Learning What Matters: Dynamic Dimension Selection and Aggregation for Interpretable Vision-Language Reward Modeling

Qiyuan Chen[1,3*], **Hongsen Huang**[3*], **Jiahe Chen**[1,4], **Qian Shao**[1,4], **Jintai Chen**[5]
Hongxia Xu[2,4†], **Renjie Hua**[3,6], **Chuan Ren**[3†], **Jian Wu**[2,7]
[1]College of Computer Science & Technology Zhejiang University
[2]State Key Laboratory of Transvascular Implantation Devices and TIDRI
[3]Soochow Securities Co.,Ltd. [4]WeDoctor Cloud and Liangzhu Laboratory [5]HKUST(GZ)
[6]Nanjing University [7]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence
qiyuanchen@zju.edu.cn

Abstract

Vision-language reward modeling faces a dilemma: generative approaches are interpretable but slow, while discriminative ones are efficient but act as opaque "black boxes." To bridge this gap, we propose VL-MDR (Vision-Language Multi-Dimensional Reward), a framework that dynamically decomposes evaluation into granular, interpretable dimensions. Instead of outputting a monolithic scalar, VL-MDR employs a visual-aware gating mechanism to identify relevant dimensions and adaptively weight them (e.g., Hallucination, Reasoning) for each specific input. To support this, we curate a dataset of 321k vision-language preference pairs annotated across 21 fine-grained dimensions. Extensive experiments show that VL-MDR consistently outperforms existing open-source reward models on benchmarks like VL-RewardBench. Furthermore, we show that VL-MDR-constructed preference pairs effectively enable DPO alignment to mitigate visual hallucinations and improve reliability, providing a scalable solution for VLM alignment.

1 Introduction

Reward Models (RMs) serve as a cornerstone for aligning Large Vision-Language Models (LVLMs) with human preferences (Christiano et al., 2017; Ouyang et al., 2022; Yang et al., 2025). Beyond driving Reinforcement Learning from Human Feedback (RLHF) (Rafailov et al., 2023), RMs are increasingly pivotal for test-time scaling strategies, such as rejection sampling, enhancing reliability without model retraining (Cui et al., 2023; Yu et al., 2024). As LVLMs transition towards post-training with AI-augmented synthetic data (Bai et al., 2025b,a), robust and scalable automated evaluation becomes essential for ensuring principled guidance throughout the model lifecycle.

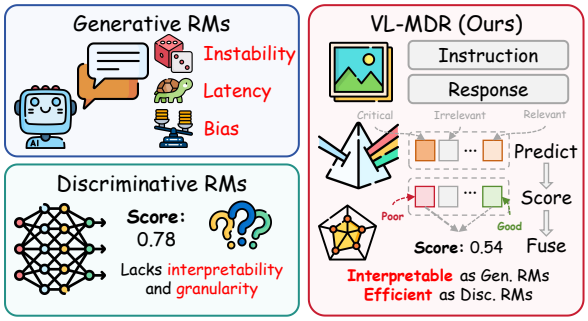

Figure 1: **Comparison of paradigms.** Unlike Generative RMs (high latency) and Discriminative RMs (opaque scalars), **VL-MDR** dynamically decomposes evaluation into granular dimensions, achieving both interpretability and efficiency.

Existing multimodal reward modeling generally falls into two paradigms, each with distinct limitations. *Generative RMs* (e.g., LLaVA-Critic (Xiong et al., 2025)) offer interpretability via textual critiques but suffer from high latency, stochasticity, and cognitive biases (e.g., position bias) (Xiong et al., 2025). Conversely, *Discriminative RMs* (e.g., Skywork-VL (Wang et al., 2025a)) formulate preference learning as scalar regression. However, they operate as opaque "black boxes": a single scalar cannot disentangle the multidimensional nature of multimodal quality (e.g., distinguishing visual faithfulness from logical reasoning) (Zang et al., 2025). Consequently, they fail to provide the granular feedback necessary for targeted model iteration.

To address this dichotomy, we take a step back to scrutinize the cognitive complexities unique to multimodal evaluation. Unlike pure text, assessing vision-language responses requires navigating a hierarchical cross-modal dependency. A response might be linguistically fluent yet visually hallucinatory; it might correctly perceive objects but fail to reason about their spatial relationships. Traditional discriminative models, by compressing these orthogonal dimensions into a single scalar, obscure the rationale behind the preference. They fail to distinguish between a model that is "blind" (per-

[*] Equal Contribution.
[†] Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19779–19792

ception failure) and one that is "biased" (reasoning failure), rendering the feedback opaque and difficult to steer. This motivates a question: *Can we design a reward model that mirrors this granular, visually-grounded cognitive process while retaining the high throughput of discriminative scorers?*

In response, we propose the **V**ision-**L**anguage **M**ulti-**D**imensional **R**eward (**VL-MDR**) model. As shown in Figure 1, VL-MDR reformulates evaluation as a dynamic disentanglement process. Our key insight is that a reward model should first "perceive" the intent of a multimodal query before "judging" the response. Specifically, VL-MDR employs a three-stage mechanism: (1) *Visual-Aware Dimension Prediction*, where the model predicts relevance probabilities over dimensions and selects the Top-k active dimensions for the current image-text pair; (2) *Fine-Grained Scoring*, which assesses the response quality across these targeted dimensions; and (3) *Adaptive Weighting*, which fuses these scores into a final reward based on their contextual importance. To support this granular supervision, we curate a large-scale fine-grained preference dataset comprising approximately 321k pairs, where each pair is annotated with fine-grained preferences on the target dimensions drawn from a 21-dimension fine-grained taxonomy. This data foundation enables VL-MDR to effectively disentangle perception errors from reasoning errors, offering the interpretability of a generative judge with the efficiency of a discriminative scorer.

Extensive experiments on three multimodal reward benchmarks show that VL-MDR achieves strong and stable preference modeling with competitive performance across diverse categories. Meanwhile, VL-MDR remains efficient with a single forward pass and a lightweight multi-dimensional head, and its annotations produce higher-quality DPO preference pairs that improve downstream LVLMs and further reduce visual hallucinations.

Our main contributions are summarized as follows: (1) We propose VL-MDR, an efficient multi-dimensional vision-language reward model that performs visual-aware dimension selection and masked adaptive aggregation to produce an interpretable reward in a single forward pass. (2) We curate a large-scale fine-grained preference dataset of $\sim$321k pairs with consistent annotations over a 21-dimension hierarchical taxonomy, enabling scalable supervision for dimension-aware reward learning. (3) Experiments on multiple benchmarks show that VL-MDR outperforms strong open-source reward models, and further provides effective preference pairs for DPO alignment to improve reliability and reduce hallucinations in downstream LVLMs.

