Multimodal LLM Alignment: Challenges, Solutions, and Research Opportunities

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated impressive potential in handling complex tasks involving visual, auditory, and textual data. However, critical issues related to truthfulness, safety, and alignment with human preference remain insufficiently addressed. This gap has spurred the emergence of various alignment algorithms. Recent studies have shown that alignment algorithms are a powerful approach to resolving the aforementioned challenges. In this paper, we aim to provide a comprehensive and systematic review of MLLM alignment algorithms. Specifically, we address four critical questions: (1) What application scenarios do existing alignment algorithms cover? (2) How are alignment datasets constructed? (3) How are alignment algorithms evaluated? (4) What are the future directions for the development of alignment algorithms? This work seeks to help researchers organize current advancements in the field and inspire better alignment methods.

1 Introduction

Large language models (LLMs) have ushered in a new era for artificial intelligence (AI), demonstrating remarkable abilities such as instruction-following and few-shot learning (Brown et al., 2020), which stem from their extensive model parameters and vast training data. These models represent a paradigm shift from traditional, task-specific models, as LLMs can handle a wide variety of general tasks with a simple prompt, without the need for task-specific training. This capability has fundamentally changed the AI landscape. However, while LLMs excel in text processing, they are limited by their inability to process multimodal data. Our world, on the other hand, is inherently multimodal, comprising visual, auditory, and other forms of data. This limitation has inspired the development of MLLMs (Fu et al., 2024a), which extend LLMs by incorporating the ability to process and understand multimodal data. MLLMs open up new opportunities for applications that require the integration and understanding of multiple types of data, expanding the potential of AI.

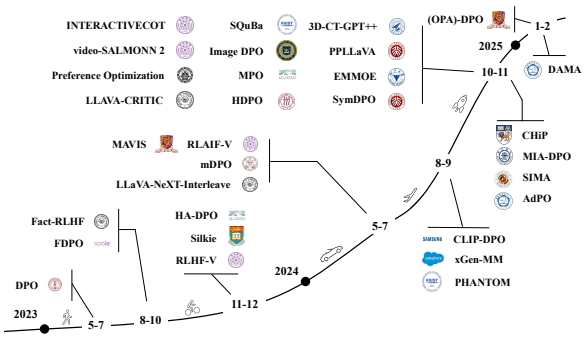

Figure 1: A timeline of MLLM alignment algorithms

Despite the impressive potential demonstrated by MLLMs in tackling complex tasks that involve visual, auditory, and textual data, the current state-of-the-art MLLMs have rarely undergone rigorous alignment with human preference (Figure 2) such as reinforcement learning from human preference (RLHF) stages (Wang et al., 2024c; Deitke et al., 2024; Chen et al., 2024e; Dai et al., 2024; Agrawal et al., 2024; Fu et al., 2025a) and direct preference optimization (DPO (Rafailov et al., 2024b)). Typically, these models only advance to the supervised fine-tuning (SFT) phase, with critical issues related to authenticity, safety, and alignment with human preferences remaining inadequately addressed. This gap has led to the emergence of various alignment algorithms, each targeting different application areas and optimization goals. However, this rapid development (Figure 1) also presents a number of challenges for researchers, particularly in areas such as benchmarking, optimizing alignment data, and introducing novel algorithms. In response, this paper provides a comprehensive and systematic review of alignment algorithms (Figure 3), focusing on the following four key questions:

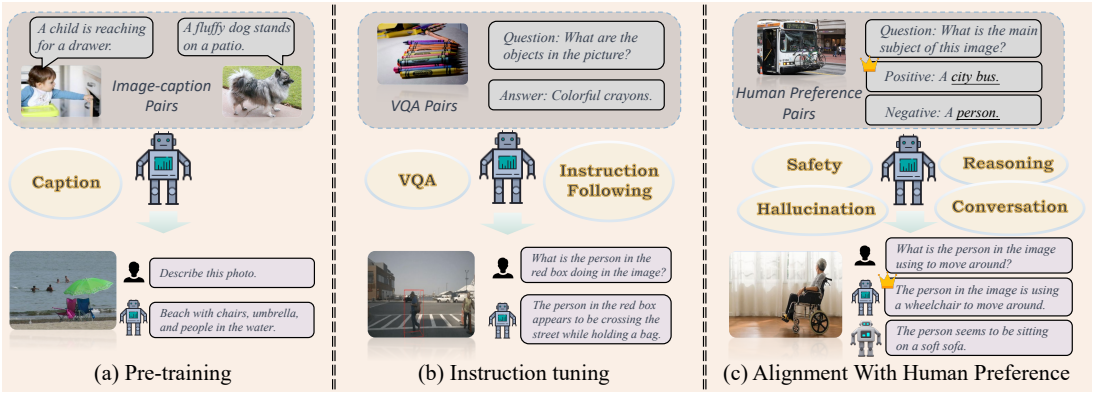

Figure 2: Comparison of pre-training, instruction tuning, and alignment with human preference.

- **What application scenarios do existing alignment algorithms cover?** We categorize current alignment algorithms based on their application scenarios, offering a clear framework for researchers across different domains. We also establish a unified symbolic system to aid researchers in understanding the distinctions between various algorithms, which is summarized in Table 1 of the appendix.

