Aligning Multimodal LLM with Human Preference: A Survey

Tao Yu, Yi-Fan Zhang†, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, Yibo Yan, Tianlong Xu, Qingsong Wen, Zhang Zhang, Yan Huang, Liang Wang, *Fellow, IEEE* and Tieniu Tan, *Fellow, IEEE*

Abstract—Large language models (LLMs) can handle a wide variety of general tasks with simple prompts, without the need for task-specific training. Multimodal Large Language Models (MLLMs), built upon LLMs, have demonstrated impressive potential in tackling complex tasks involving visual, auditory, and textual data. However, critical issues related to truthfulness, safety, o1-like reasoning, and alignment with human preference remain insufficiently addressed. This gap has spurred the emergence of various alignment algorithms, each targeting different application scenarios and optimization goals. Recent studies have shown that alignment algorithms are a powerful approach to resolving the aforementioned challenges. This paper aims to provide a comprehensive and systematic review of alignment algorithms for MLLMs. Specifically, we explore four key aspects: 1) the application scenarios covered by alignment algorithms, including general image understanding, multi-image, video, and audio, and extended multimodal applications; 2) the core factors in constructing alignment datasets, including data sources, model responses, and preference annotations; 3) the benchmarks used to evaluate alignment algorithms; and 4) a discussion of potential 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. The project page of this paper is available at https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Alignment.

Index Terms—Multimodal Large Language Model, MLLM, Alignment, Alignment with Human Preference.

1 Introduction

LLMs have ushered in a new era for artificial intelligence (AI), demonstrating remarkable abilities such as instruction-following and few-shot learning [1], 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. The world, on the other hand, is inherently multimodal, comprising visual, auditory, textual and other forms of data. This limitation has inspired the development of MLLMs [2], 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.

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 stage such as reinforcement learning from human preference (RLHF) stages [3], [4], [5], [6], [7], [8] and direct preference optimization (DPO [9]). Specifically, these models only advance to the supervised fine-tuning (SFT) phase, with critical issues related to authenticity, safety, and alignment with human preference 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, focusing on the following four key questions:

- **What application scenarios do existing alignment algorithms cover?** We categorize current alignment algorithms based on their application scenarios in Figure 2, offering a clear framework for researchers across different domains. We also establish a unified symbolic system to aid researchers in understanding the distinctions and connections between various algorithms, which is summarized in Table 1.
- **How are alignment datasets constructed?** The creation of alignment datasets involves three core factors: data sources, model responses, and preference annotations. We perform a systematic analysis and categorization of these factors (publicly available datasets are summarized in Table 2), highlighting the strengths and weaknesses of current construction methods and highlighting key considerations that need to be addressed.

- †*Yi-Fan Zhang is the project leader: yifanzhang.cs@gmail.com.*
- *Tao Yu, Yi-Fan Zhang, Zhang Zhang, Yan Huang, Liang Wang, Tieniu Tan are with the Institute of automation, Chinese academy of science. Chaoyou Fu is with the Nanjing University. Junkang Wu, and Jinda Lu are with the University of Science and Technology of China. Kun wang is with the Nanyang Technological University. Xingyu Lu is with the Shenzhen International Graduate School, Tsinghua University. Yunhang Shen is with the Tencent Youtu Lab. Guibin Zhang is with the National University of Singapore. Dingjie Song is with the Lehigh University. Yibo Yan is with the The Hong Kong University of Science and Technology. Tianlong Xu and Qingsong Wen are with the Squirrel Ai Learning.*

