qwen3-VL-32B-Ins	+MV-PRM (Ours)
Exp Group A: Base Model = InternVL3.5-38B				
Avg	53.73	51.89	52.96	**61.49** (↑7.76)
Obj. Matching	65.71	62.86	60.95	**74.29** (↑8.58)
Obj. Grounding	58.93	52.68	56.25	**65.18** (↑6.25)
Obj. Recognition	61.90	60.71	61.90	**72.62** (↑10.72)
Causal Assess.	66.67	65.48	65.48	**77.38** (↑10.71)
Quality Assess.	30.95	29.76	**34.52**	28.57 (↓2.38)
Usability Assess.	48.78	45.12	46.34	**56.10** (↑7.32)
Obj. Counting	27.16	29.63	**32.10**	29.63 (↑2.47)
What to Collab.	70.37	71.60	71.60	**81.48** (↑11.11)
Why to Collab.	65.43	67.90	61.73	**75.31** (↑9.88)
Who to Collab.	53.09	40.74	44.44	**66.67** (↑13.58)
Scene Desc.	52.63	56.14	54.39	**61.40** (↑8.77)
Scene Comp.	55.81	55.81	**60.47**	**60.47** (↑4.66)
Obs. Posture	29.63	33.33	33.33	**37.04** (↑7.41)
When to Collab.	27.59	24.14	31.03	**37.93** (↑10.34)
Exp Group B: Base Model = InternVL3.5-8B				
Avg	53.64	52.57	54.41	**61.30** (↑7.66)
Obj. Matching	57.14	60.00	62.86	**68.57** (↑11.43)
Obj. Grounding	59.82	48.21	57.14	**60.71** (↑0.89)
Obj. Recognition	52.38	54.76	57.14	**63.10** (↑10.72)
Causal Assess.	70.24	63.10	69.05	**76.19** (↑5.95)
Quality Assess.	38.10	46.43	42.86	**47.62** (↑9.52)
Usability Assess.	56.10	52.44	54.88	**58.54** (↑2.44)
Obj. Counting	**34.57**	32.10	28.40	**34.57** (—)
What to Collab.	62.96	67.90	70.37	**81.48** (↑18.52)
Why to Collab.	62.96	59.26	56.79	**64.20** (↑1.24)
Who to Collab.	50.62	56.79	58.02	**76.54** (↑25.92)
Scene Desc.	**63.16**	54.39	59.65	57.89 (↓5.27)
Scene Comp.	51.16	55.81	53.49	**58.14** (↑6.98)
Obs. Posture	22.22	22.22	29.63	**40.74** (↑18.52)
When to Collab.	**34.48**	27.59	20.69	**34.48** (—)

Note: *Obj.*, *Assess.*, *Collab.*, *Desc.*, *Comp.*, and *Obs.* denote Object, Assessment, Collaborate, Description, Comparison, and Observing, respectively.

rectly from our proposed multi-task joint optimization, which synergistically reconciles step-level quality rating and cross-view alignment.

5 Conclusion

In this paper, we address a critical yet overlooked problem in complex multi-view scenarios: MLLMs frequently struggle to align information across disparate viewpoints, leading to spurious alignment and compounding inference-time errors. To mitigate this issue in test-time scaling, we present **MV-PRM**, a novel multi-view process reward model tailored for fine-grained, step-level quality assessment and cross-view alignment verification. To ground this framework, we design a multi-dimensional supervision mechanism tailored for multi-view scenarios and propose a fully automated pipeline to yield step-level feedback at scale. Comprehensive evaluations on the collaboration-intensive AirCopBench demonstrate that incorporating MV-PRM consistently yields substantial performance gains across diverse MLLM architectures and scales, validating the generalizability and effectiveness of our paradigm.