- **How are alignment datasets constructed?** The creation of alignment datasets involves three core factors: data sources, model responses, and preference annotations. We conduct a systematic analysis and categorization of these factors(publicly available datasets are summarized in appendix Table 2), highlighting the strengths and weaknesses of current dataset construction methods and emphasizing key considerations that must be addressed.

- **How are alignment algorithms evaluated?** Given that most alignment algorithms are designed for specific tasks—such as addressing hallucinations, ensuring safety, and improving reasoning—we categorize and organize common alignment algorithm benchmarks, providing a clear framework for evaluation. The full discussion of this section is provided in Appendix A due to space limitations.

- **What are the future directions for the development of alignment algorithms?** We propose several potential future directions, such as the integration of visual information into alignment algorithms, insights from LLM alignment methods, and the challenges and opportunities posed by MLLMs as agents.

Although many existing surveys focus on the alignment of AI (Ji et al., 2024a), none of them specifically address the alignment of MLLMs. To the best of our knowledge, this survey is the first to specifically focus on the alignment of MLLMs. Our objective is to provide a comprehensive and systematic guide for researchers in both academia and industry, helping them identify appropriate tools and methodologies in the rapidly evolving field of alignment algorithms.

2 Application Scenarios

Recent advancements in MLLM alignment algorithms have significantly expanded their applicability across a variety of domains. As illustrated in Figure 3, these methods can be categorized into three tiers based on their application scenarios: (1) general image understanding, (2) alignment algorithms designed for more complex modalities (such as multi-image, video, and audio), and (3) extended applications targeting domain-specific tasks. The first tier establishes the foundational principles of MLLM alignment. The second tier addresses the challenges of integrating more diverse and complex modalities, enabling more comprehensive multimodal interactions. Finally, the third tier focuses on adapting alignment frameworks to meet the specialized requirements of specific applications. Together, these tiers represent a structured and progressive framework for advancing multimodal intelligence and broadening its practical impact.

2.1 General Image Understanding

MLLM alignment algorithms are developed to address the issue of hallucinations in multimodal systems. Recent research shows that these algorithms not only improve performance in this regard but also enhance safety, conversational capabili-

