A Survey of Verifiers for Multimodal Large Language models: Judge, Critic, and Reward model

Peng Kuang
Zhejiang University
pengkuang@zju.edu.cn

Xiaoyu Han
University of Illinois Urbana-Champaign
xiaoyu27@illinois.edu

Abstract

The rapid advancement of Multimodal Large Language Models (MLLMs) has revolutionized tasks requiring visual understanding and reasoning. However, these models frequently suffer from hallucinations, flawed logical reasoning, and misalignment with human intent. To mitigate these issues, the role of the *verifier*—an external or internal mechanism that assesses the quality, correctness, and safety of model outputs—has become paramount. This survey provides a comprehensive overview of verifiers for MLLMs, unifying distinct concepts such as Reward Models (RMs), Critics, and Judges under a single taxonomy. We categorize verifiers based on their output modality: *discriminative verifiers*, which provide scalar signals for outcome or process supervision, and *generative verifiers*, which offer interpretable critiques and reasoning traces. Furthermore, we analyze their utility across the entire model lifecycle, including training data filtering, automated evaluation (LLM-as-a-judge), reinforcement learning from human feedback (RLHF), and test-time scaling. By highlighting the shift from black-box scoring to transparent, grounded verification, this paper outlines the current landscape and future directions for building robust, self-correcting multimodal systems.

1 Introduction

The integration of vision and language has propelled Multimodal Large Language Models (MLLMs) to the forefront of artificial intelligence, enabling capabilities that range from detailed image captioning to complex visual reasoning and embodied navigation. Despite their impressive performance [1, 2, 3, 4], MLLMs remain prone to critical reliability issues, notably "hallucination"—where generated text contradicts visual evidence [5]—and logical inconsistencies in multi-step reasoning. As models scale, relying solely on Supervised Fine-Tuning (SFT) has proven insufficient for ensuring alignment with human preferences and factual grounding. Consequently, the field has increasingly turned to *verifiers*: auxiliary models or mechanisms designed to evaluate, critique, and guide the generation process.

In this survey, we define a "verifier" broadly to encompass any system that assesses MLLM outputs. This includes **Reward Models (RMs)** used in Reinforcement Learning from Human Feedback (RLHF) [6, 7], **Process Reward Models (PRMs)** that monitor intermediate reasoning steps [8, 9], and **Generative Critics** or **Judges** that produce natural language feedback [10, 11]. While early verifiers were predominantly discriminative scalar predictors, the landscape is shifting towards generative approaches that leverage the reasoning capabilities of LLMs to explain *why* an error occurred, thereby facilitating better credit assignment and self-correction.

Relation to Existing Surveys. While several recent surveys have addressed aspects of model alignment, they largely focus on unimodal domains or specific sub-tasks. For instance, Zhong et al. [12] focuses primarily on text-based RLHF taxonomy, while Zheng et al. [13] emphasizes mathematical reasoning over visual trajectories. Similarly, Venktesh et al. [14] is limited to inference-

Preprint.

time strategies, and Liu et al. [15] prioritizes logical reasoning over the critical multimodal issues of grounding and hallucination. In contrast, our work provides a holistic taxonomy covering the entire verifier lifecycle specifically tailored to the challenges of Multimodal Large Language Models.

To address the unique challenges of the modality gap and visual grounding, we structure our survey around the output modality of the verifier, which fundamentally dictates its interaction mechanism with the MLLM. We begin in Section 2 by examining **Discriminative Verifiers**, tracing the evolution from outcome-based scalar scoring to fine-grained process supervision that locates errors in reasoning trajectories. In Section 3, we explore the emerging paradigm of **Generative Verifiers**, evaluating how pair-wise ranking, point-wise critiques, and process-level reasoning traces foster interpretability and self-correction (System 2 thinking). Finally, Section 4 analyzes the practical application of these verifiers across the model lifecycle, detailing their critical roles in filtering training data, serving as automated judges for benchmarking, optimizing policies via RLHF, and enabling compute-heavy test-time scaling strategies. This structure aims to provide researchers with a clear roadmap for selecting and designing verifiers that enhance both the trustworthiness and capability of multimodal systems.