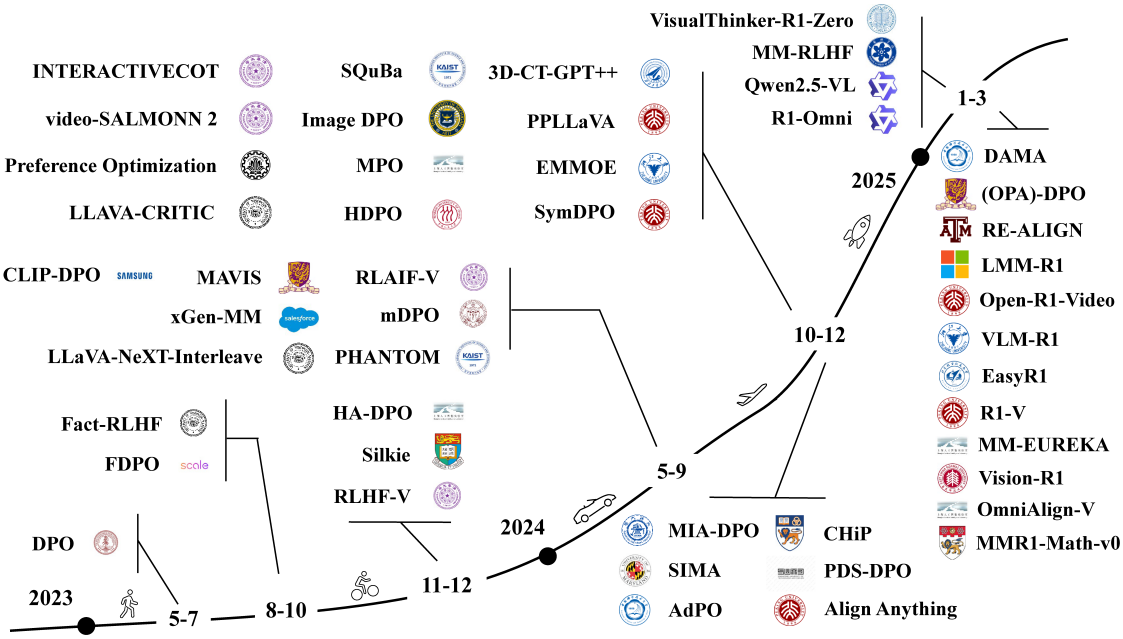

Fig. 1: A timeline of MLLM alignment algorithms.

- **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.
- **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 [10], [2], [11], 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 MLLM alignment.

2 Background: MLLM Alignment

In this section, we will provide a brief explanation of the complete training process for MLLMs, which consists mainly of three phases (Figure 3): pre-training, instruction tuning, and alignment with human preference.

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 [108], [109], while speech data and transcriptions are used for speech understanding MLLMs [110], [8]. Through this pre-training phase, the model learns to understand inputs from various modalities.

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.

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 [111]. The alignment phase, typically involving reinforcement learning (RL) strategies, is crucial for generalizing to unseen domains. However, most multimodal models neglect this step [3], [4], [5], [6], [7]. The goals of alignment stage are broad, such as reducing hallucinations [112], [21], enhancing conversational abilities [28], improving safety [113], strengthening the reasoning abilities [29], improving capabilities for long-reasoning tasks like DeepSeek-R1 [38], and overall MLLM performance [24]. This phase usually uses pair data that incorporates human preference.

3 Application Scenarios

Recent advancements in MLLM alignment algorithms have significantly expanded their applicability across a variety of domains. As illustrated in Figure 2, 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