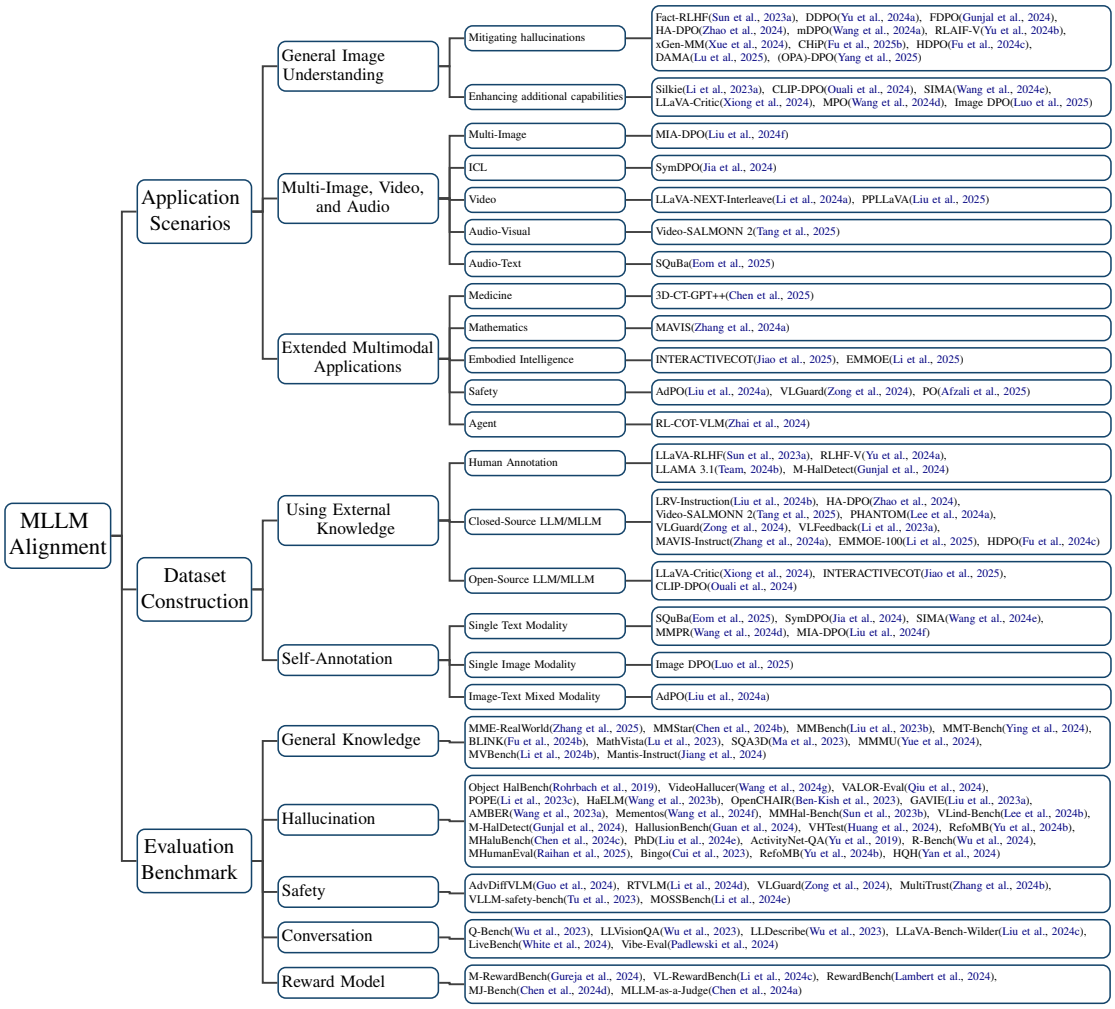

Figure 3: Categories of MLLM alignment benchmarks

ties, and a range of other functional attributes. In this section, we systematically examine innovative approaches, categorizing them based on their primary application scenarios: mitigating hallucinations and enhancing additional capabilities.

Mitigating hallucinations The original design intention of MLLM alignment algorithms is to mitigate hallucinations. Fact-RLHF (Sun et al., 2023a), the first multimodal RLHF method, integrates per-token KL penalties, factual calibration, and correctness/length constraints. DDPO (Yu et al., 2024a) focuses on fine-grained corrections (e.g., object, position, and number errors) by assigning weights to the revised data in its loss function. FDPO (Gunjal et al., 2024) modifies the architecture of InstructBLIP to obtain a clause/sentence-level reward model trained on human feedback data for the purpose of detecting hallucinations. HA-DPO (Zhao et al., 2024) leverages GPT-4 (Achiam et al., 2023) to verify whether the MLLM-generated descriptions contain hallucinations, utilizes GPT-4 to rewrite the positive and negative samples used for DPO, preventing distributional shifts. mDPO (Wang et al., 2024a) enhances DPO with a visual loss function (to counter visual information neglect) and anchoring (to prevent the decreasing in the probability of chosen response). RLAIF-V (Yu et al., 2024b) iterates DPO using GPT-4-labeled accuracy scores from open-source model responses. xGen-MM (Xue et al., 2024) employs a four-stage pipeline (pretraining, SFT, interleaved multi-image supervised fine-tuning, post-training) to holistically improve hallucinations, helpfulness, and safety. CHiP (Fu et al., 2025b) combines visual DPO (via diffused images) and hierarchical text preferences (response/segment/token levels) to refine alignment. HDPO (Fu et al., 2024c) specifically constructs hallucination-centric pairs (VDH, LCH, MCH) for targeted training. DAMA (Lu et al., 2025) refines DPO with data hardness and model responses by adaptively modifying β.

Enhancing additional capabilities In this subsec-

tion, we introduce several algorithms designed to enhance various aspects of model performance beyond hallucination reduction. For instance, Silkie aggregates responses from 12 models, evaluates them using GPT-4V, and applies DPO on diverse instruction datasets to improve perception, cognition, and faithfulness. CLIP-DPO (Ouali et al., 2024) leverages CLIP scores to label data and applies DPO loss, resulting in improvements in both hallucination mitigation and zero-shot classification tasks. SIMA (Wang et al., 2024e) constructs preference pairs by having the model self-evaluate its own responses. LLaVA-Critic (Xiong et al., 2024) employs an iterative DPO in which each round samples data from the model itself and uses a reward model to label preferences, thereby enhancing performance in hallucination reduction, image/video understanding, and open-ended dialogue. MPO (Wang et al., 2024d) automates the construction of a diverse multimodal reasoning preference dataset and blends SFT loss with various preference optimization losses, leading to improvements in reasoning. Finally, Image DPO (Luo et al., 2025) perturbs images (e.g., via blurring or pixelation) while keeping textual inputs unchanged, optimizing performance through visual-only DPO loss.

Current advancements in optimizing MLLM alignment algorithms primarily focus on two critical dimensions: data and loss functions. In the realm of data optimization, dominant strategies include manual annotation, strong model-generated data, and self-generation data. However, each of these approaches faces characteristic limitations. A persistent challenge lies in reducing annotation costs while simultaneously enhancing data quality and diversity. On the other hand, innovations in loss functions have introduced advanced variants of DPO, such as HDPO and DDPO, which demonstrate significant potential. Additionally, frameworks like Image DPO and CHiP incorporate vision-modality supervision, underscoring the importance of cross-modal alignment. Moving forward, progress in this field will hinge on two critical areas: improving data quality and diversity and optimizing multimodal loss functions to achieve more robust and efficient alignment.

2.2 Multi-Image, Video, and Audio

Compared to single-image tasks, many natural scene tasks involve multiple images, videos, or audio, introducing not only richer contextual scenarios but also greater complexity. Addressing these challenges requires specialized architectural designs and domain-specific optimizations. For instance, multi-image tasks necessitate models capable of understanding the relationships between multiple inputs, while in-context learning (ICL) requires the extraction of relevant information from multiple contextually provided images. Similarly, video processing demands the ability to perceive and analyze a large sequence of frames, and the data format of audio streams differs significantly from visual modalities. To tackle these complexities, researchers are actively investigating novel architectural modifications and specialized training paradigms tailored to these multifaceted tasks.

