Multimodal LLM Alignment: Challenges, Solutions, and Research Opportunities

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have exhibited substantial capability in performing complex tasks that integrate visual, auditory, and textual modalities. Nevertheless, they continue to exhibit notable shortcomings in truthfulness, safety, and alignment with human preferences, thereby motivating intensive research on alignment algorithms specifically tailored to MLLMs. This paper provides a systematic and comprehensive survey of MLLM alignment, structured along three principal dimensions: (i) the algorithmic pipeline and technical procedures underlying alignment methods, (ii) the application domains and usage scenarios for which these methods are designed, and (iii) the evaluation methodologies employed to measure alignment quality. The objective of this work is to furnish researchers with a coherent framework for situating recent advances in MLLM alignment and to facilitate the development of more robust, reliable, and human-aligned multimodal systems.

1 Introduction

Multimodal large language models (MLLMs) have achieved remarkable progress in integrating information from vision, language, and other modalities, and have been widely deployed across downstream applications such as embodied AI (Liang et al., 2025b), scientific reasoning (Wang et al., 2025d), and human–AI interaction (Durante et al., 2024). Despite these impressive advances, a fundamental open question remains: **how can we rigorously align the behavior of these models with human values and expectations?** Since the early development of large language models, alignment has been a central and continuously evolving theme, giving rise to a rich family of alignment algorithms shown in Fig 1. Nowadays, alignment procedures have become an indispensable component of foundation model training pipelines, referred to Qwen3-VL (Bai et al., 2025), Keye-VL (Kwai Keye Team et al., 2025) and InternVL3.5 (Wang et al., 2025c), and they increasingly exhibit a dual nature: wide applicability and task-specificity.

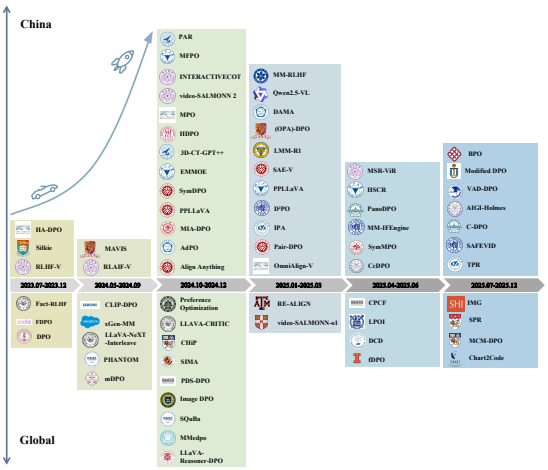

Figure 1: A timeline of MLLM alignment algorithms

Recent survey papers have focused on MLLMs alignment (Ji et al., 2025) and (Li et al., 2025d) systematically review LLMs and MLLMs, emphasizing the critical role of alignment in enabling real-world or downstream deployment. (Shu et al., 2025) approaches the problem from an explainability perspective, analyzing concrete phenomena and potential causes of alignment and misalignment at both the representation and behavior levels. However, existing work either treats multimodal alignment as just one module among many or focuses primarily on describing literal phenomena. A systematic integration of alignment methods, evaluation protocols, and application practice for MLLMs is still lacking. Against this backdrop, we revisit the alignment problem for multimodal large models from a more systematized and application-oriented perspective.

Our main contributions can be summarized as follows:

- **Technical pipeline.** We clearly present the de-

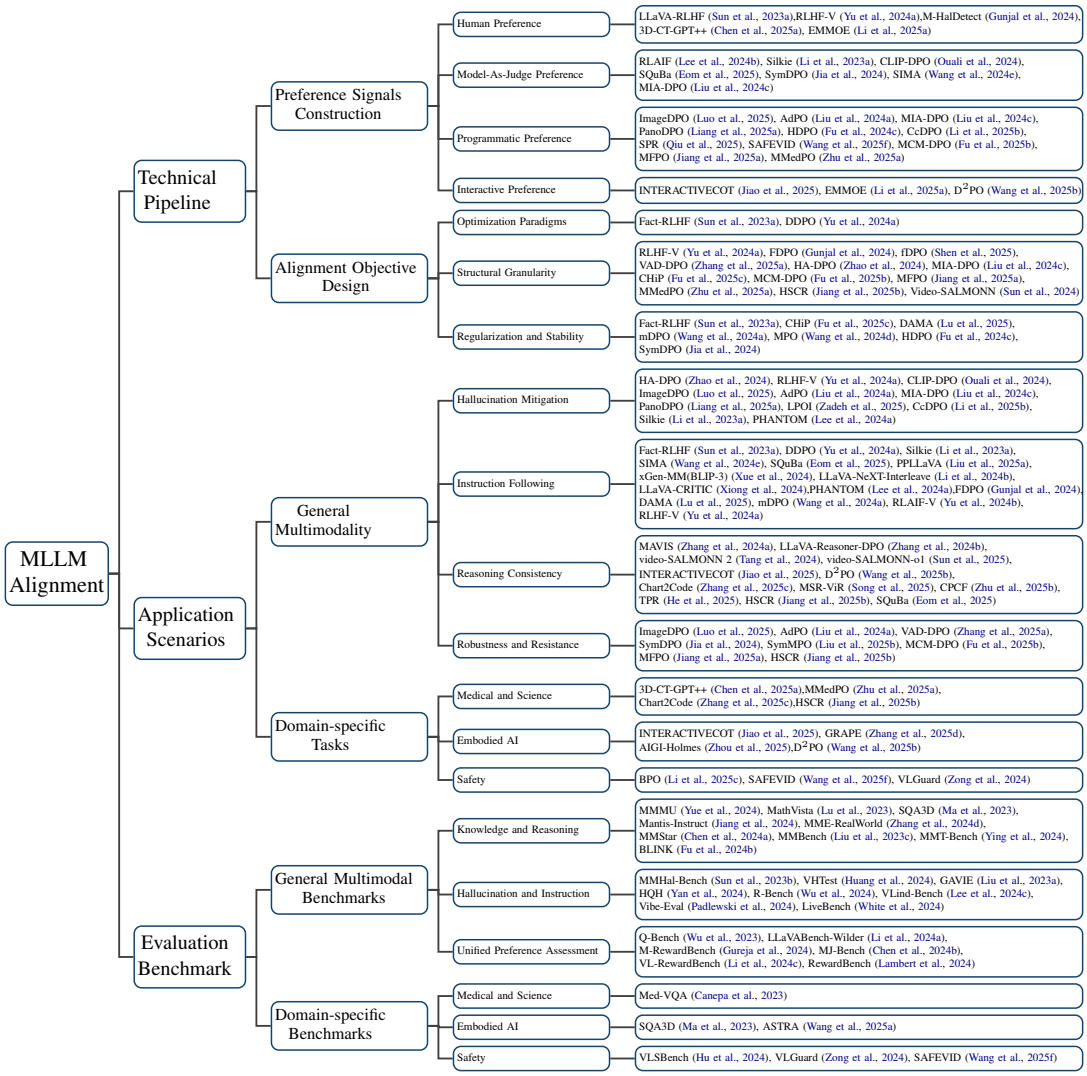

Figure 2: The overall framework of MLLM Alignment

tailed technical pipeline for multimodal alignment, covering preference signal construction and alignment objective design, thereby formalizing an end-to-end recipe for the MLLM alignment algorithm.

- **Application scenarios.** We analyze how alignment manifests across a variety of multimodal application scenarios, highlighting recurring design patterns as well as task-dependent requirements and constraints.

- **Evaluation framework.** Across diverse application scenarios, we systematically compare the relevant evaluation benchmarks and their key emphases, thereby guiding researchers to refine alignment strategies toward metric-specific objectives correspondingly.

The overall framework of our paper is shown in Fig 2. This survey aims to provide a comprehensive and systematized perspective that enables researchers in both academia and industry to identify fundamental, cross-domain challenges in multimodal alignment. By doing so, we hope it can facilitate the design and innovation of more broadly applicable alignment algorithms that can effectively address, and ultimately go beyond, the complexities of real-world environments.