6 Limitations

Despite its effectiveness, our paradigm exhibits two primary limitations. First, MV-PRM operates as an external inference-time verifier via the BoN strategy, which inherently increases computational overhead and latency compared to greedy decoding. Although this approach yields substantial gains without modifying base MLLM parameters, integrating early-stopping or rejection sampling remains essential for deployment efficiency. Second, as a reranker, **MV-PRM** is strictly bounded by the generative capacity of the underlying MLLM. If the base generator fails to output any trace containing proper cross-view alignment across all N candidates, the verification process cannot rectify such a generation bottleneck. To overcome this limitation, future work will leverage **MV-PRM** to directly optimize the generator via iterative SFT or reinforcement learning (e.g., PPO or DPO), thereby internalizing the external verification capability into intrinsic multi-view alignment.

References

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, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Qi Cao, Ruiyi Wang, Ruiyi Zhang, Sai Ashish Somayajula, and Pengtao Xie. 2025. DreamPRM: Domain-reweighted process reward model for multimodal reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Fuhao Chang, Shuxin Li, Yabei Li, and Lei He. 2025. Vlm-3d:end-to-end vision-language models for open-world 3d perception. *Preprint*, arXiv:2508.09061.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. 2024. Multi-agent collaborative perception via motion-aware robust communication network. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15301–15310.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. 2024. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*, volume 37, pages 113991–114017. Curran Associates, Inc.

Jiayi Ji, Haowei Wang, Changli Wu, Yiwei Ma, Xiaoshuai Sun, and Rongrong Ji. 2025. Jm3d & jm3d-llm: Elevating 3d representation with joint multimodal cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2475–2492.

Weijie Li, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2026. Step-grpo: Enhancing reasoning quality and efficiency via structured prm-based reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31734–31742.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Zefu Lin, Wenbo Chen, Xiaojuan Jin, Yuran Yang, Lue Fan, Yixin Zhang, Yufeng Zhang, and Zhaoxiang Zhang. 2025. Mcop: Multi-uav collaborative occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27242–27251.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *Preprint*, arXiv:2408.02718.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. R-prm: Reasoning-driven process reward modeling. *Preprint*, arXiv:2503.21295.

Lin Sun, Chuang Liu, Xiaofeng Ma, Tao Yang, Weijia Lu, and Ning Wu. 2025. Freeprm: Training process reward models without ground truth process labels. *Preprint*, arXiv:2506.03570.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Shaohong Wang, Lu Bin, Xinyu Xiao, Zhiyu Xiang, Hangguan Shan, and Eryun Liu. 2025a. Iftr: An instance-level fusion transformer for visual collaborative perception. In *Computer Vision – ECCV 2024*, pages 124–141, Cham. Springer Nature Switzerland.

Shuai Wang, Zhenhua Liu, Jiaheng Wei, Xuanwu Yin, Dong Li, and Emad Barsoum. 2025b. Athena: Enhancing multimodal reasoning with data-efficient process reward models. *Preprint*, arXiv:2506.09532.

Tai Wang, Qing Lian, Chenming Zhu, Xinge Zhu, and Wenwei Zhang. 2022. MV-FCOS3D++: Multi-View camera-only 4d object detection with pretrained monocular backbones. *arXiv preprint*.

Teng Wang, Zhangyi Jiang, Zhenqi He, Shenyang Tong, Wenhan Yang, Yanan Zheng, Zeyu Li, Zifan He, Hailei Gong, Zewen Ye, Shengjie Ma, and Jianping Zhang. 2025c. Towards hierarchical multi-step reward models for enhanced reasoning in large language models. *Preprint*, arXiv:2503.13551.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *Preprint*, arXiv:2308.04592.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025d. Visualprm: An effective process reward model for multimodal reasoning. *Preprint*, arXiv:2503.10291.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025e. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xijun Wang, Rayyan Abdalla, Junyun Huang, Chengyuan Zhang, Ruiqi Xian, and Dinesh Manocha. 2026. Bi-vlm: Binary post-training quantization for vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):10207–10215.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2025. Pointllm: Empowering large language models to understand point clouds. In *Computer Vision – ECCV 2024*, pages 131–147, Cham. Springer Nature Switzerland.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxi Tao, Feng Zhou, and Zhaoxiang Yu. 2023. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5414–5424.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. 2026. Seeing from another perspective: Evaluating multi-view understanding in mllms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):12000–12008.