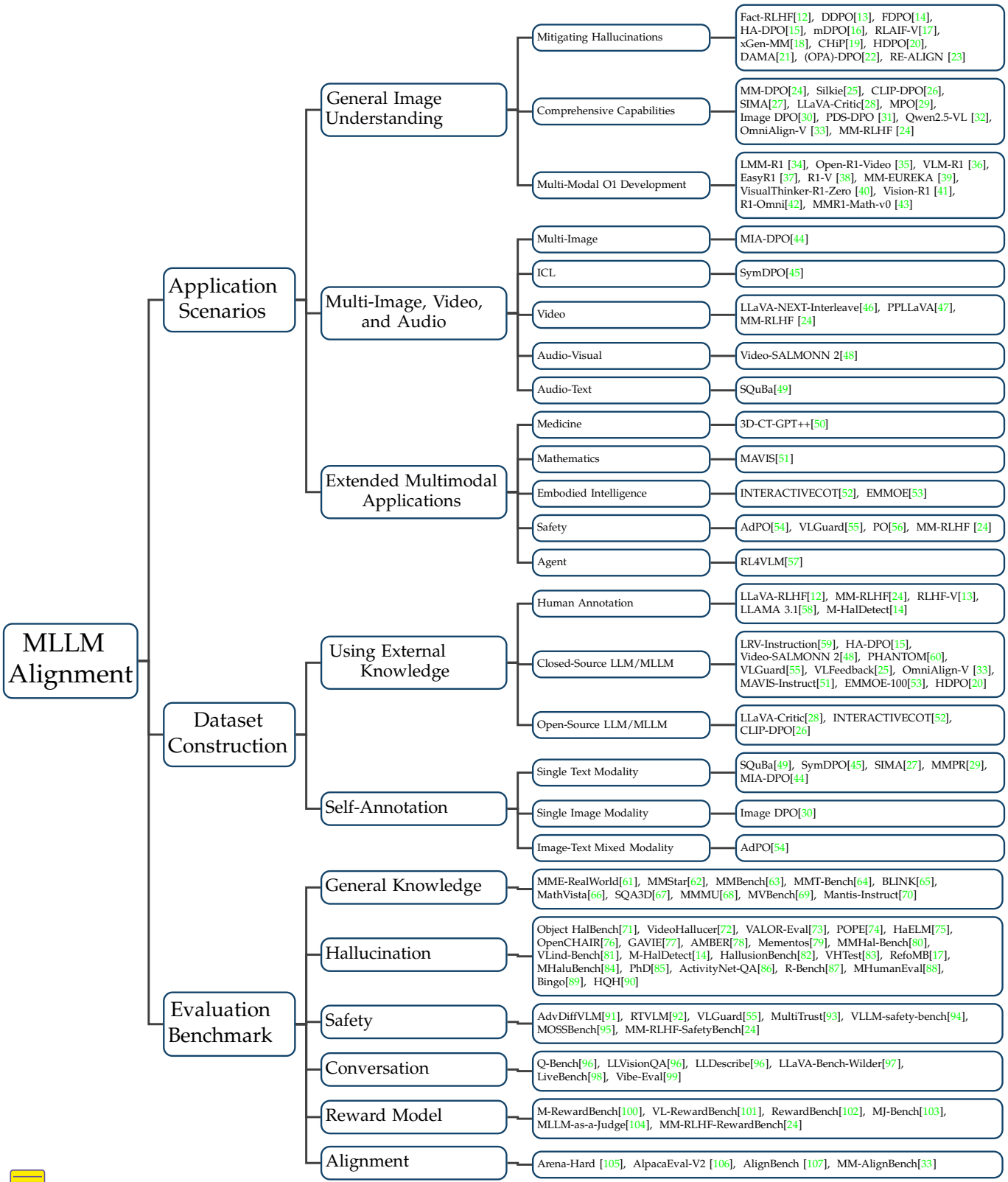

Fig. 2: Categories of MLLM alignment, including application scenarios, alignment datasets, and evaluation benchmarks.

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. We unify all existing methods using consistent notations, as shown in Table 1, making it easier for researchers to compare the differences and connections.

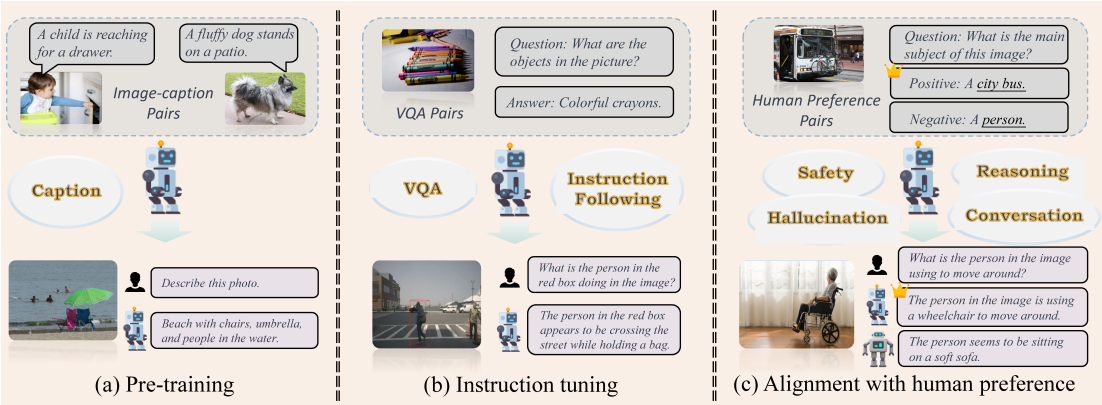

Fig. 3: Comparison of pre-training, instruction tuning, and alignment with human preference.

General Image Understanding

LLM alignment algorithms are developed to address the issue of hallucinations in multimodal systems. ent research shows that these algorithms not only improve performance in this regard but also enhance safety, conversational capabilities, reasoning abilities, and a range of other functional attributes. his section, we systematically examine innovative approaches, categorizing them based on their primary application scenarios: mitigating hallucinations and enhancing additional capabilities.