2 MLLM Alignment Technical Pipeline

Alignment in MLLM is generally an end-to-end design process. In practice, it is typically decomposed into two major stages: **preference signal construction** and **alignment objective design**. The preference signal determines the sources and representational forms of supervision used during alignment. In contrast, the alignment objective specifies how these signals are integrated into learning and the

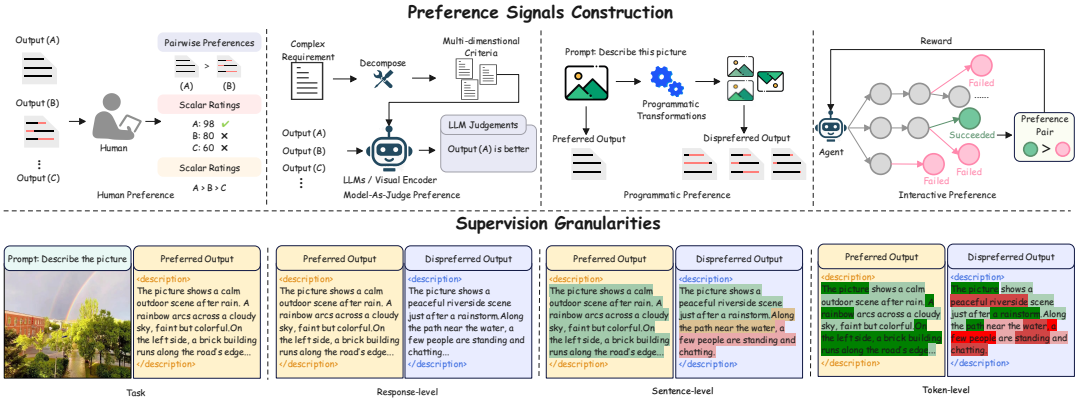

Figure 3: Preference signal construction mechanisms and supervision granularity designs.

degree or intensity of alignment is enforced. Consequently, alignment across domains and tasks should not be viewed as isolated cases; rather, it can be understood as an adaptive design grounded in diverse human preferences and expectations. This section provides a detailed decomposition and analysis of this unified design framework.

2.1 Preference Signals Construction

Multimodal settings contain rich, experience-grounded informational factors. Expected preferences vary substantially across tasks and may even be mutually conflicting. This diversity gives rise to heterogeneous, task-specific alignment needs, which in turn impose different requirements on how preference signals should be constructed, as shown in the upper part of Fig 3.

2.1.1 Human Preference

The most direct form of supervision for preference alignment comes from humans. To a considerable extent, MLLM alignment mirrored the success of RLHF in the text-only setting: annotators provide either pairwise preferences, scalar ratings, or ranked sequences to support preference labeling. Early MLLMs alignment methods, such as LLaVA-RLHF (Sun et al., 2023a), RLHF-V (Yu et al., 2024a), M-HalDetect (Gunjal et al., 2024), largely relied on carefully designed human annotations for training datasets, which substantially mitigated hallucinations and jailbreak attacks. In domains with high professional complexity or stringent safety requirements, human-in-the-loop remains indispensable. In biomedical systems, 3D-CT-GPT++ (Chen et al., 2025a) depend on expert-designed tasks and quality criteria, even when parts of the subsequent scoring process are delegated to language models. Similarly, embodied settings such as EMMOE (Li et al., 2025a) typically require the collection and analysis of expert demonstrations as a prerequisite for building imitation-learning and preference datasets.

2.1.2 Model-As-Judge Preference

Golden standard human annotations are prohibitively expensive. With the growing adoption of RLAIF (Lee et al., 2024b), increasing work has shifted to the LLM-as-judge paradigm, leveraging strong LLMs or auxiliary visual encoders to generate labels and scores for candidate samples. Closed-source models such as GPT (OpenAI., 2024; Achiam et al., 2023) and Gemini (Comanici et al., 2025; Team, 2024) have reached near-human performance on multiple tasks, therefore widely used to annotate preference data across diverse areas, including vision-language understanding (Liu et al., 2023b; Li et al., 2023a; Lee et al., 2024c), video understanding (Tang et al., 2025; Sun et al., 2024), long chain-of-thought reasoning (Zhang et al., 2025b; Wang et al., 2025e), and model safety (Zong et al., 2024; Zhao et al., 2024).

A key ingredient of Model-based annotation is prompt engineering that operationalizes target preferences: decomposing complex requirements into multi-dimensional criteria to improve reliability. For example, Silkie (Li et al., 2023a) uses GPT-4V to score responses along dimensions such as helpfulness, visual faithfulness, and ethical considerations, yielding relatively stable preference signals. CLIP-DPO (Ouali et al., 2024), in contrast, employs CLIP-like visual encoders to assist preference generation and scoring, illustrating how

modular encoders can be exploited for preference construction.

Beyond such external judges, preference signals could come from the target model itself: depending on modality settings, methods such as SQuBa (Eom et al., 2025), SymDPO (Jia et al., 2024), SIMA (Wang et al., 2024e), and MIA-DPO (Liu et al., 2024c) construct preference pairs from model responses in text-only or vision-language scenarios. Nevertheless, Model-as-judge substantially improves labeling efficiency and reduces cost, but it also inevitably transfers the biases and failure modes of the judging or generating models into the preference data.

2.1.3 Programmatic Preference

Commonly, preferences are programmatically specified through reusable rules and procedural generation. This makes programmatic preference construction especially important for MLLMs, as it directly encodes task structure, safety constraints, and definitions of hallucination into the data generation pipeline.

Visual perturbations and contrastive variants provide a practical route to constructing preference data at scale. Image DPO (Luo et al., 2025) and AdPO (Liu et al., 2024a) generate aligned positive/negative samples through semantic edits, blurring, or adversarial transformations, while MIA-DPO and PanoDPO (Liang et al., 2025a) exploit compositing, viewpoint variation, or conditional preferences to strengthen visual grounding. For hallucination and safety control, HDPO (Fu et al., 2024c), CcDPO (Li et al., 2025b), SPR (Qiu et al., 2025), and SAFEVID (Wang et al., 2025f) create contrastive pairs by modulating hallucination severity, temporal coherence, or exposure to hazardous content. Moving beyond strict pairwise setups, MCM-DPO (Fu et al., 2025c) and MFPO (Jiang et al., 2025a) define modality-aware, multi-criteria preferences, and MMedPO (Zhu et al., 2025a) introduces domain-aware weighting to emphasize clinically consequential errors.

Programmatic preference design has become central to modern MLLM alignment and offers low-cost scalability, improved control over specific failure modes, and more reproducible training signals than large-scale human annotation or generic judge-based scoring.

2.1.4 Interactive Preference

As the scope of MLLM tasks expands, preference signals no longer arise solely from static annotation pipelines; they can also be induced from reward information obtained through interaction with environments or tasks. MLLMs operate as agents that generate actions, queries, or reasoning trajectories and receive feedback in the form of task success signals, environment rewards, or trajectory-level performance scores. INTERACTIVECOT (Jiao et al., 2025) in ALFWorld, EMMOE in Replica-like 3D environments, and D^2PO (Wang et al., 2025b) for embodied planning, which derive preferences from task completion signals and decomposed subgoals. GRAPE similarly aligns policies at the trajectory level by contrasting successful and failed trials under spatiotemporal constraints.

While auxiliary LLM judges may still be employed at certain stages, the primary supervision in these approaches is inherently dynamic and trajectory-dependent rather than a one-shot evaluation of static input–output pairs, making interaction itself a central source of alignment feedback.

2.2 Alignment Objective Design

Alignment objectives for multimodal build upon the general framework of preference-based or feedback-based alignment, while introducing broader constraints and more fine-grained decompositions centered on optimization paradigms, structural granularity and stability mechanisms. The detailed optimization objectives are shown in Table 1.

2.2.1 Optimization Paradigms

Online optimization largely follows the RLHF-style formulation, treating alignment as a policy-based reinforcement learning problem. This was a commonly adopted framework in early alignment work; Fact-RLHF (Sun et al., 2023a) uses human scores as rewards to construct the RLHF training process. Related RL variants, such as DDPO (Yu et al., 2024a) and RL4VLM (Zhai et al., 2024) interact with an environment or sampler and receive reward signals provided by humans or strong baseline models.

Instead of explicitly fitting a reward model and performing on-policy policy optimization as in PPO-based RLHF, offline-based DPO directly optimizes the policy with a simple contrastive objective over offline preference pairs, while implicitly incorporating a reference-model regularization effect.

This RL-free formulation substantially reduces algorithmic and engineering complexity, improves training stability, and lowers computational cost by avoiding iterative sampling-and-optimization loops. Together with its simplicity and scalability, it has made it one of the most widely adopted approaches for preference alignment in current practice.

2.2.2 Structural Granularity

Traditional alignment typically operates at the response level, treating the output as a single response sequence. For a multimodal input $(x, \mathcal{I})$, where x is a textual query and $\mathcal{I}$ denotes the associated images. The objective is commonly defined over preference pairs, optimizing the model to favor a preferred response over a dispreferred response(i.e., (y_w, y_l)). Although this response-level formulation is effective across many tasks, recent works move beyond it by introducing fine-grained and hierarchical supervision illustrated in Fig 3 below, either by assigning non-uniform weights to tokens/spans or by defining dedicated objectives for specific structured components of multimodal outputs.