2 Discriminative Multimodal Verifiers

Discriminative verifiers output a scalar score representing the quality or correctness of a response, treating the generation capabilities of the model as secondary or utilizing a separate reward head.

2.1 Outcome Supervision

Reward modeling serves as the cornerstone of aligning Multimodal Large Language Models (MLLMs) with human preferences and ensuring the correctness of generated outputs. Early approaches primarily focused on *outcome supervision*, where a reward model evaluates the final response quality. Works like **InternLM-XComposer2.5-Reward** [16] and **Skywork-VL Reward** [7] pioneered this direction by training discriminative reward models on preference pairs to guide Reinforcement Learning from Human Feedback (RLHF). However, outcome-based supervision treats the reasoning process as a black box. While effective for general alignment, these methods often fail to locate intermediate logical errors or hallucinations in complex mathematical reasoning tasks, where a correct final answer can occasionally result from flawed reasoning (false positives).

2.2 Process Supervision

To address the opacity of outcome supervision, recent research has pivoted toward *process supervision*, which evaluates reasoning trajectories step-by-step. This paradigm has shown significant promise in text-based math reasoning and is now being adapted for multimodal contexts.

VisualPRM [9] and **URSA** [17] introduced discriminative Process Reward Models (PRMs) trained via Monte Carlo Tree Search (MCTS) derived labels. These models assign a scalar probability score to each step, guiding inference-time search algorithms like Best-of-N or Tree Search. Similarly, **Athena** [18] demonstrated that data-efficient PRMs could be trained by leveraging error-injection techniques. Despite their success, discriminative PRMs suffer from the "black-box verification" problem: they output a score without an explanation. These models struggle with fine-grained visual grounding errors and often exhibit biases toward simple question replication or outcome-based heuristics.

3 Generative Multimodal Verifiers

The primary bottleneck in traditional Reinforcement Learning from Human Feedback (RLHF) is the "alignment tax," where models optimized for scalar rewards may exploit the reward model's idiosyncrasies without genuinely adhering to human intent. This is particularly acute in multimodal settings where "hallucination"—the generation of text unsupported by visual input is rampant. Generative verifiers, also known as Generative Reward Models (GRMs), address this by restructuring the reward signal: instead of a silent scalar, the reward model acts as a critic, grounding its evaluation in observable evidence. Unlike discriminative classifiers, these models generate a structured rationale or critique before (or alongside) making a final verification judgment. This "Wait-then-Judge" or

"Think-then-Score" mechanism aligns with System 2 cognitive processes, fostering transparency and robustness. According to the input-output mechanism of the verifier, we categorize them into three types: pair-wise, point-wise, and process level. Furthermore, we explore their applications in data filtering, evaluation, reinforcement learning, and test-time scaling.

3.1 Pair-wise

Pair-wise verifiers accept pairs of generations (candidate responses) from a Multimodal Large Language Model (MLLM) and compare them by generating a rationale before concluding which is superior. This approach leverages the intuition that relative ranking is often easier for models than absolute scoring.

Recent research highlights the utility of pair-wise verifiers in resolving the "pairwise-to-pointwise" disconnect. For instance, **Generative RLHF-V** [19] posits that while pairwise comparison allows models to learn generalizable principles of preference, these must be distilled to inform stable optimization. Similarly, the **LLaVA-Critic-R1** [20] architecture employs a self-critique mechanism where the model generates multiple trajectories and selects the optimal path via a tournament of pairwise comparisons, effectively internalizing the verifier role.