Jirong Zha, Yuxuan Fan, Tianyu Zhang, Geng Chen, Yingfeng Chen, Chen Gao, and Xinlei Chen. 2026. Aircopbench: A benchmark for multi-drone collaborative embodied perception and reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2):1507–1515.

Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, and Liang Song. 2024. Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12575–12584.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025a. Generative verifiers: Reward modeling as next-token prediction. *Preprint*, arXiv:2408.15240.

Xiangxiang Zhang, Jingxuan Wei, Donghong Zhong, Qi Chen, Caijun Jia, Cheng Tan, Jinming Gu, Xiaobo Qin, Zhiping Liu, Liang Hu, Tong Sun, Yuchen Wu, Zewei Sun, Chenwei Lou, Hua Zheng, Tianyang Zhan, Changbao Wang, Shuangzhi Wu, Zefa Lin, and 15 others. 2025b. Structvrm: Aligning multimodal reasoning with structured and verifiable reward models. *Preprint*, arXiv:2508.05383.

Tong Zhao, Junping Du, Zhe Xue, Meiyu Liang, Aijing Li, Xiaolong Meng, and Dandan Liu. 2026. St-vlm: A spatial-to-image multimodal spatial-temporal prediction framework with vision-language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16441–16449.

Duo Zheng, Shijia Huang, and Liwei Wang. 2025. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8995–9006.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Eric Xin Wang, and Achuta Kadambi. 2025. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8600–8612.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2025. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4295–4305.

A Dataset & Prompt Templates

A.1 Dataset Statistics

To investigate the core characteristics of our process-level supervision, we conduct a quantitative analysis on the label proportions within the MV-PRM dataset. Comprising approximately **54k** step-level annotated instances, the fine-grained percentage distribution of labels across the three primary annotation dimensions—`rating`, `cross_view`, and `consistency`—is summarized in Table 4.

Table 4: Label distribution (%) of multi-dimensional annotations in the MV-PRM training set.

Annotation Dimension & Label	Percentage (%)
Quality Rating (`rating`)	
Score 1 (Critical Failure)	7.48%
Score 2 (Major Flaw)	13.82%
Score 3 (Mediocre/Pass)	21.90%
Score 4 (Good Quality)	24.80%
Score 5 (Flawless/Optimal)	31.99%
Cross-View Alignment (`cross_view`)	
GOOD	56.76%
WEAK (Spurious/Partial)	42.20%
ERROR	1.05%
Historical Consistency (`consistency`)	
GOOD	78.94%
WEAK	10.82%
ERROR	10.24%

The distribution highlights two localized insights regarding multi-view reasoning:

- **Linguistic-Visual Disconnect:** While 78.94% of reasoning steps maintain high historical coherence (`consistency`=GOOD), 42.20% exhibit `WEAK` cross-view alignment. This discrepancy indicates that general-purpose MLLMs frequently generate linguistically plausible trajectories that yet suffer from *spurious alignment* across viewpoints.
- **Granular Supervision:** The continuous distribution of scalar `rating` (Scores 1–5) provides dense, non-binary quality signals. This multi-level distribution prevents reward collapse and facilitates effective Best-of-N (BoN) reranking during inference.

A.2 Prompt Templates

To ensure reproducibility, we provide the exact text prompts used for data generation and multi-dimensional annotation.

- **Prompt 1: Structured CoT Generation (InternVL3.5-38B).** This prompt guides the generator model to externalize its latent multi-view reasoning process into formatted (u_i, v_i, t_i) triplets. (See Table 5).
- **Prompt 2: Multi-Dimensional Automatic Annotation (Qwen3-VL-32B).** This rule-based prompt instructs the teacher model to evaluate each step across four dimensions (`rating`, `cross-view`, `consistency`, and `comment`). (See Table 7).

B Implementation Details

To guarantee the empirical reproducibility of our work, we elaborate on the baseline configurations and the exact hyperparameter setup used during the training of MV-PRM.