2 Dataset Construction

To facilitate fine-grained capability assessment and alignment, we construct a high-quality preference dataset containing approximately **321k** samples. As illustrated in Figure 2, our construction pipeline integrates multi-source data integration, applies a hierarchical capability taxonomy, and enforces a multi-model fine-grained overall-consistency filtering mechanism to ensure high data quality.

2.1 Data Integration and Taxonomy

Our dataset is built upon a diverse, aggregated pool of approximately 414.2k preference samples collected from seven widely used VLM preference datasets. To avoid ambiguity about the provenance of supervision, we explicitly categorize these sources by the origin of their preference signals. Specifically, VLFeedback (Li et al., 2023), RLAIF-V (Yu et al., 2025), and SPA-VL (Zhang et al., 2025b) provide large-scale preferences generated via AI feedback. In contrast, VisionArena (Chou et al., 2025), WildVision (Lu et al., 2024), RLHF-V (Yu et al., 2024), and MM-RLHF (Zhang et al., 2025a) primarily contain human-annotated preferences or preferences that have been human-verified, offering higher-fidelity supervision.

Moving beyond generic quality scores, we design a hierarchical capability taxonomy to precisely characterize the skills required for each sample. As detailed in Table 1, this taxonomy consists of 7 Core Capabilities which branch into 21 Fine-grained Dimensions. This structured approach allows us to decouple complex multimodal tasks into specific, interpretable skill sets, providing a granular semantic index for the entire dataset.

2.2 Automated Annotation and Filtering

We employ a panel of strong VLM judges: `Qwen3-VL-235B-A22B-Instruct`, `GLM-4.5V`, and `InternVL3-78B` to automatically predict fine-grained dimensions and verify preference labels via multi-model consistency.

For each sample, we first prompt each judge to annotate the image-question pair with its top-3 fine-grained dimensions (the prompt is shown in Figure 8 in Appendix D). We retain a dimension annotation only when the predicted top-3 dimen-

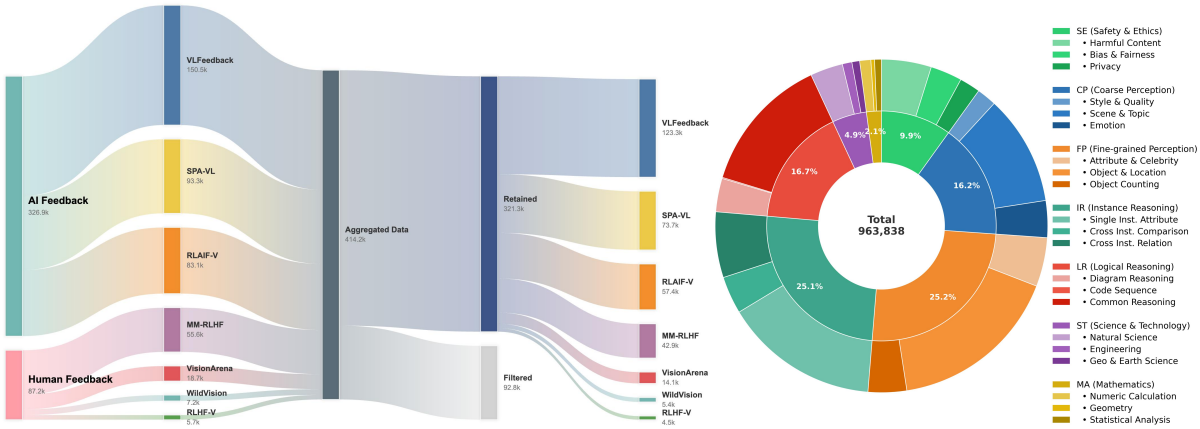

Figure 2: **Data Construction Pipeline and Capability Distribution. Left:** We aggregate ∼414.2k preference samples from 7 different VLM preference datasets, grouped by supervision provenance (AI Feedback vs. Human Feedback), and apply our multi-model fine-grained overall-consistency filtering to retain ∼321.3k samples; the rightmost nodes show the retained set's source distribution. **Right:** The distribution of capability tags in the final dataset. As each sample is annotated with its top-3 relevant dimensions, the total volume of tags reaches ∼964k.

Table 1: **Hierarchical Capability Taxonomy.** We define seven Core Capabilities, each encompassing 3 specific Fine-grained Dimensions (21 total) to capture the micro-skills required for multimodal tasks. **Tag Ratio** reports each fine-grained dimension's share of tag slots in the retained dataset, where each sample contributes its top-3 dimensions.

Core Capability	Fine-grained Dimensions	Tag Ratio
Safety & Ethics (SE)	Harmful Content Detection	4.8%
	Bias & Fairness	3.1%
	Privacy & Personal Information	2.0%
Coarse Perception (CP)	Style & Quality	1.9%
	Scene & Topic	10.7%
	Emotion	3.6%
Fine-grained Perception (FP)	Attribute & Celebrity Recognition	4.8%
	Object Location	16.7%
	Object Counting	3.7%
Instance Reasoning (IR)	Single Instance Attribute	15.1%
	Cross-instance Comparison	3.6%
	Cross-instance Relation	6.4%
Logical Reasoning (LR)	Diagram Reasoning	3.2%
	Code & Sequence Reasoning	0.1%
	Common Reasoning	13.3%
Science & Technology (ST)	Natural Science	3.2%
	Engineering	0.9%
	Geography & Earth Science	0.8%
Mathematics (MA)	Numeric Calculation	1.1%
	Geometry	0.4%
	Statistical Analysis	0.7%

sions are consistent across models, ensuring stable and reproducible capability tagging.

Given the agreed target dimensions, we then prompt each judge to compare the *chosen* and *rejected* responses along these dimensions and output an overall preference (see Figure 9 in Appendix D). A sample is retained only if (i) the overall preference is *consistent across the three judges*, and (ii) the consensus preference aligns with the original ground truth (i.e., *chosen* is preferred). As visualized in the Sankey diagram (Figure 2, Left), this multi-model fine-grained overall-consistency check refines the initial pool of 414.2k into a robust set of 321.3k samples, corresponding to a 77.6% retention rate.