To address the training instability often found in these models, **R1-Reward** [21] introduces a stable reinforcement learning framework for training pair-wise verifiers. By reformulating reward modeling as a rule-based RL task, it employs the *StableReinforce* algorithm, which refines the training process through 'Pre-CLIP' on logit ratios and an 'Advantage Filter'. Uniquely, R1-Reward enforces a strict output format containing a reasoning trace (`<think>`) followed by a judgment (`<answer>`), utilizing a consistency reward to ensure the final ranking is logically derived from the generated critique.

Furthermore, recent comparative studies suggest that pair-wise verifiers functioning in a "knockout tournament" strategy can offer greater consistency than pointwise judges, particularly when handling high-quality responses with marginal differences.

3.2 Point-wise

Point-wise verifiers accept a single generation from an MLLM and assess it by generating a rationale—often based on specific rubrics or visual evidence—before concluding with a final scalar score or rating. This category represents a significant evolution from the "black box" scalar reward models.

A seminal work in this domain is **MM-RLHF** [11], which introduces a Critique-Based Reward Model. Instead of outputting a raw score, the model produces a textual critique (e.g., identifying specific object mismatches in an image) followed by a numerical rating. This forces the model's internal representations to attend to specific visual features relevant to the error. Similarly, **LLaVA-Critic** [10] is designed as a generalist evaluator, trained on high-quality instruction-following datasets to provide reasoning traces alongside scores, effectively challenging proprietary models like GPT-4o in the "LMM-as-a-judge" role.

In domain-specific applications, **EQA-RM** (Embodied Question Answering Reward Model) [22] tailors pointwise verification to embodied agents. It utilizes a schema-based output (evaluating object identification, navigation path, and reasoning separately) to provide fine-grained feedback for 3D navigation tasks, addressing the limitations of generic reward models in understanding spatio-temporal logic.

3.3 Process level

Process-level verifiers, or Process Reward Models (PRMs), accept a single generation and judge the correctness of distinct steps within the reasoning chain. By validating the "how" rather than just the final "what," these models address the credit assignment problem inherent in complex logical or mathematical tasks.

Moving beyond simple scoring, **GM-PRM** (Generative Multimodal Process Reward Model) [23] focuses on active error correction. It analyzes step-wise intent and visual alignment; if a step is flagged as incorrect, it generates a correction to guide the reasoning trajectory back to a valid path. Similarly, **VRPRM** [24] has begun to explore generative process rewards, enabling PRMs to output reasoning traces alongside verification scores for each reasoning step.

Addressing the constraints of internal parametric knowledge, **TIM-PRM** [25] distinguishes itself by integrating *explicit tool use* into the verification process. Unlike VRPRM or GM-PRM, which rely solely on internal chain-of-thought, TIM-PRM actively queries the visual environment through independent question-asking. This agentic approach decouples perception from reasoning, mitigating the hallucination propagation found in previous generative verifiers.

4 Usage of Multimodal Verifiers

4.1 Training data filtering

Generative verifiers are increasingly employed upstream in the pipeline to filter low-quality training data, mitigating the "Garbage In, Garbage Out" problem in Multimodal Supervised Fine-Tuning (SFT).

The **VERITAS** [26] pipeline exemplifies this application. It employs a multi-expert critique system (comprising GPT-4o, Gemini-Pro, and Doubao-Pro) grounded by vision priors (such as OCR and tagging tools). These experts critique data samples, and their judgments are statistically fused to train a lightweight, open-source critic. This distilled critic is then deployed to filter massive datasets at scale, removing hallucinations and factual errors before model training begins. Similarly, **OmniVerifier** [27] utilizes automated data construction pipelines (fixing images while modifying prompts, and vice versa) to create robust verification datasets that assist in cleaning generation and reasoning data.

4.2 Evaluator

Beyond training, generative verifiers serve as scalable substitutes for human annotators in benchmarking MLLMs. **LLaVA-Critic** [10] has demonstrated high correlation with human judgments, allowing for automated evaluation across diverse multimodal tasks ranging from captioning to complex reasoning.