Fine-grained sequence. Fine-grained supervision improves credit assignment by pushing preference signals down to segment, sentence, or token units, which helps suppress localized hallucinations and makes better use of limited preference data. Dense-DPO in RLHF-V concentrates optimization on uncertain or reliability-critical spans, so the model is explicitly trained to correct the most error-prone fragments rather than merely shifting overall response style. FDPO (Gunjal et al., 2024) refines the pairwise objective by decoupling the contributions inside the sigmoid term, offering tighter control over local gradient behavior while keeping the offline preference optimization simplicity. This line also motivates token-level or span-aware DPO variants such as fDPO (Shen et al., 2025) and VAD-DPO (Zhang et al., 2025a) that retain the response-level backbone but inject finer error-targeting mechanisms.

Multiple hierarchies. Hierarchical fusion improves stability and controllability by jointly aligning global preference direction and local grounding constraints across different granularities. HA-DPO (Zhao et al., 2024) and MIA-DPO extend response-level DPO with auxiliary supervised losses from positive examples to strengthen alignment under noisy or sparse supervision. CHiP (Fu et al., 2025c) represents a more explicit multi-level design by combining visual DPO, sentence-level DPO, and a token-level KL-style sequence constraint, effectively distributing alignment pressure across textual structure and visual evidence. Even when the main objective remains response-level, MIA-DPO's attention-based multi-image pairing implicitly spreads supervision over multiple visual components, improving robustness to compositional inputs.

Multiple dimension. Multi-dimensional designs broaden alignment beyond a single "good vs bad" axis, enabling structured control over multi-component inputs and heterogeneous error modes. MCM-DPO and MFPO define preferences over combinations of answer text, images, and context, encouraging consistent cross-modal reasoning rather than modality-isolated optimization. MMedPO and HSCR (Jiang et al., 2025b) introduce multiple dispreference types and learn to rank them, which supports finer-grained governance of distinct failure patterns such as hallucination versus omission in domain-critical settings. Video-SALMONN series (Sun et al., 2024; Tang et al., 2025) extends this idea to temporal media, video-oriented objectives treat coherence and time-consistency as additional alignment dimensions, pushing preference learning toward richer, scenario-aware control.

2.2.3 Regularization and Stability

As many alignment algorithms can be formulated as contrastive preference optimization, a key concern is preventing excessive distribution shift, catastrophic forgetting, and undesirable likelihood drift relative to base models. Recent work addresses this with two recurring objective-level designs.

Reference-oriented regularization. Reference-oriented regularization constrains updates toward a trusted baseline. Fact-RLHF uses an explicit KL penalty to limit PPO deviation from the initial policy; CHiP adds KL terms between the reference and updated models for both preferred and dispreferred responses; and PO/MPO-style objectives incorporate log-ratio or reference-based penalties to keep adapted policies close to clean or accurate anchors.

Margin and asymmetry shaping. Complementing reference constraints, margin design and asymmetry-aware objectives regulate how strongly preferences are enforced. DAMA (Lu et al., 2025) rescales the DPO logit gap, mDPO (Wang et al.,

2024a) and MPO (Wang et al., 2024d) impose margin-style constraints to preserve base accuracy while enforcing a minimum preference distance, and hallucination-oriented variants such as HDPO and SymDPO adopt anchored or symmetric formulations to mitigate probability drift.

3 Application Scenarios

MLLM alignment is inherently application-driven. As the scope of MLLM deployment continues to expand, these models have demonstrated substantial value in both general-purpose multimodal settings and more domain-specific scenarios.

3.1 General Multimodality

General multimodal capability serves as a foundational application for assessing the effectiveness of multimodal alignment. Correspondingly, at the level of alignment strategy, it manifests as four system-level dimensions that recur across multimodal model designs: hallucination mitigation, instruction following, reasoning consistency, and robustness and resistance.

Hallucination mitigation Hallucination mitigation is a core objective in MLLM alignment, aiming to ensure responses are grounded in visual evidence rather than linguistic priors or spurious cues. Most methods construct preference pairs that differ in visual faithfulness and apply token- or instance-level penalties. Clause-aware approaches (Zhao et al., 2024) suppress labeled hallucinated spans; image-perturbation methods such as CLIP-DPO, Image DPO and AdPO enforce true visual dependence by altering visuals while keeping text fixed; and multi-image PanoDPO (Liang et al., 2025a) and LPOI (Zadeh et al., 2025) contrast viewpoints or occlusions to reduce shortcut reasoning. The inclusion of visual faithfulness as an explicit scoring in broader systems (Lee et al., 2024a) indicates that hallucination control is now a standard component of general MLLM alignment.

Instruction following Instruction-following alignment aims to position MLLMs as practical, well-behaved assistants that can correctly interpret complex prompts, respect constraints, and produce task-compliant outputs. RLHF-style pipelines optimize policies with user-derived rewards that reflect overall user preferences over multimodal responses. In contrast, DPO-style (Liu et al., 2025a; Xue et al., 2024; Li et al., 2024b; Xiong et al., 2024) scales instruction-response data construction and derives preference pairs from human ratings, LLM-as-judge scoring, or cross-model comparisons. Extensions such as FDPO, DAMA, and mDPO refine how strong versus weak preferences are separated, while RLAIF-V (Yu et al., 2024b) and RLHF-V incorporate dense judgments from proprietary or open evaluators to improve controllability and reliability. Across these systems, visual inputs are integrated into prompts, yet the primary alignment target remains robust multimodal instruction adherence and consistent, controllable satisfaction of user goals.

Reasoning consistency With the rise of CoT, multimodal reasoning has become an increasingly active direction in MLLM alignment. Reasoning alignment targets how models reach answers by shaping CoT traces, decomposition, and intermediate decisions rather than only final outputs. This line spans domain-focused CoT alignment for math and visual reasoning (Zhang et al., 2024a,b), step-wise optimization for temporal and audio-visual understanding (Tang et al., 2024; Sun et al., 2025), and environment-grounded CoT improved by task success signals in embodied settings (Jiao et al., 2025). In parallel, several methods align multi-step intermediate artifacts, such as plans (Song et al., 2025), code variants (Zhang et al., 2025c), or structured edits (Zhu et al., 2025b; He et al., 2025), so that preference learning directly constrains the reasoning products themselves. These approaches broaden reasoning alignment from explicit CoT supervision to outcome-centric and artifact-centric designs, encouraging MLLMs to act as more interpretable, human-aligned multimodal reasoners.

Robustness and resistance Robustness and resistance alignment aims to ensure that MLLMs remain stable under benign perturbations and resilient to adversarial or misleading inputs, such as inconsistencies across images or between modalities, thereby reducing modality inertia and spurious cross-modal reliance. Image DPO and AdPO construct preference pairs by perturbing only the visual input, rewarding responses that remain correct when semantics are preserved and appropriately change when visual meaning shifts. VAD-DPO extends this idea to videos by contrasting visually similar yet semantically opposite clips to heighten sensitivity to subtle temporal differences. Complementary robustness-oriented objectives, such as SymDPO, SymMPO (Liu et al., 2025b), adopt sym-

metric, to curb over-reliance on textual cues and preserve performance. While MCM-DPO, MFPO and HSCR use sparsity-aware formulations that generalize supervision from isolated pairs to multi-candidate comparisons and encourage reliance on the most informative visual evidence.

3.2 Domain-specific Tasks

MLLM alignment becomes more in-depth in domain-specific tasks, exhibiting strong domain specificity, and is primarily applied to Medical and science, embodied AI, and safety scenarios.

Medical and science Medical and scientific applications place a premium on factual correctness, domain expertise, and calibrated behavior under complex multimodal evidence. 3D-CT-GPT++ constructs radiology preference data scored by experts or experienced reasoning LLMs. MMedPO differentiates hallucination-type versus lesion and weights them by clinical severity with visualization-aware signals. MSR-ViR further aligns long-form scientific summarization via preferences over multi-stage outputs. HSCR (Jiang et al., 2025b) targets medical VLM alignment by automatically constructing high-quality preference pairs to improve clinical trustworthiness with limited training data.

Embodied AI MLLMs are increasingly used for spatial perception and action selection in simulated and real environments. INTERACTIVECOT collects rollouts in environments like ALFWorld and treats task success as a preference signal to optimize multimodal CoT traces for action. In robotics, GRAPE aligns manipulation policies by contrasting successful and failed trajectories under spatial constraints, while D^2PO jointly learns action selection and state prediction via preference learning, coupling policy and world model learning. Domain systems like AIGI-Holmes (Zhou et al., 2025) combine vision experts with MLLMs and apply DPO to task-specific datasets for visual forensics and inspection. Across these settings, alignment depends on environmental dynamics and long-horizon success, making robustness to distribution shift and stable interaction crucial.