2.3 Capability Distribution Analysis

Recognizing that real-world visual tasks are rarely one-dimensional, we annotate each sample with its top-3 relevant fine-grained dimensions. This results in a total of approximately 964k capability tags, offering a comprehensive view of the dataset's composition. As shown in the sunburst chart (Figure 2, Right), the distribution is led by Fine-grained Perception (25.2%) and Instance Reasoning (25.1%), indicating that the majority of alignment scenarios focus on visual grounding and reasoning about object attributes or relations. Furthermore, Logical Reasoning (16.7%) and Coarse Perception (16.2%) remain substantial, ensuring the model is trained on multi-step tasks such as diagram analysis, commonsense deduction, and holistic understanding. Notably, Safety & Ethics accounts for 9.9% of the tags, providing explicit signals for harmful content, fairness, and privacy. Domain-specific categories like Science & Technology (4.9%) and Mathematics (2.1%) are preserved to provide crucial supervision for expert-level capabilities. This balanced yet comprehensive distribution ensures our dataset provides precise training signals across the full spectrum of multimodal capabilities.

3 Methodology

In this section, we present **VL-MDR** (Vision-Language Multi-Dimensional Reward), a framework designed to decompose the holistic evaluation of multimodal responses into interpretable, fine-grained components. As illustrated in Figure 3,

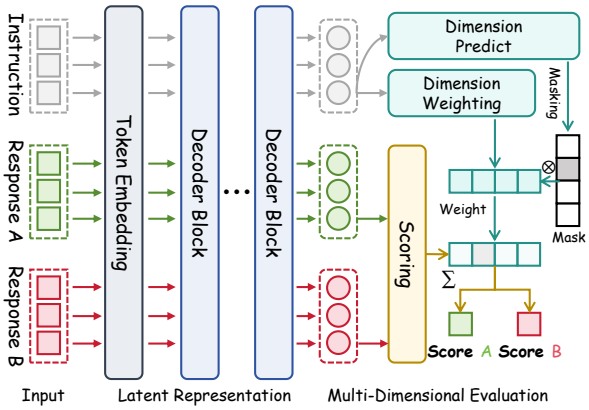

Figure 3: **Overview of the VL-MDR framework.** As shown in the diagram, the model processes the *Instruction* and candidate *Responses* (A and B) through a decoupled architecture. The backbone extracts distinct representations to feed three specialized heads: *Dimension Predict* (identifying relevant criteria based on the instruction), *Dimension Weighting* (assigning importance), and *Scoring* (evaluating quality). These components are aggregated to compute the final reward signal.

our method decouples the evaluation criteria from response generation to simulate a human-like evaluation process: identifying *what* matters via the instruction path, while assessing *how* good the response is and determining *how much* each aspect contributes via the response path.

3.1 Problem Formulation

We consider the problem of reward modeling under a pairwise preference setting. Let x denote a multimodal instruction containing both visual and textual inputs. For each instruction, we are given a pair of candidate responses (y_A, y_B). We define a set of K evaluation dimensions, denoted as $\mathcal{D} = \{1, \ldots, K\}$. The training dataset consists of tuples $(x, y_A, y_B, \mathbf{z}, \mathbf{p}, o)$, where: $\mathbf{z} \in \{0,1\}^K$ is the *dimension relevance label*, where $z_k = 1$ indicates that dimension k is relevant to the instruction x. $\mathbf{p} \in \{1, 0, -1\}^K$ represents the *fine-grained preference* as a *sparse* label: for the k-th dimension, $p_k = 1$ implies y_A is preferred over y_B, $p_k = -1$ implies y_B is preferred, and $p_k = 0$ indicates a tie; importantly, p_k is defined/used only when $z_k = 1$. $o \in \{1, 0, -1\}$ represents the *overall preference*, following the same semantics as $\mathbf{p}$.

3.2 Model Architecture

VL-MDR is built upon a pre-trained VLM backbone. A core design principle of our framework is the *Query-Response Decoupling*, which posits that the evaluation criteria should be determined solely by the instruction, while the performance assessment depends on the response content.

Given the input sequence, we extract two distinct pooled representations from the backbone's last hidden states to support our decoupled evaluation design. Specifically, we obtain the **instruction representation** $\mathbf{h}^q$ from the last token of the query to capture the user's intent independent of the generated content. Simultaneously, we extract the **response representation** $\mathbf{h}^r$ from the last token of the generated response. It is worth noting that in standard decoder-only transformers, the causal self-attention mechanism ensures that each token attends to all preceding tokens. Therefore, the final token's hidden state naturally aggregates the comprehensive information of the entire sequence, effectively encoding the full context of both the instruction and the response without the need for redundant concatenation. Based on these features, we introduce three parallel projection heads:

Dimension Prediction. To determine the active evaluation criteria, a dimension prediction head ϕ maps the instruction representation to relevance logits $\mathbf{l} \in \mathbb{R}^K$:

$$\mathbf{l} = \phi(\mathbf{h}^q). \tag{1}$$

During inference, we first obtain relevance probabilities $\sigma(\mathbf{l})$ and then generate a *Top-k* binary mask $\mathbf{m} \in \{0,1\}^K$ that keeps only the k dimensions with the largest probabilities.

Fine-Grained Scoring. To assess the quality of the response across all potential dimensions, a scoring head ψ maps the response representation to scalar scores $\mathbf{s} \in \mathbb{R}^K$:

$$\mathbf{s} = \psi(\mathbf{h}^r). \tag{2}$$

Adaptive Aggregation. To synthesize a comprehensive reward, we employ a weighting head ω that predicts raw importance logits $\mathbf{u} = \omega(\mathbf{h}^q)$. We propose a masked normalization mechanism to compute the final dimension weights $\boldsymbol{\alpha} \in \mathbb{R}^K$:

$$\alpha_k = \frac{m_k \cdot \exp(u_k)}{\sum_{j=1}^{K} m_j \cdot \exp(u_j)}. \tag{3}$$

This ensures that irrelevant dimensions (where $m_k = 0$) are strictly excluded from the aggregation. The final holistic reward $R(x, y)$ is obtained by the weighted sum of the dimension scores:

$$R(x, y) = \sum_{k=1}^{K} \alpha_k \cdot \sigma(s_k). \tag{4}$$

3.3 Hierarchical Multi-Objective Optimization

We train VL-MDR using a hierarchical objective that jointly optimizes criteria selection, fine-grained ranking, and overall alignment.