Mitigating Hallucinations

original design intention of MLLM alignment algorithms is to mitigate hallucinations. -RLHF [12] is the first multimodal RLHF algorithm, utilizing 10K human-labeled samples for the reward model and 50K hold-out data. oss function integrates: (per-token KL penalty; actual information to calibrate judgments; a (3) correctness and length penalties. PO [13] assigns higher weights to corrected data in its loss function compared to standard DPO. ses 1.4K manually refined samples covering hallucination types such as objects (41.2%), positions (20.3%), numbers (16.5%), attributes (10%), actions (5.3%), and others (6.8%). O [14] reuses InstructBLIP's [114] architecture by replacing the final embedding layer with a classification head to train clause/sentence-level reward models. r reward modeling and rejection sampling, it optimizes InstructBLIP with human-annotated data. -DPO [15] uses MLLM to generate image descriptions, validates them with GPT-4 for hallucinations, rewrites positive/negative samples for consistency, and adds an auxiliary causal language modeling loss to standard DPO. duces hallucinations and introduces the SHR metric (hallucinated sentences per total sentences). PO [16] enhances DPO with a visual loss function (to counter visual information neglect) and anchoring (to prevent the decreasing in the probability of chosen response). IF-V [17] generates multiple responses, splits each response into sentences, and reformulates each sentence as a question to query an open-source model for a trustworthiness score. total score for each response is then used to determine the DPO data. n-MM [18] employs a four-stage pipeline (pre-training, SFT, interleaved multi-image supervised fine-tuning, post-training) to holistically improve hallucinations, helpfulness, and safety. P [19] introduces visual preference optimization (e.g., diffusion, cropping, rotation for negative images) and hierarchical text preferences (response/segment/token levels). oss combines visual DPO and three text-level DPO terms, targeting hallucination reduction. PO [20] specifically constructs three hallucination-specific pairs: visual distracted hallucination, long context hallucination, and multimodal conflicts hallucination, aiming to reduce hallucinations. MA [21] refines DPO with data hardness and model responses by adaptively modifying β. ALIGN [23] intentionally injects controllable hallucinations into selected responses through image retrieval, generating rejected responses that provide more reasonable and natural preference signals regarding hallucinations. A)-DPO [22] prevents distribution shift by revising the responses generated by the pre-trained model using GPT-4V. revised responses are then mixed with the ground truth to construct a dataset. model is fine-tuned using LoRA-SFT [115], converting out-of-distribution data into in-distribution data.

Enhancing Comprehensive Capabilities

In this subsection, we introduce several algorithms designed to enhance various aspects of model performance beyond hallucination reduction. For instance, Silkie[25] collects a diverse set of instruction datasets and correspondingly generates responses from 12 models, which are then evaluated using GPT-4V to obtain preference data for applying DPO. CLIP-DPO [26] leverages CLIP [116] scores to label data and applies DPO loss, resulting in improvements in both hallucination mitigation and zero-shot classification tasks. SIMA [27] constructs preference pairs by having the model self-evaluate its own responses. LLaVA-Critic

[28] uses LLaVA-OV[117] to generate responses, fine-tunes a critic model (LLaVA-Critic) for scoring, and iteratively applies DPO, thereby enhancing performance in hallucination reduction, image/video understanding, and open-ended dialogue. MPO [29] automates the construction of a diverse multimodal reasoning preference dataset and blends SFT loss with several preference optimization losses, leading to improvements in reasoning. Finally, Image DPO [30] perturbs images (e.g., via blurring or pixelation) while keeping textual inputs unchanged, optimizing performance through visual-only DPO loss. MM-DPO [24] introduces a dynamic reward scale in DPO, where the reward model assigns higher weights to comparison pairs with larger reward margins during training. This ensures that the most informative samples have a greater impact on model updates. PDS-DPO [31] uses synthetic images generated by Stable Diffusion [118], which are evaluated by a pre-trained reward model. The highest-rated images, along with preference data generated by an open-source MLLM and scored by Llama-3-8B-ArmoRM [119], are used for DPO. This approach enhances the trustworthiness and reasoning capabilities of the MLLM. The DPO phase of Qwen2.5-VL [32] focuses on image-text data and pure text data, utilizing preference data to align the model with human preference, thereby enhancing the model's reasoning capabilities and task-specific performance. OmniAlign-V [33] improves the alignment of MLLM with human preference by constructing a comprehensive dataset featuring diverse images, complex questions, and varied response formats.

3.1.3 Multi-Modal O1 Development

Recently, the popularity of DeepSeek-R1 [120] has brought new inspiration to the MLLM community. LMM-R1 [34] is trained on a pure-text math dataset using RLOO [121] and showed improvements on multimodal math benchmarks. Open-R1-Video [35] utilizes GRPO [122] to enhance the model's performance in the video domain. VLM-R1 [36] applies the R1 method to the referring expression comprehension (REC) task. EasyR1 [37] proposed a multimodal RL training framework. R1-V [38] attempts to improve the generalization capabilities of MLLMs. MM-EUREKA [39] has built a scalable multimodal large-scale reinforcement learning framework based on OpenRLHF [123], offering stronger scalability compared to R1-V. VisualThinker-R1-Zero [40] successfully replicates R1's Aha Moment using only a 2B non-SFT model. Vision-R1 [41] is the first reasoning MLLM that combines cold-start and RL. R1-Omni [42] focuses on the emotion recognition task and is the industry's first application of reinforcement learning with verifiable reward (RLVR) with an omni-multimodal LLM. MMR1-Math-v0 [43] remarkably achieves top-tier performance with just 6k high-quality samples from public training datasets.