Safety Safety-centered alignment seeks to prevent harmful content, curb misuse, and improve risk awareness in multimodal systems. BPO (Li et al., 2025c) introduces the MMSafe-PO dataset and a text-inertia loss to deter "modality cheating," making safe, visually grounded responses strictly preferred over unsafe or ungrounded ones. Training-time safety alignment is increasingly complemented by inference-time modules that serve as multimodal filters or risk classifiers. LLaMA Guard 3 Vision (Cheng et al., 2024) provides pairwise safety gating at the response level, while VL-Guard (Zong et al., 2024) targets lightweight guard models and calibrated risk detection for VLM deployments. MultiTrust (Zhang et al.) offers a unified trustworthiness evaluation suite to audit safety and robustness under diverse stressors, and MMDT (Xu et al., 2025) enforces decoding-time trustworthiness by intervening during generation rather than relying only on post-hoc filtering. Overall, safety alignment typically adopts conservative objectives and uses adversarially constructed preference data to improve robustness and reliability against hostile or deceptive inputs.

4 Benchmark and Evaluations

Mirroring the application landscape, existing benchmarks can be broadly organized into general multimodal and domain-specific evaluations: the former are designed to reflect breadth-oriented use of MLLMs across diverse vision–language tasks and thus measure average capability and overall alignment quality, while the latter track deployment in specialized settings with task-critical depth and boundary conditions. We list representative benchmarks in Table 2.

4.1 General Multimodal Benchmarks

General-purpose multimodal benchmarks are typically designed to probe broad, cross-task capabilities over diverse vision–language settings, a trend reinforced by the growing integration of alignment into mainstream post-training pipelines. The focus of general evaluation has progressively crystallized into three dimensions: knowledge and reasoning, hallucination and instruction, and unified preference assessment.

Knowledge and reasoning. As demands for multimodal reasoning grow, general-purpose knowledge and reasoning benchmarks are increasingly used to track how alignment reshapes a model's overall reasoning profile. MMMU (Yue et al., 2024) and MathVista (Lu et al., 2023) remain representative suites for cross-disciplinary and math-oriented reasoning, while SQA3D (Ma et al., 2023)targets spatial perception and grounded reasoning. Complementing these, fine-grained real-

world visual reasoning benchmarks are often organized as broader suites (Jiang et al., 2024; Zhang et al., 2024d; Chen et al., 2024a; Liu et al., 2023c; Ying et al., 2024; Fu et al., 2024b), which probe more localized and realistic visual reasoning effects of alignment.

Hallucination and instruction. Another widely emphasized dimension is whether aligned models can produce faithful, evidence-consistent responses and remain robust against misleading or shifting contexts. Benchmarks such as MMHal-Bench (Sun et al., 2023b), VHTest (Huang et al., 2024), GAVIE (Liu et al., 2023a), and HQH (Yan et al., 2024) primarily measure factual grounding by testing consistency between visual inputs and generated responses, enabling systematic measurement of multimodal hallucinations. In contrast, R-Bench (Wu et al., 2024), VLind-Bench (Lee et al., 2024c), Vibe-Eval (Padlewski et al., 2024), and LiveBench (White et al., 2024) place greater weight on robustness and interaction behavior, assessing whether models maintain cautious, stable, and faithful response patterns under adverse conditions or distribution shifts.

Unified preference assessment. Alignment stems from the need for consistency with human preferences, so a model's alignment is often judged by how closely its outputs match human judgments. Q-Bench (Wu et al., 2023), LLVisionQA, LLDescribe, and LLaVA Bench-Wilder (Li et al., 2024a) evaluate alignment to human expectations via controlled perception-focused QA and description tasks, jointly reflecting visual ability and instruction following. Going beyond task performance, M-RewardBench (Gureja et al., 2024), MJ-Bench (Chen et al., 2024b), VL-RewardBench (Li et al., 2024c), and RewardBench (Lambert et al., 2024) assess alignment at the preference level, measuring how well multimodal reward models or LLM-as-judge systems approximate human preferences on safety, reliability, and bias in multi-task, multilingual settings.

4.2 Domain-specific Benchmarks

Domain-specific benchmarks narrow the evaluation scope and pose a more explicit question: Does the model meet the acceptable and useful standards of a given domain? These evaluations are inherently more domain-sensitive, typically grounded in concrete tasks and downstream use cases, and emphasize boundary conditions, risk profiles, and expert-defined criteria rather than broad average performance.

Medical and science. Evaluations in the medical and science fields prioritize expert-level accuracy and norms, characterized by an extremely low tolerance for error. Med-VQA (Canepa et al., 2023) and 3D-CT-GPT++ quantify proficiency in interpreting specialized imaging, whereas MMedPO and related methodologies define criteria for clinical preference alignment. Chen et al. ensure that model outputs strictly adhere to diagnostic logic and terminological standards. In immune adjuvants and materials chemistry, expert-designed standards and consistency validation were introduced to ensure the reliability of assessments.

Embodied AI. Benchmarks in the Embodied AI domain aim to validate closed-loop "perception-decision-action" capabilities, facilitating the transition from static visual processing to dynamic interaction. While SQA3D establishes rigorous standards for 3D spatial perception involving depth and occlusion, ASTRA (Wang et al., 2025a) and Engine scrutinize task planning and navigation by translating high-level instructions into executable actions, and the RL4VLM framework further validates reinforcement learning strategies based on environmental feedback.

Safety. Safety-oriented benchmarks prioritize stress-testing model defenses to construct robust adversarial protections and define the boundaries of harmlessness. VLGuard and VLSBench (Hu et al., 2024) quantify the ability to reject "Visual Jailbreak" attacks; RLAIF-V and HallusionBench focus on mitigating harmful multimodal hallucinations; and SafeVID extends these ethical compliance assessments to dynamic video generation.

5 Conclusion

MLLM alignment has progressed rapidly in recent years. In this survey, we provide a systematic and comprehensive review of existing work from three complementary perspectives: technical pipelines, application scenarios, and evaluation methodologies. Due to space constraints, we place an extended discussion of future directions in the Appendix D. We hope this survey will serve as a useful reference and help catalyze further research on principled, reliable, and scalable MLLM alignment.

Limitations

The paper retrieval, inclusion, and exclusion processes were performed by a single reviewer (the study's first author). While we implemented rigorous procedures to ensure comprehensive coverage of published works, this approach inherently carries the risk of omitting potentially relevant studies. Furthermore, classification of papers into specific categories or citation implementation might contain inadvertent errors. Nevertheless, we have performed multiple verification steps throughout the analytical process to mitigate such limitations. Although minor inconsistencies or omissions may persist, we maintain that this survey constitutes the most comprehensive review of MLLM alignment currently available, offering an objective and detailed assessment of future research directions and outstanding challenges.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Shuai Bai et al. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Louisa Canepa, Sonit Singh, and Arcot Sowmya. 2023. Visual question answering in the medical domain. *arXiv preprint arXiv:2309.11080*.

Hao Chen, Wei Zhao, Yingli Li, Wenjun Li, Zhuoyi Li, Ning Zhu, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, Jun Liu, and Tuo Zhang. 2025a. 3d-CT-GPT++: Enhancing 3d radiology report generation with direct preference optimization and large vision-language models.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025b. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024b. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *Preprint*, arXiv:2407.04842.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent ai: Surveying the horizons of multimodal interaction. *Preprint*, arXiv:2401.03568.

SooHwan Eom, Jay Shim, Eunseop Yoon, Hee Suk Yoon, Hyeonmok Ko, Mark A. Hasegawa-Johnson, and Chang D. Yoo. 2025. SQuba: Speech mamba language model with querying-attention for efficient summarization.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. 2024a. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. 2025a. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Yichong Huang, Xiaoyu Shen, Xipeng Qiu, and See-Kiong Ng. 2025b. Mcm-dpo: Multifaceted cross-modal direct preference optimization for alt-text generation. *arXiv preprint arXiv:2510.00647*.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. 2025c. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *Preprint*, arXiv:2501.16629.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. 2024c. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. *Preprint*, arXiv:2411.10436.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *ICML*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. *Preprint*, arXiv:2308.06394.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2024. M-rewardbench: Evaluating reward models in multilingual settings. *Preprint*, arXiv:2410.15522.

Lehan He, Zeren Chen, Zhelun Shi, Tianyu Yu, Jing Shao, and Lu Sheng. 2025. Systematic reward gap optimization for mitigating vlm hallucinations. *arXiv preprint arXiv:2411.17265*.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. 2024. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multimodal large language models. *arXiv:2402.14683*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2025. Ai alignment: A comprehensive survey. *Preprint*, arXiv:2310.19852.

Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, Mohan Wang, Josef Dai, Tianyi Qiu, Hua Xu, Dong Li, Weipeng Chen, Jun Song, Bo Zheng, and Yaodong Yang. 2024. Align anything: Training all-modality models to follow instructions with language feedback. *Preprint*, arXiv:2412.15838.

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. *Preprint*, arXiv:2411.11909.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Preprint*, arXiv:2405.01483.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. 2025a. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*.

Songtao Jiang, Yan Zhang, Yeying Jin, Zhihang Tang, Yangyang Wu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025b. Hscr: Hierarchical self-contrastive rewarding for aligning medical vision language models. *arXiv preprint arXiv:2506.00805*.

Kechen Jiao, Zhirui Fang, Jiahao Liu, Bei Li, Zhongjian Qiao, Yaxin Xu, Yifan Zhu, Xinyu Liu, Jingang

Wang, and Xiu Li. 2025. InteractiveCOT: Aligning dynamic chain-of-thought planning for embodied decision-making.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. Llm post-training: A deep dive into reasoning large language models. *Preprint*, arXiv:2502.21321.

Kwai Keye Team, Biao Yang, Bin Wen, et al. 2025. Kwai Keye-VL technical report. *Preprint*, arXiv:2507.01949.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. *Preprint*, arXiv:2403.13787.

Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. 2024a. Phantom of latent for large language and vision models. *Preprint*, arXiv:2409.14713.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024b. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2024c. Vlind-bench: Measuring language priors in large vision-language models. *arXiv:2406.08702*.

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.

Dongping Li, Tielong Cai, Tianci Tang, Wenhao Chai, Katherine Rose Driggs-Campbell, and Gaoang Wang. 2025a. Emmoe: A comprehensive benchmark for embodied mobile manipulation in open environments. *Preprint*, arXiv:2503.08604.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *Preprint*, arXiv:2407.07895.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. 2024c. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.

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

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. 2023b. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *Preprint*, arXiv:2312.16217.

Xudong Li, Mengdan Zhang, Peixian Chen, Xiawu Zheng, Yan Zhang, Jingyuan Zheng, Yunhang Shen, Ke Li, Chaoyou Fu, Xing Sun, and Rongrong Ji. 2025b. Zooming from context to cue: Hierarchical preference optimization for multi-image mllms. *arXiv preprint arXiv:2505.22396*.

Yongqi Li, Lu Yang, Jian Wang, Runyang You, Wenjie Li, and Liqiang Nie. 2025c. Towards harmless multimodal assistants with blind preference optimization. *arXiv preprint arXiv:2503.14189*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025d. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *Preprint*, arXiv:2501.02189.

Jiafeng Liang, Shixin Jiang, Xuan Dong, Ning Wang, Zheng Chu, Hui Su, Jinlan Fu, Ming Liu, See-Kiong Ng, and Bing Qin. 2025a. Investigating and enhancing the robustness of large multimodal models against temporal inconsistency. *arXiv preprint arXiv:2505.14405*.

Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. 2025b. Large model empowered embodied ai: A survey on decision-making and embodied learning. *Preprint*, arXiv:2508.10399.

Chaohu Liu, Gui Tianyi, Yu Liu, and Linli Xu. 2024a. AdPO: Enhancing the adversarial robustness of large vision-language models with preference optimization.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Jiaming Liu, Chenxuan Li, Guanqun Wang, Lily Lee, Kaichen Zhou, Sixiang Chen, Chuyan Xiong, Jiaxin Ge, Renrui Zhang, and Shanghang Zhang. 2024b. Self-corrected multimodal large language model for end-to-end robot manipulation. *Preprint*, arXiv:2405.17418.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, Haibo Lu, and Jiankun Yang. 2025a. PPLLaVA: Varied video sequence understanding with prompt guidance.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. 2025b. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *arXiv preprint arXiv:2506.11712*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024c. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *Preprint*, arXiv:2410.17637.

Jinda Lu, Junkang Wu, Jinghan Li, Xiaojun Jia, Shuo Wang, YiFan Zhang, Junfeng Fang, Xiang Wang, and Xiangnan He. 2025. Dama: Data- and model-aware alignment of multi-modal llms. *Preprint*, arXiv:2502.01943.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2025. vVLM: Exploring visual reasoning in VLMs against language priors.

Feipeng Ma, Yizhou Zhou, Yueyi Zhang, Siying Wu, Zheyu Zhang, Zilong He, Fengyun Rao, and Xiaoyan Sun. 2024a. Task navigator: Decomposing complex tasks for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. Sqa3d: Situated question answering in 3d scenes. In *ICLR*.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024b. A survey on vision-language-action models for embodied ai. *Preprint*, arXiv:2405.14093.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *CoRR*.

OpenAI. 2024. Introducing openai o1-preview.

Timothy Ossowski, Jixuan Chen, Danyal Maqbool, Zefan Cai, Tyler Bradshaw, and Junjie Hu. 2025. Comma: A communicative multimodal multi-agent benchmark. *Preprint*, arXiv:2410.07553.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. *Preprint*, arXiv:2408.10433.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, Ethan Yeo, Eugenie Lamprecht, Qi Liu, Yuqi Wang, Eric Chen, Deyu Fu, Lei Li, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Mikel Artetxe, and Yi Tay. 2024. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *Preprint*, arXiv:2405.02287.

Han Qiu, Peng Gao, Lewei Lu, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2025. Spatial preference rewarding for mllms spatial understanding. *arXiv preprint arXiv:2510.14374*.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. *CoRR*.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. 2025. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1713–1735, Suzhou, China. Association for Computational Linguistics.

Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. 2025. Modularized self-reflected video reasoner for multimodal LLM with application to video question answering. In *Forty-second International Conference on Machine Learning*.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun MA, and Chao Zhang. 2025. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. *arXiv preprint arXiv:2502.11775*.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024. video-salmonn: Speech-enhanced audio-visual large language models. In *International Conference on Machine Learning*, pages 47198–47217. PMLR.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023a. Aligning large multimodal models with factually augmented rlhf. *Preprint*, arXiv:2309.14525.

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

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun MA, and Chao Zhang. 2025. Enhancing multimodal LLM for detailed and accurate video captioning using multi-round preference optimization.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zujun Ma, and Chao Zhang. 2024. Enhancing multimodal llm for detailed and accurate video captioning using multi-round preference optimization. *arXiv preprint arXiv:2410.06682*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *Preprint*, arXiv:2402.10176.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *Preprint*, arXiv:2406.11839.

Han Wang, Gang Wang, and Huan Zhang. 2025a. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. *arXiv preprint arXiv:2411.16721*.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024b. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *Preprint*, arXiv:2401.16158.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Siyin Wang, Zhaoye Fei, Qinyuan Cheng, Shiduo Zhang, Panpan Cai, Jinlan Fu, and Xipeng Qiu. 2025b. World modeling makes a better planner: Dual preference optimization for embodied task planning. *arXiv preprint arXiv:2503.10480*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024d. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *Preprint*, arXiv:2411.10442.

Weiyun Wang et al. 2025c. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xinming Wang, Jian Xu, Aslan H Feng, Yi Chen, Haiyang Guo, Fei Zhu, Yuanqi Shao, Minsi Ren, Hongzhu Yi, Sheng Lian, et al. 2025d. The hitchhiker's guide to autonomous research: A survey of scientific agents. *TechRxiv.August 07, 2025. DOI:10.36227/techrxiv175459840.02185500/V1*.

Xinming Wang, Jian Xu, Bin Yu, Sheng Lian, Hongzhu Yi, Yi Chen, Yingjian Zhu, Boran Wang, Hongming Yang, Han Hu, et al. 2025e. Mr-align: Metareasoning informed factuality alignment for large reasoning models. *arXiv preprint arXiv:2510.24794*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, and Cao Xiao. 2024e. Enhancing visual-language modality alignment in large vision language models via self-improvement. *Preprint*, arXiv:2405.15973.

Yixu Wang, Jiaxin Song, Yifeng Gao, Xin Wang, Yang Yao, Yan Teng, Xingjun Ma, Yingchun Wang, and Yu-Gang Jiang. 2025f. Safevid: Toward safety aligned video large multimodal models. *arXiv preprint arXiv:2505.11926*.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. Livebench: A challenging, contamination-free llm benchmark. *Preprint*, arXiv:2406.19314.

Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. 2025. Dissecting adversarial robustness of multimodal lm agents. *Preprint*, arXiv:2406.12814.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv:2309.14181*.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in lvlms. *arXiv:2406.16449*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan

Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *Preprint*, arXiv:2410.02712.

Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, Peiyang Xu, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Xuandong Zhao, Francesco Pinto, Zhen Xiang, Yu Gai, Zinan Lin, Dan Hendrycks, Bo Li, and Dawn Song. 2025. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. 2024. xgen-mm (blip-3): A family of open large multimodal models. *Preprint*, arXiv:2408.08872.

Bei Yan, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. 2024. Evaluating the quality of hallucination benchmarks for large vision-language models. *arXiv:2406.17115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024a. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Yulong Yang, Xinshan Yang, Shuaidong Li, Chenhao Lin, Zhengyu Zhao, Chao Shen, and Tianwei Zhang. 2024b. Security matrix for multimodal agents on mobile devices: A systematic and proof of concept study. *Preprint*, arXiv:2407.09295.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *Preprint*, arXiv:2404.16006.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. 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*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

Fatemeh Pesaran Zadeh, Yoojin Oh, and Gunhee Kim. 2025. Lpoi: Listwise preference optimization for vision language models. *arXiv preprint arXiv:2505.21061*.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Preprint*, arXiv:2405.10292.

Menghao Zhang, Huazheng Wang, Pengfei Ren, Kangheng Lin, Qi Qi, Haifeng Sun, Zirui Zhuang, Lei Zhang, Jianxin Liao, and Jingyu Wang. 2025a. Do LVLMs truly understand video anomalies? revealing hallucination via co-occurrence patterns. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. 2024a. Mavis: Mathematical visual instruction tuning with an automatic data engine. *Preprint*, arXiv:2407.08739.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024b. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

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. 2025b. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv*.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024c. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li,

Kun Wang, Qingsong Wen, Zhang Zhang, et al. 2024d. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025c. Boosting chart-to-code generation in mllm via dual preference-guided refinement. *arXiv preprint arXiv:2504.02906*.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. 2025d. Grape: Generalizing robot policy via preference alignment. *Preprint*, arXiv:2411.19309.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2024. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *Preprint*, arXiv:2311.16839.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. 2025. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. *arXiv preprint arXiv:2507.02664*.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2025a. Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. *arXiv preprint arXiv:2412.06141*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, De Wen Soh, and Jun Liu. 2025b. CPCF: A cross-prompt contrastive framework for referring multimodal large language models. In *Forty-second International Conference on Machine Learning*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

Appendix

A Preliminary of Optimization Paradigm

A.1 RLHF

RLHF is a technique that combines RL with human feedback to align AI models (such as LLMs) with human preferences or values. It is widely used in alignment tasks and has played a crucial role in fine-tuning generative models (e.g., ChatGPT). RLHF typically consists of three stages: pretrained model, reward modeling, and RL fine-tuning.

In the field of MLLM alignment, RLHF-based methods are relatively rare, with Fact-RLHF being a representative example. Fact-RLHF undergoes three training phases (supervised fine-tuning, human preference collection & preference modeling, and factually-augmented RLHF) to align with human preferences. In Table 1, we present the loss function of the Fact-RLHF algorithm and compare it with other alignment algorithms. The training objective of RLHF is as follows:

$$\mathcal{L}_{\text{RLHF}} = -\mathbf{E}_{(\mathcal{I},x)\in D, y\sim\pi_\phi(y|\mathcal{I},x)} [r_{\boldsymbol{\theta}}(\mathcal{I},x,y) - \beta\cdot\mathbb{D}_{KL}(\pi_\phi(y|\mathcal{I},x) \parallel \pi^{\text{INIT}}(y|\mathcal{I},x))], \tag{1}$$

where $\mathcal{L}$ denotes the loss function, $\mathcal{D}$ represents the dataset, $\mathcal{I}$ represents the image, x represents the question, y represents the response, $\mathbb{D}_{KL}$ represents the KL divergence, π_ϕ represents the trained MLLM policy, π^{INIT} represents the fixed initial policy model, $r_{\boldsymbol{\theta}}$ represents the reward model, and β represents a hyperparameter.

A.2 DPO

DPO is an emerging model alignment approach designed to replace traditional RLHF by directly optimizing model policies using preference data, eliminating the need for explicit reward model training or complex reinforcement learning algorithms. Its core idea is to transform human preferences into optimization objectives for probabilistic models, thereby simplifying the alignment process and improving efficiency.

Since DPO does not require an explicit reward model, it reduces the number of hyperparameters, making it widely adopted in the MLLM alignment field. In Table 1, we present the DPO loss function and all its variants in MLLM alignment, allowing readers to quickly and intuitively understand the differences and connections between different algorithms. The DPO training objective is defined as follows:

$$\mathcal{L}_{\text{dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}} [\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w|\mathcal{I},x))}{\pi_{\text{ref}}(y_w|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l|\mathcal{I},x))}{\pi_{\text{ref}}(y_l|\mathcal{I},x))})], \tag{2}$$

where $\mathcal{L}$ denotes the loss function, $\mathcal{D}$ represents the dataset, $\mathcal{I}$ represents the image, x represents the question, y_w, y_l represents the chosen and rejected responses respectively, $\mathbb{D}_{KL}$ represents the KL divergence, π_ϕ represents the trained MLLM policy, π_{ref} represents the reference MLLM policy, $r_{\boldsymbol{\theta}}$ represents the reward model, and β represents a hyperparameter, σ represents the sigmoid function.

B Background of MLLM Alignment

In this section, we will provide a brief explanation of the complete training process for MLLMs, which consists mainly of three phases: pre-training, instruction tuning, and alignment with human preference.

B.1 Pre-Training

The pre-training phase of MLLMs primarily aims to align the feature spaces of different modalities with that of the language model. The data used in this phase is typically simple caption data. For instance, image-caption pairs are commonly used for image/video understanding MLLMs (Bai et al., 2023; Chen et al., 2023), while speech data and transcriptions are used for speech understanding MLLMs (Fu et al., 2024a, 2025a). Through this pre-training phase, the model learns to understand inputs from various modalities.

B.2 Instruction Tuning

Building on the pre-training phase, the SFT phase aims to teach the model how to interact with humans by focusing on understanding questions and providing responses in a specified format, i.e., instruction-following ability. The data used in this phase is typically high-quality and diverse dialogue data. For example, in the commonly seen visual question answering (VQA) task, given an image and its corresponding instruction, the trained model will provide the correct answer for the task.

B.3 Alignment with Human Preference

Previous works have shown that SFT tends to make the model memorize training data and try to generalize across diverse scenarios (Chu et al., 2025). The alignment phase, typically involving reinforcement learning (RL) strategies, is crucial for generalizing to unseen domains. However, most multimodal models neglect this step (Wang et al., 2024c; Deitke et al., 2024; Chen et al., 2024c; Dai et al., 2024; Agrawal et al., 2024). The goals of the alignment stage are broad, such as reducing hallucinations (Zhang et al., 2024c; Lu et al., 2025), enhancing conversational abilities (Xiong et al., 2024), improving safety (Zong et al., 2024), strengthening the reasoning abilities (Wang et al., 2024d), improving capabilities for long-reasoning tasks like DeepSeek-R1 (Chen et al., 2025b), and overall MLLM performance (Zhang et al., 2025b). This phase usually uses pair data that incorporates human preference.

C Positioning Relative to Prior Works

Here, we reiterate the systematic and innovative aspects of our work compared to other surveys.

First, previous alignment review papers have primarily focused on AI (Ji et al., 2025) or LLM (Kumar et al., 2025) alignment, with no existing work specifically addressing MLLM alignment. Our paper is the first survey dedicated to MLLM alignment. Among related works, (Kumar et al., 2025) is the most relevant, but it provides only a broad overview of LLM post-training methods without delving into their specific characteristics, making it more of a general summary. In contrast, we offer a detailed analysis of each method, enabling readers to intuitively understand their distinctions and quickly grasp the fundamentals of MLLM alignment. While MLLMs demonstrate immense potential in handling complex tasks involving visual, auditory, and textual data, state-of-the-art MLLMs are rarely rigorously aligned with human preferences. Our paper aims to provide researchers with a comprehensive and systematic guide to entering the field of MLLM alignment.

Second, we systematically organize MLLM alignment algorithms based on their application scenarios, data construction, and evaluation methods, addressing key questions such as how algorithms are applied, how datasets are built, and how methods are evaluated. This forms a clear and structured framework for MLLM alignment:

(1) **Algorithms**: In Table 1, we establish a unified notation system to formally describe different algorithms, allowing readers to grasp their differences and connections quickly.