Scaling this concept to a comprehensive framework, **VLMEvalKit** [28] provides an open-source toolkit supporting over 200 models and 80 benchmarks. It standardizes the evaluation process by employing generation-based evaluation across diverse tasks. Crucially, it integrates Large Language Models (LLMs) as choice extractors and judges to mitigate the impact of response formatting errors and employs strategies like circular evaluation to reduce variance, ensuring that evaluations reflect true model capability rather than just instruction-following strictness.

However, the reliability of these evaluators is itself a subject of scrutiny. The **VideoReward-Bench** [29] study reveals that even top-tier models like GPT-4o struggle with temporal reasoning in video, achieving only moderate accuracy. This underscores the necessity of specialized benchmarks like **VisualProcessBench** [9] (focusing on step-wise reasoning accuracy) and **ViVerBench** [27] (assessing universal verification capabilities across diverse categories) to "judge the judges."

4.3 Reinforcement learning

Generative verifiers are pivotal in modern Reinforcement Learning (RL) frameworks, providing the high-quality signals necessary to optimize policies.

In the **MM-RLHF** [11] framework, the generative critique enables *Dynamic Reward Scaling*. This technique weights the optimization loss based on the clarity and confidence of the critique, ensuring the model learns more aggressively from clear errors than from ambiguous signals.

EQA-RM [22] introduces *Contrastive Group Relative Policy Optimization (C-GRPO)* for embodied agents. By utilizing rule-based contrastive rewards derived from temporal and spatial perturbations, the verifier guides the agent to avoid specific failure modes like chronological confusion or object hallucination.

Furthermore, **LLaVA-Critic-R1** [20] proposes a unified architecture where the critic and policy are isomorphic. By training the model via RL to optimize its preference judgments, the system improves both its generation and evaluation capabilities simultaneously, suggesting that strong critics are fundamentally strong policies. **MCM-DPO** [30] extends this to multifaceted optimization, using complex reward functions to capture competing qualities like visual grounding versus textual fluency.

4.4 Test-time scaling

Test-time scaling utilizes generative verifiers to enhance performance during inference, often through iterative refinement or search strategies.

OmniVerifier-TTS [27] demonstrates Sequential Test-Time Scaling for image generation. Rather than simple parallel sampling, it employs a loop where the verifier critiques a generated image, produces an edit prompt if errors are detected, and guides the model to refine the output.

In reasoning tasks, **GM-PRM** [23] employs a *Refined Best-of-N* framework. Instead of merely discarding incorrect trajectories, it actively repairs flawed steps using generative feedback, improving sample efficiency. Similarly, the **Policy as Generative Verifier (PAG)** [31] framework alternates the LLM between a proposer and a verifier mode. It uses a "Verify-then-Revise" workflow where revisions are only triggered if the generative verification step detects an error, preventing model collapse due to excessive editing.

5 Conclusion and Future Directions

As Multimodal Large Language Models (MLLMs) continue to scale, the reliance on purely supervised learning has revealed significant limitations in grounding, logical reasoning, and safety. This survey has systematized the emerging landscape of *verifiers*—mechanisms designed to assess and guide model outputs. We have categorized these systems into **discriminative verifiers**, which offer efficient scalar signals for outcome and process supervision, and **generative verifiers**, which leverage the reasoning capabilities of LLMs to provide interpretable critiques and self-corrections.

The trajectory of recent research indicates a decisive shift from "black-box" scoring toward "glass-box" reasoning. While early reward models treated verification as a classification task, the state-of-the-art is moving toward System 2 thinking, where verifiers explicitly deliberate on visual evidence and reasoning steps before rendering a judgment. This evolution is enabling new paradigms in the model lifecycle, from filtering pre-training data to enabling compute-intensive test-time scaling.

Despite this progress, several open challenges and research directions remain:

Unified Training of Policy and Verifier. Current pipelines often treat the policy (generator) and the verifier (reward model) as distinct entities. A promising direction is the unification of these roles, where a single model improves its generation capabilities through self-verification and refines its verification skills by learning from its own generation errors. Frameworks like LLaVA-Critic-R1 suggest that the distinction between "actor" and "critic" may eventually dissolve into self-evolving systems.