Current advancements in optimizing MLLM alignment algorithms primarily focus on two critical dimensions: data and loss functions. In the realm of preference data collection, 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 [20] and DDPO [13], which demonstrate significant potential. Additionally, frameworks like Image DPO [30] and CHiP [19] 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.

3.2 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 tasks.

3.2.1 Multi-Image

While existing open-source MLLMs perform well on single-image tasks, they often struggle with multi-image contextual understanding. MIA-DPO [44] addresses the challenges of hallucination in multi-image scenarios by concatenating unrelated single-image data into sequential, grid, and picture-in-picture images 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.

3.2.2 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 [45] employs the few-shot idea by replacing original text answers in examples with unrelated words. This modification reduces the influence of information provided by the text modality, encouraging the model to rely more on visual information for answers, thereby successfully improving performance on tasks such as image captioning and VQA.

3.2.3 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 [46]), which enhances multi-frame reasoning by combining interleaved visual instructions with

TABLE 1: Various preference optimization objectives given preference data $\mathcal{D} = (x, \mathcal{I}, y_w, y_l, r_w, r_l)$, where x is the question, $\mathcal{I}$ is the Image, y_w and y_l are winning and losing responses, and r_w and r_l are there rewards scored by a reward model.

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	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathbf{E}_{(\mathcal{I},x,y)\sim\mathcal{D}_{\text{SFT}}}[-\log\text{P}(\text{y}\|\mathcal{I},\text{x};\pi_{\boldsymbol{\theta}})]$
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	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathcal{L}_{\text{visual-dpo}} + \lambda \cdot \mathcal{L}_{\text{sentence-dpo}} + \gamma \cdot \mathbf{E}_{(\mathcal{I},x,y_w^{\text{Token}},y_l^{\text{Token}})\sim\mathcal{D}_{\text{Token}}}$ $\beta\mathbb{D}_{\text{SeqKL}}[\pi_{\text{ref}}(y_w\|\mathcal{I},x) \parallel \pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)] - \beta\mathbb{D}_{\text{SeqKL}}[\pi_{\text{ref}}(y_l\|\mathcal{I},x) \parallel \pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)]$
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	$\mathcal{L} = \mathcal{L}_{\text{dpo}} + \mathbf{E}_{(\mathcal{I}_w,\mathcal{I}_l,x,y_w,y_l)\sim\mathcal{D}}[-\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x)}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x)} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I}_l,x)}{\pi_{\text{ref}}(y_l\|\mathcal{I}_l,x)})]$ $-\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I}_w,x)}{\pi_{\text{ref}}(y_w\|\mathcal{I}_w,x)} - \delta)$
MPO	$\mathcal{L} = \alpha_1\cdot\mathcal{L}_{\text{dpo}} - \alpha_2\cdot\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)}{\pi_{\text{ref}}(y_w\|\mathcal{I},x)} - \delta)\right]$ $-\alpha_2\cdot\left[\log\sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)}{\pi_{\text{ref}}(y_l\|\mathcal{I},x)} - \delta)\right] - \alpha_3\cdot\left[\frac{\log\pi_{\text{ref}}(y_w\|\mathcal{I},x)}{\|y_w\|}\right]$
MM-RLHF	$\mathcal{L} = -\mathbf{E}_{(\mathcal{I},x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\,(r_w - r_l)\log\frac{\pi_{\boldsymbol{\theta}}(y_w\|\mathcal{I},x)}{\pi_{\text{ref}}(y_w\|\mathcal{I},x)} - \beta\,(r_w - r_l)\log\frac{\pi_{\boldsymbol{\theta}}(y_l\|\mathcal{I},x)}{\pi_{\text{ref}}(y_l\|\mathcal{I},x)}\right)\right]$

DPO loss; granular video-text alignment (e.g., PPLLaVA [47]), employing fine-grained vision-prompt alignment, context length expansion via asymmetric positional encoding, and DPO optimization. These frameworks advance the performance of MLLMs on video tasks.

3.2.4 Audio-Visual

While real-world videos typically contain audio, existing MLLMs lack audio processing capabilities. Video-SALMONN 2 [48] 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; and (4) capability restoration via "Rebirth" fine-tuning with self-generated high-quality data, enhancing audio-visual understanding capability in video analysis.

3.2.5 Audio-Text

Abstract speech summarization struggles with redundancy in outputs. SQuBa [49] 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.

3.3 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 domain-specific alignment 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.

3.3.1 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 [50] addresses this issue through a DPO-based approach, utilizing GPT-4 [124] 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.

3.3.2 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 [51] 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.

3.3.3 Embodied Intelligence

Embodied intelligence research leverages MLLMs to advance agents' reasoning through CoT optimization and hierarchical task decomposition. INTERACTIVECOT [52] enhances contextual reasoning via dynamic CoT optimization with domain-specific fine-tuning and real-time interaction feedback, boosting task success; EMMOE [53] 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.