Multi-image While existing open-source MLLMs perform well on single-image tasks, they often struggle with multi-image contextual understanding. MIA-DPO (Liu et al., 2024f) addresses the challenges of hallucination in multi-image scenarios by leveraging synthetic multi-image compositions and constructing preference data. Specifically, the method analyzes the model's attention patterns across multiple images to assign scores and extract positive-negative pairs. This approach not only achieves state-of-the-art performance on multi-image benchmarks but also maintains robustness in single-image tasks.

ICL Recent advancements in ICL for LLMs have inspired adaptations in MLLMs, but these models often suffer from textual over-reliance, which leads to the neglect of visual information. To address this issue, SymDPO (Jia et al., 2024) introduces semantic decoupling through few-shot demonstrations that include intentionally irrelevant text. This strategy reduces the dominance of the text modality, encouraging models to prioritize visual-evidential reasoning, thereby improving performance on tasks such as image captioning and VQA.

Video Video understanding introduces greater risks of hallucinations compared to image-based tasks due to the added complexity of temporal dynamics. However, DPO-based alignment methods have demonstrated effectiveness in mitigating these errors. Current advancements adopt two strategic pathways: Interleaved Visual Instruction Tuning (e.g., LLaVA-NeXT-Interleave (Li et al., 2024a)), which enhances multi-frame reasoning by combining interleaved visual instructions with DPO loss; Granular Video-Text Alignment (e.g., PPLLaVA (Liu et al., 2025)), employing fine-grained vision-prompt alignment, context length expansion via asymmetric positional encoding, and DPO opti-

mization. These frameworks advance the performance of MLLMs on video tasks.

Audio-visual While real-world videos typically contain audio, existing MLLMs lack audio processing capabilities. Video-SALMONN 2 (Tang et al., 2025) addresses audio modality blindness in MLLMs through a hierarchical framework: (1) audio-visual representation alignment via an audio aligner, (2) semantic fusion through joint audio-visual SFT, (3) generation optimization using multi-round reinforcement learning(RL), and (4) capability restoration via "Rebirth" fine-tuning with self-generated high-quality data, enhancing audio-visual understanding in video analysis.

Audio-text Abstract speech summarization struggles with redundancy in outputs. SQuBa (Eom et al., 2025) overcomes this through a three-phase framework: (1) aligning speech-text representations via ASR-focused projector training, (2) jointly fine-tuning LLM and projector, (3) using the SFT responses and answers generated by the fine-tuned model as pairs for DPO. This phased optimization synergizes speech understanding and conciseness while preserving inference efficiency.

The application of alignment algorithms in emerging multimodal domains is still in its early stages, highlighting two critical areas for exploration: designing task-specific data for novel fields and developing alignment algorithms that leverage the structural properties of specific modalities.

2.3 Extended Multimodal Applications

Most MLLMs are not originally designed with specific downstream tasks in mind, such as medical diagnostics, mathematical reasoning, embodied AI, safety-critical systems, and autonomous agents. However, their powerful multimodal processing capabilities have drawn significant interest from researchers and practitioners across various fields. Recently, several alignment-related frameworks have been proposed to better adapt these models to downstream tasks. It is worth noting that these domain-specific applications exhibit substantial gaps compared to general image understanding tasks, necessitating specialized alignment paradigms to address their unique operational constraints and ethical considerations.

Medicine The deployment of MLLMs in clinical settings is often hindered by the high risk of erroneous medical diagnoses or other domain-specific errors. The 3D-CT-GPT++ framework (Chen et al., 2025) addresses this issue through a DPO-based approach, utilizing GPT-4 to score SFT model-generated medical reports and construct preference datasets for alignment. This human-free method significantly reduces diagnostic misalignments while achieving clinical-grade accuracy and coherence in AI-assisted imaging analysis.

Mathematics MLLMs struggle with math-vision integration due to dual challenges: insufficient domain-optimized training frameworks and fragile chain-of-thought(CoT) reasoning where minor errors trigger cascading solution failures. MAVIS (Zhang et al., 2024a) addresses challenges in multimodal mathematical reasoning by enhancing MLLMs through a four-phase framework: (1) fine-tuning a math-specialized vision encoder through contrastive learning; (2) align the encoder with LLM; (3) Instruction tuning strengthens step-by-step reasoning; (4) DPO refines logical coherence by aligning annotated CoT paths. This integrated approach achieves high performance in visual mathematical problem-solving benchmarks.

Embodied intelligence Embodied intelligence research leverages MLLMs to advance agents' reasoning through CoT optimization and hierarchical task decomposition. INTERACTIVECOT (Jiao et al., 2025) enhances contextual reasoning via dynamic CoT optimization with domain-specific fine-tuning and real-time interaction feedback, boosting task success; EMMOE (Li et al., 2025) decomposes complex tasks into 966 subtasks, leveraging GPT-4 to create semantic-augmented datasets that improve embodied metrics like path efficiency. Together, they demonstrate how adaptive reasoning architectures and structured multimodal data engineering bridge the gap between semantic interpretation and actionable decision-making in embodied AI.