(2) **Datasets**: In Table 2, we summarize existing open-source datasets in terms of size, categories, response-generating models, data sources, and annotation models, providing readers with a clear and concise overview of MLLM alignment-specific datasets.

(3) **Technical Pipeline**: In Fig 1 and Fig 3, we summarize the end-to-end technical pipeline of MLLM alignment, covering benchmark-driven evaluation and preference construction, and highlight the key design choices and method characteristics at each stage to provide a practical reference for researchers.

Third, we thoroughly discuss emerging research directions and open challenges in MLLM alignment, such as visual challenges, comprehensive evaluation, full-modality alignment, MLLM reasoning, insights from LLM alignment, and MLLM as agents. These are critical issues uniquely faced by MLLM alignment today, and our paper is the first to systematically propose them in Appendix D. By doing so, we aim to inspire potential research ideas and foster growth within the alignment community.

D Future Directions

Despite rapid progress in multimodal post-training, several open challenges remain in scaling alignment beyond image–text, transferring reasoning-oriented RL or data advances to MLLMs, mitigating overoptimization and reward hacking, and enabling robust, secure agentic behaviors under multimodal interaction. Given these circumstances, we outline key directions spanning full-modality alignment, reasoning-centric optimization, insights

from LLM alignment, and MLLM-based agents.

Full-modality alignment Align-anything(Ji et al., 2024) pioneers full-modality alignment through the multimodal dataset "align-anything-200k", which spans text, images, audio, and video. This study demonstrates the complementary effects between different modalities. However, their work is still in its early stages. The dataset for each modality is relatively small, limiting its ability to cover a wide range of tasks. Additionally, the proposed algorithm is only a preliminary improvement on the DPO method, and it does not fully exploit the unique structural information inherent in each modality. Moving forward, the design of alignment algorithms beyond image/text domains, particularly for other modalities, to enhance multimodal model capabilities, will be a key trend.

MLLM reasoning Recent advancements in reasoning LLMs, such as OpenAI's O1 and DeepSeek-R1, highlight the importance of RL algorithms and preference data in enhancing performance in complex tasks. Key insights can be categorized as follows: (1) *Data*: (a) *Scale & Quality*: From small-model resampling (e.g., OpenMathInstruct (Toshniwal et al., 2024)) to large-scale synthetic data (e.g., Qwen-2.5-MATH (Yang et al., 2024a)), datasets now include millions of samples. (b) *Efficiency*: Approaches like "less is more" alignment (e.g., LIMA (Zhou et al., 2023)) demonstrate that minimal, high-quality data can optimize pretrained capabilities. (2) *Optimization Framework*: (a) *Sampling Strategies*: Online RL techniques (e.g., DeepSeek V3 (DeepSeek-AI, 2024)) mitigate distributional shifts. (b) *Training Paradigms*: Multi-stage, collaborative optimization (e.g., Llama 3's DPO iteration) improves model performance. (c) *Algorithms*: Advancements in PPO techniques, such as DPO and GRPO, focus on reducing parameter count and refining reward functions (e.g., PRIME (Cui et al., 2025)). These trends emphasize efficiency, generalization, and precision in unlocking LLMs' reasoning potential.

Insight from LLM alignment The development of LLM alignment highlights three key insights and opportunities for improvement: (1) training efficiency—PPO-based methods require simultaneous loading of policy and reference models, slowing training; reference-free approaches like SimPO (Meng et al., 2024) could accelerate optimization by eliminating dependency on reference models, though their role in MLLM alignment needs deeper analysis. (2) overoptimization mitigation (Gao et al., 2023; Rafailov et al., 2024)—DPO/RLHF risks reward hacking where proxy metrics improve while real-world performance degrades, exacerbated by biased or low-quality data. Solutions include diversifying training datasets, early stopping, and regularization to balance generalization. Addressing these challenges requires rethinking optimization architectures, robust data curation, and synergistic integration of RL paradigms.

MLLM as agents Combine the advanced reasoning capabilities of LLMs with multimodal perception—encompassing text, images, and audio—enabling cross-modal knowledge synthesis and task decomposition for complex real-world applications (Xi et al., 2023; Wang et al., 2024b; Ma et al., 2024b; Durante et al., 2024; Ma et al., 2024a). These capabilities position MLLMs as promising agents for various domains, such as autonomous driving and industrial robotics (Li et al., 2023b; Liu et al., 2024b). However, designing MLLMs as effective agents presents several unresolved challenges: (1) *Multi-agent Collaboration*: Lack of mature frameworks for multimodal communication (Ossowski et al., 2025), shared memory, and coordination in MLLM-based multi-agent systems. (2) *Robustness*: Vulnerability to adversarial attacks (e.g., image perturbations hijacking agent behavior (Wu et al., 2025)) in open environments, necessitating systematic robustness testing and defense mechanisms. (3) *Security*: Expanded attack surfaces across multimodal perception, reasoning, and memory modules, requiring comprehensive safeguards against privacy breaches and malicious hijacking (Yang et al., 2024b).

E Leveraging Visual Information for Alignment

For clarity, we use the following notation to represent the composition of current alignment data: preference data $\mathcal{D} = (x, \mathcal{I}, y_w, y_l)$, where x is the question, $\mathcal{I}$ is the image, and y_w and y_l represent the winning and losing responses, respectively. In current research, three main approaches are employed to leverage visual information in order to enhance alignment performance, though each has its limitations:

1. **Using corrupted or irrelevant images as alignment phase negative samples.** Researchers create new images $\mathcal{I}_{neg}$ and use $(y_w|x, \mathcal{I}_{neg})$ as a negative sample. This approach improves MLLM robustness to differ-

ent images and reduces hallucinations. However, visual negatives often rely on diffusion algorithms or image modifications that lack robust quality metrics, incurring high computational costs.

2. **Generating new questions and answers based on corrupted images.** In this method, researchers create a new image $\mathcal{I}_{neg}$, use it to generate additional response y_{neg}, and then treat $(y_{neg}|x, \mathcal{I})$ as a negative sample. This method also essentially compares textual outputs, but it adds more variety to the textual comparison. However, the process of generating additional negative samples incurs extra computational overhead.

3. **Using cosine similarity metrics from models like CLIP to assess text-image matching.** This approach uses a similarity score between the text and the image to filter data or as part of the reinforcement learning reward function. While this can help reduce data noise, the quality of the score depends on the evaluation model's quality, which may be subject to model bias.

Each of these methods plays a role in enhancing MLLM alignment with visual data, but they come with trade-offs in terms of efficiency, cost, and the potential for biases.