B.1 Baseline MLLMs as Verifiers

To ensure a rigorous and comprehensive evaluation, the baselines integrated into our benchmarking suite are meticulously configured as follows:

1) Base MLLMs (Direct Inference): We evaluate the zero-shot multi-view reasoning performance of several cutting-edge, open-source multimodal large language models (MLLMs) across varying parameter scales. Specifically, this includes **InternVL3.5-8B/38B**, **Qwen3-VL-8B/32B-Instruct**, and **Ovis2.5-9B**. These models act as the core task performers (Generators). We report their task accuracy under standard greedy decoding configurations to establish a robust empirical baseline for native multi-view consistency and spatial reasoning.

2) Zero-shot MLLM Evaluators (for Ablation): To decouple the performance gains of MV-PRM and prove they originate from our specialized step-level multi-task joint optimization rather than the latent capability of scaled backbones, we ablate our framework against off-the-shelf advanced MLLMs.

Box 1 | Task Prompt Template for Structured Multi-View CoT Generation

You are an expert multi-view collaborative perception system.

Input Multi-View Images:
⟨`Multi_View_Images`⟩ *(e.g., View v1: ⟨image_1⟩, View v2: ⟨image_2⟩)*
The available valid view IDs are: ⟨`View_IDs`⟩. *Use ONLY these IDs in your reasoning.*

Task Description:
You are given the multi-view images specified above capturing the same scene. You will answer the provided multiple-choice question by reasoning about what you observe, how the different views complement or support each other, and record your inference trajectory across the designated functional units. Your reasoning MUST be strictly guided by the question.

You MUST output EXACTLY 5 lines in plain English, nothing else, with no empty lines or extra newlines between them:
Step 1 [Perceive] (views: ⟨*View_IDs_Used*⟩): `[Your concise, single-sentence description of relevant objects and scene elements, approx. 20 words]`
Step 2 [Align] (views: ⟨*View_IDs_Used*⟩): `[Your concise, single-sentence matching or comparison of elements across views, approx. 20 words]`
Step 3 [Integrate] (views: ⟨*View_IDs_Used*⟩): `[Your concise, single-sentence synthesis of multi-view evidence for unified understanding, approx. 25 words]`
Step 4 [Conclude] (views: ⟨*View_IDs_Used*⟩): `[Your concise, single-sentence deduction of the final choice with a brief justification, approx. 20 words]`
⟨`Option`⟩

Strict rules (must follow exactly):
– Only use view IDs from: ⟨`View_IDs`⟩.
– Each step line (Steps 1–4) must contain exactly ONE grammatically complete sentence and strictly adhere to the approximate word budget.
– Do NOT output JSON, bullet points, explanations, or any extra text before/after/between these 5 lines. No empty lines.

Question: ⟨`Question_Text`⟩

Options:
A: ⟨`Option_A`⟩ **B:** ⟨`Option_B`⟩ **C:** ⟨`Option_C`⟩ **D:** ⟨`Option_D`⟩

Think step by step using the provided images and output only the 5 required lines.

Table 5: The exact system prompt template mapped to the structural data generation for InternVL3.5-38B.

We deploy **Qwen3-VL-8B-Instruct** and **Qwen3-VL-32B-Instruct** to act as zero-shot process critics. They are supplied with identical trajectory steps and evaluation rubrics to score the reasoning chains, serving as a direct comparison to showcase the necessity of our fine-grained process reward training.

Hyperparameter Specification	Value / Choice
Optimization & Schedule	
Optimization Algorithm	AdamW
Initial Learning Rate (η)	1.2×10^{-4}
Weight Decay (λ)	0.01
Learning Rate Schedule	Cosine Decay with Warmup
Warmup Step Ratio	3.0%
Training Epochs	9 (Early-stopped at Epoch 5)
Gradient Clipping Limit ($\|\mathbf{g}\|_2$)	0.5
Parameter-Efficient Fine-Tuning (LoRA)	
Target Projection Modules	`q, k, v, o, gate, up, down`
LoRA Rank (r)	32
LoRA Alpha (α_{lora})	64
LoRA Dropout Rate	0.05

Table 6: Hyperparameter and training configuration.