3.3.4 Safety

The advancement of MLLMs introduces adversarial risks (e.g., harmful hallucination generation), several works propose their own solutions. AdPO [54] strengthens robustness through contrastive DPO training on perturbed images, enhancing the resistance to attacks. VLGuard [55] curates multimodal harmful content datasets and employs post-hoc fine-tuning to suppress unsafe behavior. In contrast, Preference Optimization (PO) [56] 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. MM-RLHF [24] artificially constructs datasets related to adversarial attacks, privacy, and security, and mix them with a large amount of general capability data. This approach simultaneously enhances both the model's security and its general capabilities, revealing that there is no strict trade-off between the two.

3.3.5 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 [57] introduces a proximal policy optimization (PPO [125])-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.

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, and agent), these multimodal models can get greater effectiveness and precision.

4 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 presents crucial information about publicly available datasets, including data sources, response generation methods, annotation techniques, and dataset sizes, providing a convenient reference.

4.1 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.

4.1.1 Human Annotation

Multiple datasets employ distinct human annotation strategies for training: LLaVA-RLHF [12] collects 10k examples by having annotators select positive/negative responses from model-generated pairs. RLHF-V [13] creates 1.4k positive

TABLE 2: MLLM alignment datasets, including data size, categories, data sources, response model: the model to generate responses for training by given image and prompt, and annotation model: the model to annotate the 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
MM-RLHF	120K	Hallucination Math, Video, Safety Conversation	GPT-4o, QwenVL2-72B, LLaVA-Video-72B, LLaVA-ov-72B Claude3.5-Sonnet	7 Dataset & Self-constrcuted	Human
Open-R1-Video-4k	4K	o1-Reasoning	GPT-4o	LLaVA-Video-178K	Not mentioned
MM-Eureka-Dataset	54K	o1-Reasoning	InternVL2.5-8B-instruct	15 Datasets	Not mentioned
MMR1-Math-RL-Data-v0	7K	o1-Reasoning	Not mentioned	Not mentioned	Not mentioned

examples by manually correcting hallucinated responses. LLAMA 3.1 [58] incorporates 7-point ratings and optional human edits for "chosen" responses from a model pool. M-HalDetect [14] introduces clause-level hallucination analysis (16k examples) to synthesize preference data but remains in the exploratory stage. MM-RLHF [24] covers three domains: image, video understanding, and MLLM safety. Through a rigorous pipeline construction, MM-RLHF ensures high-quality, fine-grained human annotations.

4.1.2 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 [59] uses GPT-4 to generate a large, diverse dataset of 400k visual instructions covering 16 vision and language tasks. The dataset includes both positive and negative data, but the positive and negative data are generated separately. HA-DPO [15] uses GPT-4 to modify negative responses from MLLMs into positive ones, and then has both the positive and negative responses further corrected by GPT-4 to ensure that the positive and negative examples remain within the same distribution. This method collects 10,000 data annotated by GPT-4. Video-SALMONN 2 [48] employs GPT-3.5/4o and Gemini-1.5-Pro [126] for caption generation. PHANTOM [60] extracts 2.8 million visual instruction-tuning data from multiple datasets, using GPT-4o-mini to generate ambiguous responses for queries as negative examples and filtering them with GPT-4o to improve data quality. PHANTOM's approach of generating ambiguous responses to obtain negative data is novel, but its effectiveness remains to be discussed. Task-specific datasets include VLFeedback [25] (80k GPT-4V-scored responses across 12 MLLMs), MAVIS-Instruct [51] (math CoT preference data), and EMMOE-100 [53] (3.7k SFT data and 10k DPO data of embodied AI). OmniAlign-V [33] filters images based on image complexity and the number of meaningful objects in the images, and generates positive pairs for DPO using GPT-4o.

4.1.3 Open-Source LLM/MLLM

Considering the invocation cost of GPT-4 series models in constructing large-scale alignment data, current methods use open-source models for preference data construction. INTERACTIVECOT [52] builds an agent in ALFWorld[127] using predefined scores for embodied intelligence preference datasets. CLIP-DPO [26] argues that scoring based on MLLMs lacks stable evaluation metrics, so it substitutes CLIP [116] scores with clear meanings 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 can achieve a balance between scalability and reliability.

4.2 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 [128]. As a result, several approaches have emerged that do not rely on external models for data generation or reward signals, instead depending 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).

4.2.1 Single Text Modality

SQuBa [49] uses SFT data as questions and positive samples, and employs the responses generated by the fine-tuned model as negative samples for DPO. SymDPO [45] reorganizes VQA/classification data into ICL format with meaningless text symbols to enhance visual learning and select DPO pairs. SIMA [27] avoids the use of third-party data and models by having the model evaluate its own generated responses to rank the answers. MMPR [29] 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 [44] concatenates single-image data into multi-image formats and selects preferences via attention values, improving multi-image task performance.