Method	Loss
Fact-RLHF	$\mathcal{L}_{\text{RLHF}} = -\mathbf{E}_{(\mathcal{I},x)\in D, y\sim\pi_\phi(y\|\mathcal{I},x)}[r_{\boldsymbol{\theta}}(\mathcal{I},x,y) - \beta\cdot\mathbb{D}_{KL}(\pi_\phi(y\|\mathcal{I},x)\parallel\pi^{\text{INIT}}(y\|\mathcal{I},x))]$
SILKIE SIMA CLIP-DPO RLAIF-V 3D-CT-GPT++ MAVIS EMMOE xGen-MM(BLIP-3) LLaVA-NeXT-Interleave LLAVA-CRITIC SQuBa PPLLaVA HDPO SymDPO INTERACTIVECOT SAE-V TPR SAFEVIS CPCF LLaVA-Reasoner-DPO OmniAlign-V AIGI-Holmos IPA MM-IFEngine SPR C-DPO Modified DPO	$\mathcal{L}_{\text{dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$
RLHF-V	$\mathcal{L}_{\text{Dense-dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\mathbb{I}_{y_i\notin y_u}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$ $+\mathbb{I}_{y_i\in y_u}[\gamma\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]]$
F-DPO	$\mathcal{L}_{\text{Fine grained-dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))}) - \log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$
HA-DPO & MSR-ViR	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathbf{E}_{(\mathcal{I},x,y)\sim\mathcal{D}_{\text{SFT}}}[-\log\text{P}(\text{y}\|\mathcal{I},\text{x};\pi_{\boldsymbol{\theta}})]$
MIA-DPO	$\text{Loss} : \mathcal{L} = \mathcal{L}_{\text{dpo}} + \gamma\cdot\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[-\log(\text{y}_\text{w}\|\mathcal{I},\text{x})]$
CHiP	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathcal{L}_{\text{Image dpo}} + \lambda\cdot\mathcal{L}_{\text{Sentence dpo}} + \gamma\cdot\mathbf{E}_{(\mathcal{I},x,y_w^{\text{Token}},y_l^{\text{Token}})\sim\mathcal{D}_{\text{Token}}}$ $\beta\mathbb{D}_{\text{SeqKL}}[\pi_{\text{ref}}(y_w\|\mathcal{I},x)\parallel\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)] - \beta\mathbb{D}_{\text{SeqKL}}[\pi_{\text{ref}}(y_l\|\mathcal{I},x)\parallel\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)]$
Image DPO	$\mathcal{L}_{\text{Image dpo}} = -\mathbf{E}_{(\mathcal{I}_w,\mathcal{I}_l,x,y_w)}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_l,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_l,x))})]$
AdPO	$\mathcal{L} = -\mathbf{E}_{(\mathcal{I}_w,\mathcal{I}_l,x,y_w,y_l)}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I}_l,x))}{\pi_{\text{ref}}(y_l\|\mathcal{I}_l,x))})]$ $+\Sigma_{t=1}^{\text{T}}\log\pi_{\boldsymbol{\theta}}(y_w^t\|\mathcal{I}_l, x_t^{1:t-1})$
PHANTOM	$\mathcal{L} = \mathcal{L}_{\text{SFT}} - \mathbf{E}_{(\mathcal{I}_w,\mathcal{I}_l,x,y_w)}[\log\sigma(\frac{\beta}{\|y_w\|}\log\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x)) - (\frac{\beta}{\|y_w\|}\log\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))]$
video-SALMONN 2	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \lambda\mathbf{E}_{(\mathcal{I},x,y_{\text{gt}})\sim\mathcal{D}_{\text{gt}}}\log\pi_{\boldsymbol{\theta}}(y_{\text{gt}}\|\mathcal{I},x)$
Preference Optimization	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \lambda\mathbf{E}_{(\mathcal{I},x,y)\sim\mathcal{D}_{\text{reg}}}[\log\frac{\pi_{\boldsymbol{\theta}}(\text{y}\|\text{x})}{\pi_{\text{ref}}(\text{y}\|\text{x})}]$
DAMA	$\mathcal{L} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\alpha\cdot\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \alpha\cdot\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$
mDPO LPOI	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathbf{E}_{(\mathcal{I}_w,\mathcal{I}_l,x,y_w,y_l)\sim\mathcal{D}}[-\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I}_l,x))}{\pi_{\text{ref}}(y_l\|\mathcal{I}_l,x))})]$ $-\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x))} - \delta)$
MPO	$\mathcal{L} = \alpha_1\cdot\mathcal{L}_{\text{dpo}} - \alpha_2\cdot\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)}{\pi_{\text{ref}}(y_w\|\mathcal{I},x)} - \delta)\right]$ $-\alpha_2\cdot\left[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)}{\pi_{\text{ref}}(y_l\|\mathcal{I},x)} - \delta)\right] - \alpha_3\cdot\left[\frac{\log\pi_{\text{ref}}(y_w\|\mathcal{I},x)}{\|y_w\|}\right]$
CcDPO MMedPO MFPO HSCR IMG BPO	$\mathcal{L}_{\text{Image-Text DPO}} = \alpha_1\cdot\mathcal{L}_{\text{DPO}} + \alpha_2\cdot\mathcal{L}_{\text{Image DPO}}$
PanoDPO	$\mathcal{L}_{\text{Image-2-Text DPO}} = \mathcal{L}_{\text{Image-Text DPO}} + \alpha_3\cdot -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x+c))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x+c))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$
D^2PO PAR	$\alpha_1\mathcal{L}_{\text{Action DPO}} + \alpha_2\mathcal{L}_{\text{State DPO}}$ $\mathcal{L}_{\text{simPO}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\beta\log\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)) - \beta\log\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))]$
DCD	$\mathcal{L}_{\text{DCD}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\pi_\theta(y_l\|x,\mathcal{I}_l) + \log\pi_\theta(y_w\|x,\mathcal{I}_w)]$
fDPO	$\mathcal{L}_{\text{fDPO}} = \beta_1\cdot\mathcal{L}_{\text{DescDPO}} + \beta_2\cdot\mathcal{L}_{\text{ReasonDPO}}$
Pair-DPO	$\mathcal{L}_{\text{Pair-dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\Delta\text{Und}\Delta\text{Gen})]$, $\Delta\text{Und} = \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|,x))}{\pi_{\text{ref}}(y_w\|x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|,x))}{\pi_{\text{ref}}(y_l\|x))}, \Delta\text{Gen} = \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|,x))}{\pi_{\text{ref}}(y_w\|x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|,x))}{\pi_{\text{ref}}(y_l\|x))}$

Table 1: Various preference optimization objectives given preference data $\mathcal{D} = (x, \mathcal{I}, y_w, y_l)$, where x is the question, $\mathcal{I}$ is the Image, and y_w and y_l are winning and losing responses.

Dataset	Size	Categories	Response Model	Data Sources	Annotation Model
LLaVA-RLHF	10K	Hallucination	LLaVA-SFT	LLaVA-Instruct	Human
RLHF-V	1.4K	Hallucination	Muffin	UniMM-Chat	Human
VLFeedback	80K	Hallucination	12 Models	9 Datasets	GPT-4
CLIP-DPO	750K	Hallucination	MobileVLM-v2	12 Datasets	CLIP
M-HalDetect	16K	Hallucination	InstructBLIP	MS COCO	Human
HA-DPO	6K	Hallucination	3 Models	Visual Genome	GPT-4
SIMA	17K	Hallucination	LLaVA-1.5	LLaVA-Instruct	LLaVA-1.5
RLAIF-V	83K	Hallucination	3 Models	7 Datasets	2 Models
xGen-MM (BLIP-3)	62.6K	Hallucination	xGen-MM-4B	open-source	-
MIA-DPO	52K	Multi-Image	LLaVa-v1.5 & InternLM-XC 2.5	Not mentioned	Not mentioned
MAVIS	88K	Math	MAVIS-7B	Self-Constructed	GPT-4
EMMOE-100	10K	Embodied AI	Video-LLaVA	Self-Constructed	GPT-4
Image-DPO	60K	visual reasoning	Cambrian-8B & LLaVA-1.5	3 Datasets	Stable Diffusion
LLAVA-CRITIC	40.1K	Multiple tasks	LLaVA-OneVision	3 Datasets	LLaVA-OneVision
MMPR	3.25M	Reasoning	InternVL2-8B	Not mentioned	automate pipeline
video-SALMONN-o1	100K	Reasoning	Gemini-1.5-pro	Self-Constructed	GPT4
MMedPO	19 K	Medical	Med-LVLM & GPT-4o	Self-Constructed	GPT-4o
CPCF	50K	Image Referring	Fine-tuned Forret	Open Images	MLLM Itself
SHAREGPT-4O-REASONING	193K	Reasoning	Fine-tuned LLaVA-Next	9 Datasets	Golden Answer
OmniAlign-V	205K	Multiple tasks	LLaVA-Next & GPT-4o	Self-Constructed	automate pipeline
D^2PO	15K	Embody AI	automate pipeline	Self-Constructed	automate pipeline
Holmes-DPOSet	49K	Multiple tasks	Fine-tuned MLLMs	Self-Constructed	automate pipeline
Pick-a-Pic	851K	Generation	Self-Generated	Generative Models	
IPA	89K	Instruction Following	Qwen2VL-7B	15 Datasets	LLM-as-judge
MM-IFInstruct	23k	Instruction Following	Intern-VL-2.5 7B	automate pipeline	
PAR	5K	General	LLaVA-1.5 7B & Qwen-VL-chat-9B	LLaVA-80K	KL divergence
SPR	10K	Spatial Understanding	Ferret	Object365	automate pipeline
SENTINEL	8.6K	Hallucination	LLaVA-1.5	Visual Genome	automate pipeline
TAlt & PAlt	18K	alt-text generation	Gemini	Social Networks	automate pipeline
Chart2Code	23K	code generation from chart	GPT-4o	ReachQA	Reward Models
MMSafe-PO	5.6K	Safety	human & MLLM	Anthropic-HH	Human
Fin-APT	470	Financial Advisory Videos	Gemma2-9B	Youtube	GPT-4o
Pair-DPO	5K	Multimodal Understanding and Generation	Show-o	Journey-DB	automate pipeline
SafeVid-350K	350K	Video Understanding Safety	LLaVA-Next & GPT-4	Self-Constructed	automate pipeline
TPR	20K	hallucination	LLaVA-1.5-7B	7 datasets	LLaVA-Next-34B
MultiScope	42k	Multi-Image	3 datasets	LLaVA-OneVision	automated-pipeline

Table 2: Preference optimization dataset construction, including dataset, data size, categories: usage of the data, data sources, response model: the model to generate responses y_w and y_l by given image $\mathcal{I}$ and prompt x, and annotation model: the model to annotate y_w and y_l.