Efficiency in System 2 Verification. Generative verifiers, particularly those utilizing Chain-of-Thought or test-time search, incur high inference costs. Future work must address the trade-off between verification depth and computational latency. Techniques such as distilling the reasoning traces of heavy generative verifiers into lightweight discriminative models or optimizing token-efficient critique formats will be essential for real-time applications.

Temporal and Embodied Consistency. Most current verifiers excel at static image-text alignment but struggle with the temporal dynamics of video and the spatial logic of embodied agents. Developing verifiers that can track object permanence, causality, and navigational consistency over long horizons is a critical frontier. This requires moving beyond frame-level analysis to video-native or 4D-native reward modeling.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, February 2025. URL `http://arxiv.org/abs/2502.13923`. arXiv:2502.13923 [cs].

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi,

Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling, September 2025. URL http://arxiv.org/abs/2412.05271. arXiv:2412.05271 [cs].

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

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, October 2024. URL http://arxiv.org/abs/2408.03326. arXiv:2408.03326 [cs].

[5] Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. Unveiling the Ignorance of MLLMs: Seeing Clearly, Answering Incorrectly. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 9087–9097, June 2025.

[6] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th international conference on neural information processing systems*, Nips '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8. Number of pages: 15 Place: New Orleans, LA, USA tex.articleno: 2011.

[7] Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork-VL Reward: An Effective Reward Model for Multimodal Understanding and Reasoning, June 2025. URL http://arxiv.org/abs/2505.07263. arXiv:2505.07263 [cs].

[8] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step, May 2023. URL http://arxiv.org/abs/2305.20050. arXiv:2305.20050 [cs].

[9] 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. VisualPRM: An Effective Process Reward Model for Multimodal Reasoning, March 2025. URL https://arxiv.org/abs/2503.10291v1.

[10] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLaVA-Critic: Learning to Evaluate Multimodal Models, March 2025. URL http://arxiv.org/abs/2410.02712. arXiv:2410.02712 [cs].

[11] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment, February 2025. URL http://arxiv.org/abs/2502.10391. arXiv:2502.10391 [cs].

[12] Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. A Comprehensive Survey of Reward Models: Taxonomy, Applications, Challenges, and Future, April 2025. URL http://arxiv.org/abs/2504.12328. arXiv:2504.12328 [cs].

[13] Congming Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. A Survey of Process

Reward Models: From Outcome Signals to Process Supervisions for Large Language Models, October 2025. URL http://arxiv.org/abs/2510.08049. arXiv:2510.08049 [cs].

[14] V. Venktesh, Mandeep Rathee, and Avishek Anand. Trust but Verify! A Survey on Verification Design for Test-time Scaling, September 2025. URL http://arxiv.org/abs/2508.16665. arXiv:2508.16665 [cs].

[15] Qiyuan Liu, Hao Xu, Xuhong Chen, Wei Chen, Yee Whye Teh, and Ning Miao. Enhancing Large Language Model Reasoning with Reward Models: An Analytical Survey, October 2025. URL http://arxiv.org/abs/2510.01925. arXiv:2510.01925 [cs].

[16] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model, 2025. URL https://arxiv.org/abs/2501.12368. _eprint: 2501.12368.

[17] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin Zeng, and Yujiu Yang. URSA: Understanding and Verifying Chain-of-thought Reasoning in Multimodal Mathematics, 2025. URL https://arxiv.org/abs/2501.04686. _eprint: 2501.04686.

[18] Shuai Wang, Zhenhua Liu, Jiaheng Wei, Xuanwu Yin, Dong Li, and Emad Barsoum. Athena: Enhancing Multimodal Reasoning with Data-efficient Process Reward Models, November 2025. URL http://arxiv.org/abs/2506.09532. arXiv:2506.09532 [cs].