Dimension Prediction Loss. Since the relevance of evaluation dimensions is intrinsic to the instruction x, both responses y_A and y_B share the same ground-truth labels $\mathbf{z}$. We optimize the dimension predictor using the binary cross-entropy loss:

$$\mathcal{L}_{\text{dim}} = -\frac{1}{K}\sum_{k=1}^{K}[z_k \log \sigma(l_k) + (1-z_k)\log(1-\sigma(l_k))]. \tag{5}$$

Unified Pairwise Ranking Loss. For preference learning, we must handle both strict preference (ranking) and equality (tie) scenarios. We formulate a unified pairwise loss function $\ell(\delta, y)$ for a score difference δ and a label $y \in \{1, 0, -1\}$:

$$\ell(\delta, y) = \mathbb{I}[y \neq 0]\cdot\max(0, \xi - y\cdot\delta) + \mathbb{I}[y=0]\cdot\delta^2, \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function and ξ is the margin. The first term enforces separation when a clear preference exists, while the second term (MSE) encourages score alignment when the responses are tied.

We apply this unified loss to both the fine-grained dimensions and the overall reward. For the fine-grained scores, let $\Delta s_k = s_{A,k} - s_{B,k}$. Since fine-grained preferences are only labeled on relevant (top-3) dimensions, the ranking loss is averaged over labeled dimensions:

$$\mathcal{L}_{\text{rank}} = \frac{1}{\sum z_k}\sum_{k=1}^{K} z_k \cdot \ell(\Delta s_k, p_k). \tag{7}$$

Similarly, for the overall reward difference $\Delta R = R(x, y_A) - R(x, y_B)$, the overall loss is defined as:

$$\mathcal{L}_{\text{overall}} = \ell(\Delta R, o). \tag{8}$$

Total Objective. The final training objective is a weighted combination of the three components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{dim}}\mathcal{L}_{\text{dim}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{overall}}\mathcal{L}_{\text{overall}}. \tag{9}$$

This multi-task formulation ensures that VL-MDR not only aligns with human preferences globally but also provides mathematically grounded justifications through its internal dimensional predictions.

4 Experiment

In this section, we conduct extensive experiments to evaluate the performance and interpretability of VL-MDR. Specifically, we aim to answer the following research questions:

- **RQ1 (Effectiveness):** How does VL-MDR compare with state-of-the-art discriminative and generative reward models across diverse multimodal benchmarks?
- **RQ2 (Granularity):** Does the proposed fine-grained dimensional supervision offer superior preference modeling performance compared to coarse or monolithic scalar rewards?
- **RQ3 (Efficiency):** To what extent does VL-MDR reduce computational overhead compared to generative approaches while maintaining competitive performance?
- **RQ4 (Utility):** Can VL-MDR be used to construct preference pairs for DPO-based alignment, improving the reliability and safety of downstream LVLM generations?

4.1 Implementation Details and Evaluation

We use `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025b) as the backbone, freezing the vision tower and projector while fine-tuning the LLM and reward heads for 2 epochs with a learning rate of 5×10^{-7}, cosine scheduling, and 0.1 warmup. Training is performed on 8 NVIDIA H100 GPUs with per-device batch size 2 and 4 gradient accumulation steps (global batch size 64). For VL-MDR, we set the $k = 3$, ranking margin $\xi = 0.3$, and reward-head dropout 0.1, with $\lambda_{dim} = \lambda_{rank} = \lambda_{overall} = 1.0$. We evaluate on VL-RewardBench (Ruan et al., 2025), Multimodal RewardBench (Yasunaga et al., 2025), and MM-RLHF-Reward Bench (Zhang et al., 2025a), reporting Overall/Macro Average Accuracy on VL-RewardBench, and Acc/Acc+ on MM-RLHF-Reward Bench (Acc+ requires correctly ranking all response pairs within each sample).

4.2 RQ1: Overall Effectiveness on Multimodal Reward Benchmarks

We first assess the overall effectiveness of VL-MDR by comparing it with state-of-the-art proprietary judges, open-source LVLMs, and representative generative/discriminative reward models on three diverse benchmarks (Table 2 and Table 4). Across VL-RewardBench, VL-MDR consistently ranks among the strongest open-source reward

Table 2: **VL-RewardBench.** Performance comparison of our reward model (VL-MDR) with existing open-source and proprietary counterparts.

Models	#Param	General	Hallucination	Reasoning	Overall Acc	Macro Acc
Proprietary Models						
GPT-4o-mini	-	41.70	34.50	58.20	41.50	44.80
GPT-4o	-	49.10	67.60	70.50	65.80	62.40
Gemini-1.5-Flash	-	47.80	59.60	58.40	57.60	55.30
Gemini-1.5-Pro	-	50.80	72.50	64.20	67.20	62.50
Claude 3.5 Sonnet	-	43.40	55.00	62.30	55.30	53.60
Claude 3.7 Sonnet	-	68.08	70.70	60.81	66.31	66.53
Open-Source Models						
LLaVA-OneVision-7B	7B	32.20	20.10	57.10	29.60	36.50
Qwen2-VL-7B	7B	31.60	19.10	51.10	28.30	33.90
Qwen2.5-VL-7B	7B	34.25	21.76	54.57	31.92	36.86
InternVL3-8B	8B	60.22	43.93	62.46	51.00	55.54
Llama-3.2-11B	11B	33.30	38.40	56.60	42.90	42.80
Qwen2-VL-72B	72B	38.10	32.80	58.00	39.50	43.00
Qwen2.5-VL-72B	72B	48.07	46.73	63.41	51.16	52.73
InternVL3-78B	78B	69.61	52.47	64.35	57.98	62.15
Llama-3.2-90B	90B	42.60	57.30	61.70	56.20	53.90
Generative RMs						
LLaVA-Critic	7B	54.60	38.30	59.10	41.20	44.00
UnifiedReward	7B	76.24	58.61	64.98	62.79	66.61
UnifiedReward-Think	7B	77.35	72.50	65.62	71.45	71.82
Discriminative RMs						
Skywork-VL-Reward	7B	65.75	79.84	60.88	72.98	68.82
IXC-2.5-Reward	7B	80.11	65.29	60.25	66.16	68.55
MM-RLHF-Reward	7B	45.04	50.45	57.55	50.15	51.01
VL-MDR (Ours)	7B	71.27	75.17	69.09	**73.06**	**71.84**

models and achieves the best overall balance across categories, indicating reliable preference identification beyond skewed task distributions. On Multimodal RewardBench, VL-MDR remains competitive under broad evaluation axes (e.g., general correctness, knowledge, reasoning, safety, and VQA), demonstrating robust cross-domain generalization rather than overfitting to a single capability. Finally, on the more challenging MM-RLHF-Reward Bench, VL-MDR shows clear advantages on strict ranking criteria, reflecting improved sensitivity to subtle preference differences and hard cases. Overall, the results verify that VL-MDR delivers strong and stable reward modeling performance across heterogeneous multimodal settings, while remaining competitive with both generative and discriminative baselines.