4.2.2 Single Image Modality

Image DPO [30] constructs DPO preference pairs by perturbing images (e.g., gaussian blur, or pixelation) while keeping text unchanged, creating negative examples.

4.2.3 Image-Text Mixed Modality

AdPO [54] aligns adversarial training with DPO by constructing preference pairs from original/adversarial images (generated via methods like PGD [129]) and their model responses, where both images and texts 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 automated data enhancement specifically designed for self-annotation approaches to improve data quality.

5 Evaluation

Existing MLLM alignment evaluation benchmarks are categorized into six 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), reward model (evaluating the performance of the reward model), and alignment with human preference.

5.1 General Knowledge

Most benchmarks prioritize high-quality, human-annotated datasets tailored for real-world applications. Examples include MME-RealWorld's [61] 29K QA pairs from 13K images and MMMU's [68] 11.5K questions from academic sources. MMStar [62] enhances reliability by minimizing data leakage and emphasizing visual dependency. Many benchmarks introduce novel methodologies, such as MMBench's [63] bilingual evaluation with CircularEval, MMT-Bench's [64] task graphs for in/out-of-domain analysis, and BLINK's [65] focus on visual perception tasks. These frameworks enhance evaluation precision and reveal model limitations. Tasks often require advanced multimodal reasoning, such as MathVista's [66] mathematical-visual integration, SQA3D's [67] 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 [69], multi-image processing in Mantis-Instruct [70]), these benchmarks aim to advance models' ability to solve real-world problems requiring nuanced perception and reasoning.

5.2 Hallucination

These benchmarks systematically identify and categorize hallucinations in multimodal models, including object hallucinations (Object HalBench [71]), intrinsic and extrinsic hallucinations (VideoHallucer [72]), and associative biases (VALOR-Eval [73]). They emphasize granular evaluation across visual, textual, and sequential contexts. Many propose novel frameworks, such as polling-based queries (POPE [74]), LLM-driven scoring (HaELM [75], RefoMB [17]), open-vocabulary detection (OpenCHAIR [76]), annotation-free assessment (GAVIE [77]), LLM-free pipelines (AMBER [78]), and GPT-4-assisted reasoning analysis (Mementos [79]). They emphasize automated, scalable evaluation while addressing limitations like data leakage (MMHal-Bench [80]) and language priors (VLind-Bench [81]). Datasets prioritize fine-grained human annotations (M-HalDetect [14], HallusionBench [82]) and synthetic data generation (VHTest [83], MHaluBench [84]). They balance real-world complexity (PhD's [85] counter-commonsense images, ActivityNet-QA's [86] 58K QA pairs) and controlled challenges (R-Bench's [87] robustness analysis). Some target specialized tasks like multilingual support (MHumanEval [88]), while others address broad issues like bias and interference (Bingo [89]). All aim to enhance model robustness in practical scenarios. By proposing alignment strategies (RLAIF-V's [17] open-source feedback) and proposing unified framework (HQH [90]), these benchmarks guide the development of more reliable multimodal systems.

5.3 Safety

Some studys introduce novel techniques, such as diffusion-based adversarial attacks (AdvDiffVLM [91]), red teaming frameworks (RTVLM [92]), and post-hoc fine-tuning strategies (VLGuard [55]). These approaches enhance evaluation rigor by simulating real-world threats or improving model resilience. Benchmarks like MultiTrust [93] and RTVLM

unify trustworthiness assessment across multiple dimensions (e.g., truthfulness, fairness), while others target specific challenges like out of distribution (OOD) generalization (VLLM-safety-bench [94]) or oversensitivity (MOSSBench [95]). Together, they provide holistic insights into model limitations. MM-RLHF-SafetyBench [24] samples from existing datasets, further covering areas such as adversarial attacks, privacy, red team attacks, and harmful content detection.

5.4 Conversation

These benchmarks prioritize evaluating foundational visual skills, such as low-level perception ability (Q-Bench [96], LLVisionQA [96]), description ability on low-level information (LLDescribe [96]), 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 [97]), cross-domain tasks (LiveBench's [98] math/news integration), and adversarial prompts (Vibe-Eval's [99] high-difficulty questions). They reveal model adaptability beyond standard datasets.