Safety The advancement of MLLMs introduces adversarial risks (e.g., harmful hallucination generation), several works propose their own solutions.: AdPO (Liu et al., 2024a) strengthens robustness through contrastive DPO training on perturbed images, enhancing the resistance to attacks; VLGuard (Zong et al., 2024) curates multimodal harmful content datasets and employs post-hoc fine-tuning to suppress unsafe behavior. In contrast, Preference Optimization (PO) (Afzali et al., 2025) frames contrastive learning as a one-step Markov decision process, combining preference data for discrimination and regularization data for stability, primarily boosting robustness. These methods synergize adversarial resilience and safety alignment to address evolving security threats.

Agent The application of MLLMs in multi-step interactive decision-making is often limited, preventing their direct application in complex decision-making scenarios. To address this limitation, existing work (Zhai et al., 2024) introduces a proximal policy optimization (PPO (Schulman et al., 2017))-driven alignment framework designed to optimize MLLMs for multi-round interactive decision-making. This approach effectively bridges the gap between semantic comprehension and actionable agent behaviors in dynamic, real-world scenarios.

Future breakthroughs in domain-specialized MLLMs The development of domain-specialized MLLMs will likely be driven by a synergistic co-evolution of alignment frameworks and domain-specific expertise. By tailoring alignment architectures to leverage the unique attributes and constraints of specific domains (e.g., healthcare, robotics, mathematics), these models can achieve greater effectiveness and precision.

3 MLLM Alignment Dataset

In this section, we classify existing MLLM alignment datasets into two categories based on their construction approach: datasets that introduce external knowledge and those that rely on self-annotation. Table 2 of the appendix presents crucial information about publicly available datasets, including data sources, response generation methods, annotation techniques, and dataset sizes, providing a convenient reference for researchers.

3.1 Introducing External Knowledge

Introducing high-quality external knowledge during data construction can enhance the quality of the generated alignment data. However, balancing data quality, quantity, and cost is a key consideration. Several works have explored data construction based on external knowledge.

Human annotation Multiple datasets employ distinct human annotation strategies for training: LLaVA-RLHF (Sun et al., 2023a) collects 10k examples by having annotators select positive/negative responses from model-generated pairs. RLHF-V (Yu et al., 2024a) creates 1.4k positive examples by manually correcting hallucinated responses. LLAMA 3.1 (Team, 2024b) incorporates 7-point ratings and optional human edits for "chosen" responses from a model pool. M-HalDetect (Gunjal et al., 2024) introduces clause-level hallucination analysis (16k examples) to synthesize preference data but remains in the exploratory stage.

Closed-source LLM/MLLM As the best-performing MLLMs currently available, GPT-4 series models have achieved near-human accuracy across many tasks. To reduce costs, current methods use them for preference data construction. LRV-Instruction (Liu et al., 2024b) uses GPT-4 to create 400k diverse visual instructions to mitigate hallucinations. HA-DPO repurposes MLLM outputs into aligned positive/negative pairs (10k examples), maintaining distribution consistency. Video-SALMONN 2 employs GPT-3.5/4o and Gemini-1.5-Pro (Team, 2024a) for caption generation. PHANTOM (Lee et al., 2024a) extracts 2.8M ambiguous negative examples via GPT-4o-mini, filtered with GPT-4o. VLGuard generates 3k safety-focused instruction-response pairs using GPT-4V, proposing post-hoc fine-tuning. Task-specific datasets include VLFeedback (Li et al., 2023a) (80k GPT-4V-scored responses across 12 MLLMs), MAVIS-Instruct (Zhang et al., 2024a) (math CoT preference data), and EMMOE-100 (Li et al., 2025) (3.7k SFT data and 10k DPO data).

Open-source LLM/MLLM Considering the invocation time and cost of GPT-4 series models in constructing large-scale alignment data, current methods use open-source models for preference data construction. INTERACTIVECOT builds an agent in ALFWorld(Shridhar et al., 2021) using predefined scores for embodied intelligence preference datasets. CLIP-DPO replaces MLLM evaluations with CLIP scores to select DPO pairs and constructs a 750k dataset (mixed QA/caption pairs).

Overall, manual annotation ensures high-quality, preference-aligned data but is constrained by challenges such as subjectivity and high costs. Both closed-source models (e.g., GPT-4V) and open-source models reduce costs and enable the large-scale construction of datasets; however, they often compromise on data quality. Looking ahead, we look forward to the development of more efficient methods that can achieve a balance between scalability and data reliability.

3.2 Self-Annotation

Data generated with the assistance of humans or models like GPT-4 may exhibit significant distributional differences from the target model, leading to issues such as overlooking image details.(Zhou et al., 2024) As a result, several approaches have emerged that do not rely on external models for data generation or reward signals, instead depend-

ing on the target model itself to construct preference pairs. Based on the modality differences in preference pair data, we categorize them into three types: single-text modality (where preference pairs differ only in the text modality), single-image modality (where preference pairs differ only in the image modality), and image-text mixed modality (where preference pairs differ in both modalities).