4.3 RQ2: Impact of Granularity and Dimensional Supervision

To identify what drives VL-MDR, we conduct controlled ablations. All variants are trained on the same 200k randomly sampled subset with identical training setup, and evaluated on VL-RewardBench.

Table 3: **Ablation study on supervision granularity and gating**. **Gran.** denotes the number of supervision dimensions; $\mathcal{L}_{\text{rank}}$ indicates whether per-dimension ranking supervision is used; **Gate** indicates instruction-aware dimension gating. Results are reported on VL-RewardBench (Overall/Macro accuracy).

Method	Config			Results	
	Gran.	$\mathcal{L}_{\text{rank}}$	Gate	Overall	Macro
Scalar	1	–	–	64.55	60.87
Implicit	21	–	✓	68.72	65.94
Coarse-7D	7	✓	✓	67.12	64.20
Fine w/o Gate	21	✓	–	69.77	68.51
VL-MDR	**21**	✓	✓	**70.81**	**69.96**

From the results in Table 3, we draw the following conclusions. Explicit dimensional supervision matters: removing per-dimension ranking supervision while keeping the 21-head structure (Implicit) improves over the Scalar baseline but still falls short of VL-MDR, suggesting that structural capacity alone is insufficient. Finer granularity also helps: with supervision enabled, Fine w/o Gate outperforms Coarse-7D on both Overall and Macro, indicating that 21D labels provide more specific

Table 4: **Performance on Multimodal RewardBench (Left) and MM-RLHF-Reward Bench (Right).** Comparison of our reward model (VL-MDR) with existing open-source and proprietary counterparts.

Model	#Param	Multimodal RewardBench								MM-RLHF-Reward Bench						
		Overall	General		Know.	Reasoning		Safety	VQA	Mcq	Long	Short	Safety	Video	Acc	Acc+
			Corr.	Pref.		Math	Code									
Proprietary Models																
GPT-4o	-	70.8	62.6	69.0	72.0	67.6	62.1	74.8	87.2	64.3	78.4	44.1	56.3	40.0	58.2	26.1
Claude 3.5 Sonnet	-	71.5	62.6	67.8	73.9	68.6	65.1	76.8	85.6	64.3	67.6	55.9	65.6	60.0	62.9	26.1
Claude 3.7 Sonnet	-	71.9	58.4	60.7	78.1	76.3	71.3	72.0	86.8	66.7	91.9	91.2	87.5	76.0	82.4	65.2
Generative RMs																
LLaVA-Critic	7B	54.1	50.1	53.2	51.3	49.2	49.3	78.0	52.4	42.9	75.7	47.1	53.1	48.0	53.5	23.9
UnifiedReward	7B	61.6	63.1	58.4	55.9	64.4	44.7	50.4	77.2	59.5	83.8	58.8	59.4	84.0	68.2	37.0
UnifiedReward-Think	7B	66.7	63.9	68.7	61.1	65.8	57.7	55.3	79.5	73.8	94.6	79.4	62.5	84.0	78.8	54.4
Discriminative RMs																
Skywork-VL-Reward	7B	65.7	64.5	62.4	54.3	70.8	44.7	62.0	83.5	50.0	94.6	70.6	68.8	72.0	70.6	58.7
IXC-2.5-Reward	7B	66.6	60.7	64.2	56.8	63.0	50.5	89.9	81.1	52.4	91.9	67.7	62.5	88.0	71.2	50.0
MM-RLHF-Reward	7B	67.1	61.7	67.5	54.3	58.4	57.9	92.9	76.8	83.3	97.3	73.5	68.8	88.0	82.3	63.0
VL-MDR (Ours)	7B	69.0	69.8	68.7	67.8	78.2	45.2	51.2	84.4	83.3	91.9	79.4	81.3	92.0	85.3	69.6

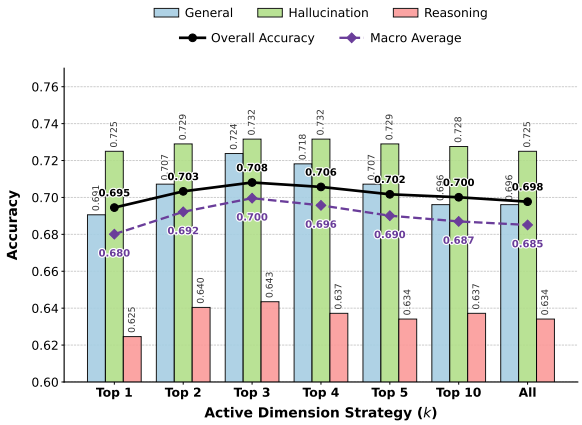

Figure 4: **Impact of Active Dimension Count (k).** Performance peaks at $k = 3$, demonstrating that selecting a focused set of relevant dimensions strikes an optimal balance: it filters out noise from irrelevant criteria while retaining sufficient evaluation signals.

guidance than coarse capability groups. Gating further improves performance: Fine w/o Gate underperforms VL-MDR, and the Top-k sweep in Figure 4 peaks at $k = 3$ before declining as more dimensions are included, showing that selecting a moderate number of relevant dimensions is optimal while aggregating too many introduces noise.

4.4 RQ3: Computational Efficiency and Parameter Analysis

We evaluate the computational efficiency of VL-MDR on VL-RewardBench by comparing its parameter overhead and inference cost with representative reward models. For inference cost, we run each evaluation five times on a single NVIDIA H20 GPU using HuggingFace `transformers` and report the average GPU-hours in Table 5. VL-MDR adds a lightweight multi-dimension reward

Table 5: **Computational cost and performance trade-off on VL-RewardBench.** We report the total inference cost (in GPU Hours) required to evaluate the full benchmark alongside the Macro Acc. While Generative RMs suffer from high latency due to decoding, VL-MDR maintains the efficiency of Discriminative RMs.