[19] Jiayi Zhou, Jiaming Ji, Boyuan Chen, Jiapeng Sun, Wenqi Chen, Donghai Hong, Sirui Han, Yike Guo, and Yaodong Yang. Generative RLHF-V: Learning Principles from Multi-modal Human Preference, May 2025. URL http://arxiv.org/abs/2505.18531. arXiv:2505.18531 [cs].

[20] Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. LLaVA-Critic-R1: Your Critic Model is Secretly a Strong Policy Model, August 2025. URL http://arxiv.org/abs/2509.00676. arXiv:2509.00676 [cs].

[21] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, and Liang Wang. R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning, May 2025. URL http://arxiv.org/abs/2505.02835. arXiv:2505.02835 [cs].

[22] Yuhang Chen, Zhen Tan, and Tianlong Chen. EQA-RM: A Generative Embodied Reward Model with Test-time Scaling, June 2025. URL http://arxiv.org/abs/2506.10389. arXiv:2506.10389 [cs].

[23] Jianghangfan Zhang, Yibo Yan, Kening Zheng, Xin Zou, Song Dai, and Xuming Hu. GM-PRM: A Generative Multimodal Process Reward Model for Multimodal Mathematical Reasoning, August 2025. URL http://arxiv.org/abs/2508.04088. arXiv:2508.04088 [cs].

[24] VRPRM: Process Reward Modeling via Visual Reasoning. October 2025. URL https://openreview.net/forum?id=sj9jmrBjMf.

[25] Peng Kuang, Xiangxiang Wang, Wentao Liu, Jian Dong, Kaidi Xu, and Haohan Wang. TIM-PRM: Verifying multimodal reasoning with Tool-Integrated PRM, November 2025. URL http://arxiv.org/abs/2511.22998. arXiv:2511.22998 [cs] version: 1.

[26] Tingqiao Xu, Ziru Zeng, and Jiayu Chen. VERITAS: Leveraging Vision Priors and Expert Fusion to Improve Multimodal Data, October 2025. URL http://arxiv.org/abs/2510.15317. arXiv:2510.15317 [cs] version: 1.

[27] Xinchen Zhang, Xiaoying Zhang, Youbin Wu, Yanbin Cao, Renrui Zhang, Ruihang Chu, Ling Yang, and Yujiu Yang. Generative Universal Verifier as Multimodal Meta-Reasoner, October 2025. URL http://arxiv.org/abs/2510.13804. arXiv:2510.13804 [cs].

[28] Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Yuxuan Qiao, Mo Li, Amit Agarwal, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Xiaozhe Li, Chaoyou Fu, Junbo Cui, Jixuan Chen, Enxin Song, Song Mao, Shengyuan Ding, Tianhao Liang, Zicheng Zhang, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models, August 2025. URL http://arxiv.org/abs/2407.11691. arXiv:2407.11691 [cs].

[29] Zhihong Zhang, Xiaojian Huang, Jin Xu, Zhuodong Luo, Xinzhi Wang, Jiansheng Wei, and Xuejin Chen. VideoRewardBench: Comprehensive Evaluation of Multimodal Reward Models for Video Understanding, August 2025. URL http://arxiv.org/abs/2509.00484. arXiv:2509.00484 [cs].

[30] Jinlan Fu, Shenzhen Huangfu, Hao Fei, Yichong Huang, Xiaoyu Shen, Xipeng Qiu, and See-Kiong Ng. MCM-DPO: Multifaceted Cross-Modal Direct Preference Optimization for Alt-text Generation, October 2025. URL http://arxiv.org/abs/2510.00647. arXiv:2510.00647 [cs].

[31] Yuhua Jiang, Yuwen Xiong, Yufeng Yuan, Chao Xin, Wenyuan Xu, Yu Yue, Qianchuan Zhao, and Lin Yan. PAG: Multi-Turn Reinforced LLM Self-Correction with Policy as Generative Verifier, June 2025. URL http://arxiv.org/abs/2506.10406. arXiv:2506.10406 [cs].