Single text modality SQuBa uses SFT data as questions and positive samples, and employs the responses generated by the fine-tuned model as negative samples for DPO. SymDPO reorganizes VQA/classification data into ICL format with meaningless text symbols to enhance visual learning and select DPO pairs. SIMA avoids the use of third-party data and models by having the model evaluate its own generated responses to rank the answers. MMPR (Wang et al., 2024d) uses the model's responses generated based on images as positive examples, and truncates these positive examples to create negative samples by continuing the response without providing the image. MIA-DPO concatenates single-image data into multi-image formats and selects preferences via attention values, improving multi-image task performance.

Single image modality Image DPO constructs DPO preference pairs by perturbing images (e.g., gaussian blur, or pixelation) while keeping text unchanged, creating negative examples through image-text mismatches.

Image-text mixed modality AdPO aligns adversarial training with DPO by constructing preference pairs from original/adversarial images (generated via methods like PGD) and their model responses, where both images and text differ between positive and negative examples during optimization.

The construction of self-annotated positive and negative samples helps mitigate distribution shifts. However, due to performance limitations of MLLMs, current data quality remains relatively low. We look forward to future developments will introduce technologies such as data enhancement specifically designed for self-annotation approaches to improve data quality.

4 Future Work and Open Challenges

As MLLMs evolve, aligning them with human preference has become a key focus. However, several challenges remain in evaluating these alignment algorithms. First, there is a lack of unified, high-quality, and diverse datasets. Existing alignment datasets often define different capability dimensions, leading to inconsistencies across studies. Second, most methods fail to effectively utilize visual information, relying mainly on text for positive and negative sample classification, and using simple loss functions like DPO without fully leveraging the multimodal nature of the data. Finally, there is a lack of comprehensive evaluation standards, with current methods often validated only on limited datasets like hallucination or dialogue, making it difficult to assess their generalizability. Further more, by drawing on advancements in LLMs and agent research, we can identify issues and limitations in current MLLM approaches. Addressing these challenges is crucial for developing more powerful and holistic alignment methods.

Data challenges The alignment of MLLMs faces two critical data-related challenges: data quality and coverage. First, the availability of high-quality MLLM alignment data is limited. Compared to LLMs, acquiring and annotating multimodal data is significantly more complex due to the inherent difficulties of handling multiple modalities. Second, existing datasets lack sufficient coverage of diverse multimodal tasks, such as OCR. Constructing a comprehensive dataset that addresses this wide array of tasks is extremely challenging, and to the best of our knowledge, no publicly available, fully human-annotated multimodal dataset exceeds 100,000 samples. These limitations in data quality and coverage pose significant barriers to effectively aligning MLLMs with human preference.

Visual information There are three methods that leverage visual information to enhance alignment performance, but all of them have certain limitations: (1) visual loss function—while positive samples are diverse, visual negatives often rely on diffusion algorithms or image modifications lacking robust quality metrics, incurring high computational costs; (2) methods of using text responses generated from visual negative samples as negative samples (Fu et al., 2024c) also fails to address the core challenge of constructing meaningful visual negative samples; (3) DPO data filtering using cosine similarity metrics from models like CLIP introduces vision-text alignment biases and quality uncertainties, limiting generalizability.

Comprehensive evaluation Current research on MLLM alignment focuses on a limited set of tasks. Most studies primarily evaluate their algorithms on a few key areas, such as hallucination detection, conversational abilities, and safety. However, we

argue that aligning MLLMs with human preference should not be restricted to these specific tasks. Future research should adopt a more comprehensive evaluation approach, assessing alignment methods across a broader range of tasks to better demonstrate their generalizability and effectiveness.

Full-modality alignment Align-anything(Ji et al., 2024b) 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., 2024a)—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., 2024d). 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).

5 Conclusion

The field of MLLM alignment is developing rapidly. In this paper, we conduct a systematic and comprehensive survey of existing research on MLLM alignment, focusing on four key questions: what application scenarios can be covered, how to construct datasets, how to evaluate algorithms, and where the direction of the next alignment lies. This paper is the first systematic survey dedicated to MLLM alignment. We hope that this survey will facilitate further research in this area.

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*.

Amirabbas Afzali, Borna khodabandeh, Ali Rasekh, Mahyar JafariNodeh, Sepehr Kazemi Ranjbar, and Simon Gottschalk. 2025. Aligning visual contrastive learning models via preference optimization. In *The Thirteenth International Conference on Learning Representations*.

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*.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2023. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv:2312.03631*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *Preprint*, arXiv:2402.04788.

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. 2025. 3d-CT-GPT++: Enhancing 3d radiology report generation with direct preference optimization and large vision-language models.

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

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Jinjie Gu, and Huajun Chen. 2024c. Unified hallucination detection for multimodal large language models. *arXiv:2402.03190*.

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. 2024d. 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. 2024e. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv:2311.03287*.

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, 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*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024a. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. 2025b. 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*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

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

Qi Guo, Shanmin Pang, Xiaojun Jia, and Qing Guo. 2024. Efficiently adversarial examples generation for visual-language models under targeted transfer scenarios using diffusion models. *arXiv:2404.10335*.

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.

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, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2024a. 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. 2024b. 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.

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.

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.

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

Dongping Li, Tielong Cai, Tianci Tang, Wenhao Chai, Katherine Rose Driggs-Campbell, Hongwei Wang, and Gaoang Wang. 2025. Homiebot: an adaptive system for embodied mobile manipulation in open environments.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. 2024c. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *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. *Preprint*, arXiv:2312.10665.

Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024d. Red teaming visual language models. *Preprint*, arXiv:2401.12915.

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.

Xirui Li, Hengguang Zhou, Ruochen Wang, Tianyi Zhou, Minhao Cheng, and Cho-Jui Hsieh. 2024e. Mossbench: Is your multimodal language model oversensitive to safe queries? *arXiv:2406.17806*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *EMNLP*.

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. 2024b. Mitigating hallucination in large multi-modal models via robust instruction tuning. *Preprint*, arXiv:2306.14565.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

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

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024e. Phd: A prompted visual hallucination evaluation dataset. *arXiv:2403.11116*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. 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. 2024f. 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*.

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.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *arXiv:2404.13874*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024b. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. mhumaneval – a multilingual benchmark to evaluate large language models for code generation. *Preprint*, arXiv:2410.15037.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. *Preprint*, arXiv:2010.03768.

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.

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

Llama3 Team. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2023. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *Preprint*, arXiv:2311.16101.

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.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv:2311.07397*.

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.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023b. Evaluation and analysis of hallucination in large vision-language models. *arXiv:2308.15126*.

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*.

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.

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.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024f. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv:2401.10529*.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024g. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv:2406.16338*.

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.

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.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. *Preprint*, arXiv:2501.09695.

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, and Tat-Seng Chua. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *Preprint*, arXiv:2312.00849.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. *Preprint*, arXiv:2405.17220.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*.

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*.

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.

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.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *Preprint*, arXiv:2408.13257.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. 2024b. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Preprint*, arXiv:2406.07057.

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.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *Preprint*, arXiv:2402.11411.

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. *Preprint*, arXiv:2402.02207.

A Evaluation

Existing MLLM alignment evaluation benchmarks are categorized into five key dimensions: general knowledge (assessing foundational capabilities), hallucination (measuring the inconsistency of generated content with facts), safety (evaluating the ability to mitigate risks in responses), conversation (testing whether the model can output the content required by users), and reward model (evaluating the performance of the reward model).

A.1 General Knowledge

Most benchmarks prioritize high-quality, human-annotated datasets tailored for real-world applications. Examples include MME-RealWorld's (Zhang et al., 2025) 29K QA pairs from 13K images and MMMU's (Yue et al., 2024) 11.5K questions from academic sources. MMStar (Chen et al., 2024b) enhances reliability by minimizing data leakage and emphasizing visual dependency. Many benchmarks introduce novel methodologies, such as MMBench's (Liu et al., 2023b) bilingual evaluation with CircularEval, MMT-Bench's (Ying et al., 2024) task graphs for in/out-of-domain analysis, and BLINK's (Fu et al., 2024b) focus on visual perception tasks. These frameworks enhance evaluation precision and reveal model limitations. Tasks often require advanced multimodal reasoning, such as MathVista's (Lu et al., 2023) mathematical-visual integration, SQA3D's (Ma et al., 2023) 3D situational QA, and MMMU's coverage of charts, and maps. These benchmarks push models to handle interdisciplinary challenges. By curating challenging, fine-grained tasks (e.g., temporal understanding in MVBench (Li et al., 2024b), multi-image processing in Mantis-Instruct (Jiang et al., 2024)), these benchmarks aim to advance models' ability to solve real-world problems requiring nuanced perception and reasoning.

A.2 Hallucination

These benchmarks systematically identify and categorize hallucinations in multimodal models, including object hallucinations (Object HalBench (Rohrbach et al., 2019)), intrinsic and extrinsic hallucinations (VideoHallucer (Wang et al., 2024g)), and associative biases (VALOR-Eval (Qiu et al., 2024)). They emphasize granular evaluation across visual, textual, and sequential contexts. Many propose novel frameworks, such as polling-based queries (POPE (Li et al., 2023c)), LLM-driven scoring (HaELM (Wang et al., 2023b), RefoMB (Yu et al., 2024b)), open-vocabulary detection (OpenCHAIR (Ben-Kish et al., 2023)), annotation-free assessment (GAVIE (Liu et al., 2023a)), LLM-free pipelines (AMBER (Wang et al., 2023a)), and GPT-4-assisted reasoning analysis (Mementos (Wang et al., 2024f)). They emphasize automated, scalable evaluation while addressing limitations like data leakage (MMHal-Bench (Sun et al., 2023b)) and language priors (VLind-Bench (Lee et al., 2024b)). Datasets prioritize fine-grained

human annotations (M-HalDetect (Gunjal et al., 2024), HallusionBench (Guan et al., 2024)) and synthetic data generation (VHTest (Huang et al., 2024), MHaluBench (Chen et al., 2024c)). They balance real-world complexity (PhD's (Liu et al., 2024e) counter-commonsense images, ActivityNet-QA's (Yu et al., 2019) 58K QA pairs) and controlled challenges (R-Bench's (Wu et al., 2024) robustness analysis). Some target specialized tasks like multilingual support (MHumanEval (Raihan et al., 2025)), while others address broad issues like bias and interference (Bingo (Cui et al., 2023)). All aim to enhance model robustness in practical scenarios. By proposing alignment strategies (RLAIF-V's (Yu et al., 2024b) open-source feedback) and proposing unified framework (HQH (Yan et al., 2024)), these benchmarks guide the development of more reliable multimodal systems.