Method	Add. Param. (M)	Cost (GPU Hours) ↓	Macro Acc. (%) ↑
`Generative RMs`			
LLaVA-Critic (7B)	N/A	0.423 ± 0.09	44.0
UnifiedReward (7B)	N/A	0.399 ± 0.08	66.6
UnifiedReward-Think (7B)	N/A	0.518 ± 0.11	71.8
`Discriminative RMs`			
Skywork-VL (7B)	0.0036	0.216 ± 0.03	68.8
IXC-2.5-Reward (7B)	603.98	0.273 ± 0.05	68.6
MM-RLHF-Reward (7B)	12.85	0.218 ± 0.03	51.0
VL-MDR (Ours)	**17.86**	$\mathbf{0.218 \pm 0.04}$	**71.8**

head on top of `Qwen2.5-VL-7B`, including a Dimension Predictor f_{dim}, a Scoring Module f_{score}, and a Weighting Module f_{weight}. This introduces only **17.86M** additional parameters, i.e., **0.25%** of the 7B backbone, leading to negligible deployment overhead compared to standard scalar discriminative reward models. For generative RMs, we report *Add. Param.* as N/A since they do not introduce an explicit reward head and are obtained by fine-tuning the full backbone. In terms of inference, VL-MDR avoids autoregressive decoding and computes dimension-wise rewards in a single forward pass, whereas GenRMs typically generate critique or score tokens and thus incur higher cost as the output length grows; compared to scalar discriminative RMs, the extra computation of the MD head is marginal because it only adds a few MLP layers on top of the same backbone activations. Appendix B provides the detailed parameter accounting.

Table 6: **Image Understanding DPO Comparison.** We compare our method with generative and discriminative RMs for DPO based on LLaVA-OneVision-7B.

Method	LLaVABench	WildVision	LLaVABenchWilder	MMHal
Generative RMs				
OV-7B	90.3	54.9	67.8	3.2
w/ LLaVA-Critic	100.3	67.3	71.6	3.9
w/ UnifiedReward	101.4	67.8	75.0	4.0
w/ UnifiedReward-Think	101.8	68.3	77.5	4.2
Discriminative RMs				
w/ Skywork-VL	101.9	68.2	77.6	4.2
w/ IXC-2.5-Reward	101.6	67.9	77.0	4.1
w/ MM-RLHF-Reward	101.4	67.1	76.2	4.1
w/ VL-MDR (Ours)	**101.9**	**68.3**	**77.9**	**4.2**

4.5 RQ4: Utility for DPO Alignment

Following the pipeline in UnifiedReward (Wang et al., 2025c), we construct preference pairs from the LLaVA-RLHF dataset (Sun et al., 2023) using VL-MDR annotations. These pairs are subsequently utilized to fine-tune LLaVA-OneVision-7B (OV-7B) via DPO. All generation parameters and training configurations align strictly with (Wang et al., 2025c), with detailed hyperparameters provided in Appendix C. As shown in Table 6, VL-MDR achieves the best or comparable performance to generative and discriminative baselines across all benchmarks. This indicates that VL-MDR's fine-grained dimensional supervision effectively filters out hallucinations and subtle errors that scalar reward models might miss, providing higher-quality signals for alignment.

5 Related Works

5.1 Vision-Language Reward Modeling

Current research on vision-language reward modeling is primarily categorized into discriminative and generative paradigms. Discriminative RMs, such as Skywork-VL Reward (Wang et al., 2025a) and InternLM-XComposer2.5-Reward (Zang et al., 2025), typically predict a single scalar reward via a linear head. Although this formulation facilitates efficient large-scale ranking, it functions as a "black box" by compressing diverse error types into a monolithic score, thereby obscuring the rationale behind preferences. Conversely, Generative RMs (or "VLM-as-a-Judge"), including Prometheus-Vision (Lee et al., 2024) and LLaVA-Critic (Xiong et al., 2025), provide natural language critiques to explain their judgments, with recent models like LLaVA-Critic R1 (Wang et al., 2025b) and UnifiedReward (Wang et al., 2025c) further incorporating explicit reasoning chains for verification. However, despite offering superior interpretability, these generative methods are computationally prohibitive due to decoding latency and remain susceptible to inherent biases, such as length preference, which limits their scalability in practical alignment pipelines.

5.2 Vision-Language Preference Data

Recent VLM alignment has shifted from supervised fine-tuning to preference learning to improve safety and reduce hallucinations. Early works like VLFeedback (Li et al., 2023) and LLaVA-Critic (Xiong et al., 2025) used proprietary models to generate ranking data for general alignment. However, RLHF-V (Yu et al., 2024) demonstrated that dense, segment-level corrections are more efficient than holistic rankings for fixing fine-grained visual errors. RLAIF-V (Yu et al., 2025) further improved this by using open-source models to generate high-quality feedback, reducing reliance on proprietary judges. Finally, datasets like WildVision (Lu et al., 2024) and Vision Arena (Chou et al., 2025) incorporate real-world user interactions to better reflect practical usage scenarios. Nevertheless, these existing resources predominantly rely on holistic rankings or sparse textual corrections, lacking the systematic, fine-grained dimensional annotations required to explicitly disentangle orthogonal failure modes (e.g., perception vs. reasoning) for interpretable reward modeling.

6 Conclusion

This paper addresses the critical trade-off in vision-language reward modeling: generative models provide clear reasoning but are slow, while discriminative models are fast but act as opaque black boxes. We proposed VL-MDR, a framework that solves this by dynamically decomposing evaluation into specific, interpretable dimensions. By predicting which criteria matter for each instruction, VL-MDR achieves the clarity of a judge with the speed of a standard scorer. To support this approach, we curated a dataset of 321k preference pairs annotated across 21 fine-grained dimensions. Experiments show that VL-MDR consistently outperforms open-source baselines on benchmarks like VL-RewardBench. Furthermore, we demonstrated that VL-MDR-constructed preference pairs effectively enable DPO alignment to improve LVLM reliability and safety. We hope this framework offers a scalable and transparent solution for future VLM alignment.

Acknowledgements

This research was partially supported by the National Key Research and Development Program of China under Grant No. 2024YFF0907802, the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C02120 and No. 2024SSYS0026, and the Transvascular Implantation Devices Research Institute (TIDRI) under Grant No. KY052025003.

Limitations

While our results are promising, several limitations should be noted. First, part of our supervision relies on automated multi-judge annotations and filtering; although scalable and consistent, these signals may still reflect biases or gaps of current judge models, and additional human verification could further improve reliability. Second, our 21-dimension hierarchical taxonomy offers a structured view of multimodal quality, but it is not necessarily complete and may miss criteria that matter in specific domains or contexts. Third, our dimension gating and aggregation involve practical design choices (e.g., the number of active dimensions); further study is needed to better understand robustness and calibration under distribution shifts. Finally, our experiments focus on image-text tasks and a limited set of benchmarks and backbones; future work should test broader modalities and more diverse real-world settings to better assess generalization.