B.2 Hyperparameter Configuration

The core backbone is initialized from the **Qwen3-VL-8B-Instruct** architecture and fine-tuned via Low-Rank Adaptation (LoRA). The multi-task optimization pipeline jointly constrains a continuous scalar regressor for step-level quality prediction alongside a multi-class categorizer for view alignment auditing. Table 6 enumerates the primary hyperparameters utilized in our training process, while the convergence dynamics are visualized in Figure 2.

Box 2 | Prompt Template for Step-Level Teacher Annotations

You are an objective evaluator performing a standardized quality audit on ONE reasoning step within a multi-view chain-of-thought trajectory.

Evaluation Context & Inputs:
Evaluate ONLY the CURRENT step S_i by cross-referencing:
– **ALL** provided multi-view images (depicting the identical scene).
– The core question and exact options.
– **ALL** historical steps $S_{<i}$ (for strict chronological and state consistency verification).
– If `step_index == 4`, you are also provided with the student's `final_answer` and ground-truth `gt_answer`.

Base your evaluation strictly on the factual density of spatial propositions and logical veracity, independent of linguistic style or verbosity.

Output Format Requirement:
Output ONLY a single-line valid JSON object with no extra conversational text before or after:

```
{
  "step_index": ⟨int⟩,
  "rating": ⟨int⟩,                                   // 1–5, mapped strictly via the rubric below
  "cross_view": "GOOD"|"WEAK"|"ERROR",
  "consistency": "GOOD"|"WEAK"|"ERROR",
  "comment": "⟨English explanation, max 35 words⟩"
}
```

Criteria Definitions:
[CROSS_VIEW Alignment]
– **GOOD:** (*Step 1*): Accurately anchors observations to correct individual view IDs. (*Steps 2–3*): Explicitly names $\geq$2 views **AND** explicitly binds objects across perspectives or leverages complementary multi-view information (e.g., occlusion resolution, relative positioning). (*Step 4*): Correctly inherits prior multi-view evidence.
– **WEAK:** Mentions view IDs but lacks exact spatial anchoring, or fails to execute explicit inter-view comparisons when required.
– **ERROR:** Misattributes objects to the wrong view ID, or directly contradicts geometric layouts visible in the imagery.
[CONSISTENCY]
– **GOOD:** Complete alignment with question premises, zero historical state conflicts, and absolute fidelity to the physical imagery.
– **WEAK:** Minor spatial ambiguity or non-critical omissions, but remains logically plausible and non-contradictory.
– **ERROR:** Contradicts previous steps/question/images; hallucination; ignores obvious evidence; logical contradiction

Rating Rubric (1–5):
– **5 (Optimal):** Both dimensions are `[GOOD]`. Flawless task execution of the designated functional unit.
– **4 (Valid):** Both dimensions are `[GOOD]`. Sound multi-view reasoning with negligible minor details for improvement.
– **3 (Mediocre):** At least one dimension is `[WEAK]`, but neither is `[ERROR]`. The reasoning is shallow or overly generic.
– **2 (Defective):** Either dimension contains an `[ERROR]`, or both are `[WEAK]`. Noticeable logical gaps or weak spatial coordination.
– **1 (Fatal):** Either dimension triggers a critical `[ERROR]`, causing an irrecoverable breakdown of the reasoning chain.

Outcome Alignment (`step_index == 4` ONLY):
– If `final_answer == gt_answer` **AND** Step 4 logically supports the correct choice $\rightarrow$ Score **4** or **5**.
– If `final_answer != gt_answer`:
  ∗ If Step 4 rationalizes toward the incorrect choice $\rightarrow$ Score **1** or **2**.
  ∗ If Step 4 is neutral $\rightarrow$ Score $\leq$**3**.
– *A Conclude step leading to an incorrect final option cannot receive a score of 4 or 5.*

Output only the JSON object.

Table 7: The exact system prompt template used for generating step-level teacher annotations in the MV-PRM training set.