A.3 Safety

Several introduce novel techniques, such as diffusion-based adversarial attacks (AdvDiffVLM (Guo et al., 2024)), red teaming frameworks (RTVLM (Li et al., 2024d)), and post-hoc fine-tuning strategies (VLGuard (Zong et al., 2024)). These approaches enhance evaluation rigor by simulating real-world threats or improving model resilience. Benchmarks like MultiTrust (Zhang et al., 2024b) and RTVLM unify trustworthiness assessment across multiple dimensions (e.g., truthfulness, fairness), while others target specific challenges like OOD generalization (VLLM-safety-bench (Tu et al., 2023)) or oversensitivity (MOSSBench (Li et al., 2024e)). Together, they provide holistic insights into model limitations.

A.4 Conversation

These benchmarks prioritize evaluating foundational visual skills, such as low-level perception ability (Q-Bench (Wu et al., 2023), LLVisionQA (Wu et al., 2023)), description ability on low-level information (LLDescribe (Wu et al., 2023)), and quality assessment. They emphasize the model's ability to interpret and articulate fine-grained visual information. Several benchmarks test generalization to challenging scenarios, including unconventional images (LLaVA Bench-Wilder (Liu et al., 2024c)), cross-domain tasks (LiveBench's (White et al., 2024) math/news integration), and adversarial prompts (Vibe-Eval's (Padlewski et al., 2024) high-difficulty questions). They reveal model adaptability beyond standard datasets.

A.5 Reward Model

Each benchmark targets specific evaluation dimensions, such as multilingual capabilities (23 languages in M-RewardBench (Gureja et al., 2024)), alignment/safety/bias (MJ-Bench (Chen et al., 2024d)), and ability of MLLMs in assisting judges across diverse modalities (MLLM-as-a-Judge's (Chen et al., 2024a) scoring vs. pairwise comparisons). These frameworks reveal model strengths and weaknesses in structured and out-of-distribution scenarios. High-quality datasets are curated through human-AI collaboration (VL-RewardBench's (Li et al., 2024c) annotation pipeline) or structured triplet designs (RewardBench (Lambert et al., 2024)). Tasks range from simple preference ranking to complex reasoning, pushing models to handle nuanced challenges like hallucination detection and ethical alignment.

Overall, for MLLM alignment algorithms, many current works focus on their ability to prevent models from generating hallucinations, while also exploring how to leverage alignment algorithms to enhance MLLMs' general knowledge and conversation capability, which is an important direction for the future. Some researchers treat unsafe responses as misaligned with human preferences, thereby applying MLLM Alignment algorithms to address safety issues. Additionally, the effectiveness of reward models in these frameworks, particularly their performance in guiding alignment, warrants further investigation.

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	$\mathcal{L}_{\text{dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$
RLHF-V	$\mathcal{L}_{\text{Dense-dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)}[\mathbb{I}_{y_i\notin y_u}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$ $+\mathbb{I}_{y_i\in y_u}[\gamma\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]]$
F-DPO	$\mathcal{L}_{\text{Fine grained-dpo}} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)}[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x))}{\pi_{\text{ref}}(y_w\|\mathcal{I},x))}) - \log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x))}{\pi_{\text{ref}}(y_l\|\mathcal{I},x))})]$
HA-DPO	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathbf{E}_{(\mathcal{I},x,y)\sim\mathcal{D}_{\text{SFT}}}[-\text{logP}(\text{y}\|\mathcal{I},\text{x};\pi_{\boldsymbol{\theta}})]$
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	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathcal{L}_{\text{visual-dpo}} + \lambda \cdot \mathcal{L}_{\text{sentence-dpo}} + \gamma \cdot \mathbf{E}_{(\mathcal{I},x,y_w^{\text{Token}},y_l^{\text{Token}})\sim\mathcal{D}_{\text{Token}}}$ $\beta\mathbb{D}_{\text{SeqKL}}\left[\pi_{\text{ref}}(y_w\|\mathcal{I},x) \parallel \pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)\right] - \beta\mathbb{D}_{\text{SeqKL}}\left[\pi_{\text{ref}}(y_l\|\mathcal{I},x) \parallel \pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)\right]$
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	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathbf{E}_{(\mathcal{I}_w,\mathcal{I}_l,x,y_w,y_l)\sim\mathcal{D}}[-\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x))} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I}_l,x))}{\pi_{\text{ref}}(y_l\|\mathcal{I}_l,x))})]$ $-\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x))}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x))} - \delta)$
MPO	$\mathcal{L} = \alpha_1\cdot\mathcal{L}_{\text{dpo}} - \alpha_2\cdot\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)}{\pi_{\text{ref}}(y_w\|\mathcal{I},x)} - \delta)\right]$ $-\alpha_2\cdot\left[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)}{\pi_{\text{ref}}(y_l\|\mathcal{I},x)} - \delta)\right] - \alpha_3\cdot\left[\frac{\log\pi_{\text{ref}}(y_w\|\mathcal{I},x)}{\|y_w\|}\right]$

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

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.