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, 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. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E Gonzalez, and Wei-Lin Chiang. 2025. Visionarena: 230k real world user-vlm conversations with preference labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3877–3887.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pages 11286–11315.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. 2025. Vlrmbench: A comprehensive and challenging benchmark for vision-language reward models. *arXiv preprint arXiv:2503.07478*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, et al. 2025a. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*.

Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. 2025b. Llava-critic-r1: Your critic model is secretly a strong policy model. *arXiv preprint arXiv:2509.00676*.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025c. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628.

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

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. 2025a. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. 2025b. Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19867–19878.

A Dataset Analysis

This appendix presents analysis of the fine-grained judgment results, revealing patterns about dimension consistency, relationships, and difficulty.

Dimension Abbreviations: Style: Style & Quality, Scene: Scene & Topic, Emotion: Emotion, Celebrity: Attribute & Celebrity, Location: Object Location, Counting: Object Counting, Attribute: Single Instance Attribute, Comparison: Cross-instance Comparison, Relation: Cross-instance Relation, Diagram: Diagram Reasoning, Code: Code & Sequence Reasoning, Common: Common Reasoning, Natural: Natural Science, Engineering: Engineering, Geo: Geo & Earth Science, Calc: Numeric Calculation, Geometry: Geometry, Statistics: Statistical Analysis, Harmful: Harmful Content, Bias: Bias & Fairness, Privacy: Privacy.

A.1 Dimension Consistency

We analyze the consistency between per-dimension judgments and the overall judgment. For each sample, we count how many of its three fine-grained dimension judgments agree with the overall preference label. Table 7 presents the results.

Consistency Level	Count	Percentage
All 3 dimensions match overall	266,444	64.3%
2 dimensions match overall	108,259	26.1%
1 dimension matches overall	37,252	9.0%
0 dimensions match overall	2,177	0.5%
Total	414,132	100%

Table 7: Dimension-Overall Consistency

As shown, 64.3% of samples have all three dimensions aligned with the overall judgment, indicating strong internal consistency. Only 0.5% show complete disagreement. This high consistency rate validates the reliability of our fine-grained annotation approach.

A.2 Dimension Co-occurrence

We analyze which dimensions tend to appear together in the same samples. Figure 6 shows the top co-occurring dimension pairs.

The most frequent pair is Location + Attribute (125,535 samples), suggesting that localization and attribute description are naturally coupled tasks. Other notable pairs include Scene + Location (visual grounding) and Common + Location (reasoning about positioned objects).

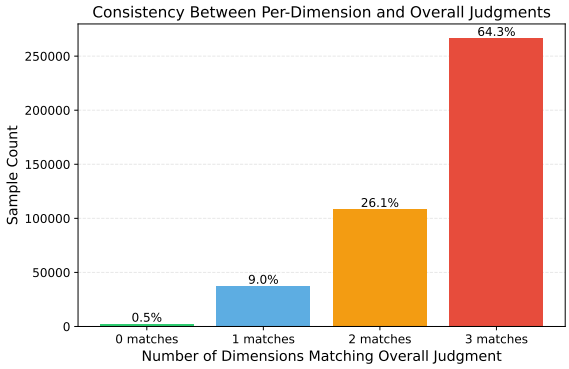

Figure 5: Dimension-overall consistency.

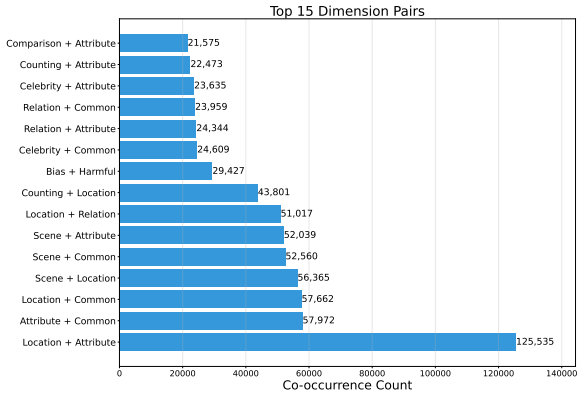

Figure 6: Top dimension co-occurrence pairs.

A.3 Dimension Difficulty

Table 8 ranks dimensions by their tie rate, which serves as a proxy for judgment difficulty. Higher tie rates indicate more ambiguous or subjective dimensions.

Dimension	Tie Rate	Count
Celebrity	44.7%	61,897
Calc	19.2%	11,049
Location	18.3%	172,433
Counting	17.4%	44,919
Attribute	4.7%	182,596
Geometry	0.7%	5,148

Table 8: Dimension Difficulty (Tie Rate)

Celebrity recognition has the highest tie rate (44.7%), suggesting high ambiguity in defining "celebrity" and edge cases in public figure identification. In contrast, mathematical dimensions (Geometry, Calc) have near-zero tie rates, indicating objective criteria and high inter-judge agreement.

B Detailed Parameter Analysis

In this section, we present the detailed breakdown of the Multi-Dimensional (MD) Reward Head ar-

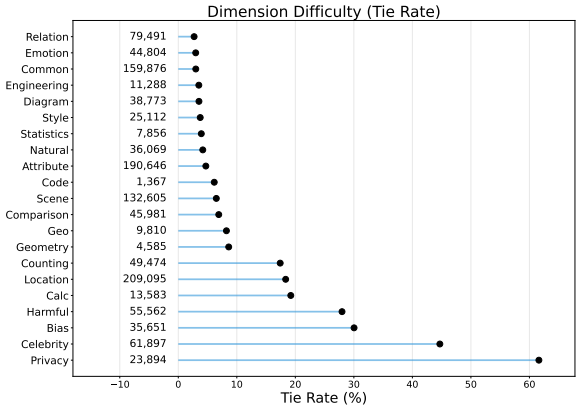

Figure 7: Dimension difficulty ranked by tie rate.

chitecture and parameter calculations. The backbone model is `Qwen2.5-VL-7B`, with a hidden size of $d_{in} = 3584$. The output dimension corresponds to the $K = 21$ fine-grained evaluation dimensions defined in our taxonomy. The MD Reward Head utilizes standard Linear layers (bias=False) for all transformations. The detailed layer configurations and parameter counts are listed in Table 9.

C Details of DPO Experiment

In this section, we provide a comprehensive description of the implementation details, training configurations, and evaluation benchmarks for the DPO experiments discussed in the main text.

C.1 Preference Data Construction

We adhere strictly to the data construction pipeline proposed in UnifiedReward to ensure a fair and rigorous comparison. The process utilizes image-question pairs from the LLaVA-RLHF dataset as the initial prompt source. For every prompt in the dataset, we sample six distinct candidate responses from the policy model, LLaVA-OneVision-7B (Li et al., 2024b), using a sampling temperature of 0.7 to encourage diversity in the generated outputs.

Once the candidates are generated, we employ VL-MDR to assign a comprehensive quality score to each response. This score is derived from the weighted aggregation of our fine-grained dimensional predictions. To maximize the preference signal, we form a training pair for each prompt by designating the response with the highest reward score as the chosen sample and the response with the lowest score as the rejected sample. This filtering process results in a high-quality preference dataset comprising approximately 14,000 pairs.

C.2 Training Configuration

The DPO fine-tuning is conducted on the LLaVA-OneVision-7B backbone using the LLaVA-NeXT codebase. We adopt the specific hyperparameters reported in the UnifiedReward study. The model is trained for three epochs using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We utilize a global batch size of 128, achieved through gradient accumulation, and set the learning rate to 5×10^{-7} with a cosine decay scheduler and a warmup ratio of 0.03. The Kullback-Leibler (KL) penalty coefficient β for DPO is set to 0.1. All experiments are performed on a cluster of 8 NVIDIA H100 GPUs.

C.3 Evaluation Benchmarks

We evaluate the aligned models using the VLMEvalKit across four diverse benchmarks to assess various capabilities. LLaVABench (Liu et al., 2023) serves as a standard metric for general visual reasoning in diverse indoor and outdoor scenes. To appraise performance in more complex and uncontrolled environments, we employ LLaVABench-Wilder (Li et al., 2024a). For a proxy of real-world user preference, we use WildVision, which is derived from the WildVision-Arena (Lu et al., 2024) and correlates well with Chatbot Arena Elo ratings. Finally, to specifically verify the effectiveness of our model in mitigating hallucinations, we report results on MMHal-Bench (Sun et al., 2023), where a higher score indicates a lower rate of hallucinatory content.

D Prompt Templates for Data Construction

To facilitate reproducibility, we strictly follow the prompts illustrated below. Figure 8 presents the instruction used for identifying the top-3 relevant dimensions, while Figure 9 displays the prompt used for the fine-grained comparison and overall judgment.

Table 9: Detailed layer-wise parameter breakdown for the VL-MDR Reward Head components.

Component	Layer Type	Shape (Input $\rightarrow$ Output)	Params
Dimension Predictor (f_{dim})	Linear	$3584 \rightarrow 1024$	3,670,016
	Linear	$1024 \rightarrow 512$	524,288
	Linear	$512 \rightarrow 512$	262,144
	Linear	$512 \rightarrow 21$	10,752
	Subtotal		*4.47 M*
Scoring Module (f_{score})	Linear	$3584 \rightarrow 2048$	7,340,032
	Linear	$2048 \rightarrow 1024$	2,097,152
	Linear	$1024 \rightarrow 1024$	1,048,576
	Linear	$1024 \rightarrow 512$	524,288
	Linear	$512 \rightarrow 21$	10,752
	Subtotal		*11.02 M*
Weighting Module (f_{weight})	Linear	$3584 \rightarrow 512$	1,835,776
	Linear	$512 \rightarrow 512$	262,144
	Linear	$512 \rightarrow 512$	262,144
	Linear	$512 \rightarrow 21$	10,752
	Subtotal		*2.37 M*
Total			17.86 M

```
# ROLE
You are an expert Multimodal Benchmark Evaluator.
# TASK
Your task is to analyze a given Image-Text pair. The Text is a question that can only be answered by analyzing the Image.
Your goal is to classify this pair by identifying the **3** (three) most relevant **Detailed Axes** required to answer the question. You must base your classification *strictly* on the definitions provided below.
---
# DETAILED AXES DEFINITIONS
{axes_definitions}
# Question to Analyze
{query}
-----
# INSTRUCTIONS
You will be given an image and text. Follow these steps precisely:
1. **Analyze:** Read the Text (question) and carefully examine the Image.
2. **Reason:** Determine the *specific micro-skills* that are essential to answer the question. (e.g., "To answer this, I must first locate the cat [fp_object_location], then count the books [fp_object_counting], and finally compare the cat's size to the books [ir_cross_instance_comparison].")
3. **Classify:** From the list of 21 Detailed Axes, select the **3** codes that are most essential to the task.
4. **Format:** You *must* provide your answer in the exact JSON format specified below. Do not include any other text or explanations outside the JSON structure.
```

Figure 8: The prompt template used for **Visual-Aware Dimension Prediction**. The model is instructed to analyze the image-text pair and select the top-3 relevant fine-grained axes from the defined taxonomy.

```
# ROLE
You are an expert Multimodal Response Judge.
# TASK
Given an Image-Question pair and two candidate assistant responses (A and B), you must:
1) Judge which response is better on EACH target dimension first.
2) Then provide an overall judgement.
# TARGET DIMENSIONS
{target_dimensions}
# DIMENSION DEFINITIONS
{dimension_definitions}
# INPUT
Question:
{query}
Response A:
{response_a}
Response B:
{response_b}
# INSTRUCTIONS
- The order of responses is randomized; do NOT assume A is preferred.
- Use the image as evidence when judging correctness.
- For each target dimension:
  - give integer scores 0-10 for A and B (higher is better)
-  choose winner: "A", "B", or "tie"
-  avoid ties as much as possible: only output "tie" if A and B are truly indistinguishable on that dimension.
-  if both are wrong, pick the less wrong one; if uncertain, make your best guess and still pick a winner.
-  winner must be consistent with scores: A if score_a > score_b, B if score_b > score_a, and "tie" only if equal.
-  provide a short rationale focused on that dimension
- For overall:
-  primarily consider the target dimensions.
-  avoid "tie" unless overall quality is truly indistinguishable; use the same score-consistency rule as above.
- You must output exactly one judgement per target dimension, and each "dimension" must be one of TARGET DIMENSIONS.
```

Figure 9: The prompt template used for **Fine-Grained Response Comparison**. The model evaluates two candidate responses on the specific target dimensions identified in the previous step before providing an overall preference.