5.5 Reward Model

Each benchmark targets specific evaluation dimensions, such as multilingual capabilities (23 languages in M-RewardBench [100]), alignment/safety/bias (MJ-Bench [103]), leveraging human annotations to enhance explainability and the final model scoring capability (MM-RLHF-RewardBench [24]), and ability of MLLMs in assisting judges across diverse modalities (MLLM-as-a-Judge's [104] scoring vs. pairwise comparisons). These frameworks reveal model strengths and weaknesses in structured and OOD scenarios. High-quality datasets are curated through human-AI collaboration (VL-RewardBench's [101] annotation pipeline) or structured triplet designs (RewardBench [102]). Tasks range from simple preference ranking to complex reasoning, pushing models to handle nuanced challenges like hallucination and ethical alignment.

5.6 Alignment

Some benchmarks investigate the ability of models to align with human preference. Arena-Hard [105] is a comprehensive, multi-dimensional benchmark designed to evaluate the alignment of Chinese LLMs. AlpacaEval-V2 [106] proposes a simple regression analysis method to control for length bias in self-assessment. Arena-Hard [105] increases the separation of model performance by three times and achieves a 98.6% correlation with human preference rankings. MM-AlignBench[33] is a manually annotated benchmark specifically designed to assess the alignment with human values.

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. The effectiveness of reward models in these frameworks, particularly their performance in guiding alignment, warrants further investigation. Additionally, benchmarks for alignment with human preference have also evolved from the LLM domain to the MLLM domain.

6 Future Work and Open Challenges

As MLLMs rapidly advance, aligning them with human preferences has become a central focus. However, several challenges persist. First, there is a scarcity of high-quality and diverse datasets. Second, many methods fail to effectively utilize visual information, often relying primarily on text for constructing positive and negative samples, and neglecting the full potential of multimodal data. Additionally, there is a lack of comprehensive evaluation standards, with current methods often being validated only on specific types of benchmarks, such as hallucination or dialogue tasks, which makes it difficult to assess their generalizability. Furthermore, by drawing on advancements in LLM post-training strategies and agent research, we can pinpoint limitations in existing MLLM alignment approaches. Overcoming these challenges is essential for developing more robust and comprehensive alignment methods.

6.1 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 optical character recognition, mathematical problems, and chart understanding, among others. Constructing a comprehensive dataset that addresses this wide array of tasks is an extremely challenging endeavor. To the best of our knowledge, there is currently no publicly available, fully human-annotated multimodal dataset that exceeds 200,000 samples. These limitations in data quality and coverage pose significant barriers to effectively aligning MLLMs.

6.2 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 different 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.

Comprehensive Evaluation

Most MLLM alignment studies primarily evaluate their algorithms on a few key areas, such as hallucination, conversational abilities, or 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[130] 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 reasoning LLMs represented by OpenAI (O1) [131] and DeepSeek-R1 [120] have demonstrated that RL algorithms and preference data are crucial for improving LLMs in complex problem-solving, long-context understanding, and generation tasks. Here we will explore the insights gained from LLM reasoning enhancement research and their implications for aligning MLLMs from two dimensions: data and optimization framework.

(1) *Data. Scale & quality.* Corresponding approaches have gradually evolved from the small-model resampling (e.g., OpenMathInstruct[132]) to high-quality synthetic data [124] (e.g., AceMath[133]), progressed to using cutting-edge models (e.g., OpenAI (O1)[131]), and synthesized data via domain-specialized models for scalable knowledge transfer (e.g., DeepSeek-V3 [134]). This demonstrates a clear iterative path in data construction strategies. Currently, datasets used for reasoning enhancement are generally at the million-sample scale (e.g., Qwen-2.5-MATH [135]). *Efficiency.* Adoption of "less is more" alignment (e.g., LIMA's[136] 1k curated samples for 65B Llama [137], S1-32B's[138] elite 1k high-diversity data comparable with O1-preview), proving minimal high-quality data optimally activates pretrained capabilities while reducing dependency on data scale.

(2) *Optimization Framework. Sampling Strategies.* Recent advancements have seen a shift toward online reinforcement learning (RL), with approaches such as DeepSeek-V3 [134] and Qwen-2.5-MATH's [135] online sampling methods, which effectively mitigate distributional shifts. Additionally, Mini-Max employs an offline+online sampling strategy to enhance model performance. *Training Paradigms.* Multi-stage, collaborative optimization has become the dominant approach. For example, Llama 3 incorporates a six-round DPO iteration, while DeepSeek optimizes reasoning depth (long-CoT) and conciseness through temperature-varied sampling and reflection/verification prompts. *Algorithms.* RL algorithms have evolved from early policy gradient methods to more complex PPO. Recent improvements based on PPO follow two main directions: one is to remove the critic model and train the policy using sparse rewards, thereby reducing the parameter count by half, as seen in DPO and GRPO; another is to refine the design of the critic, such as PRIME [139], which introduces a ratio as the advantage function, and OREAL[140], which reshapes the rewards of positive and negative examples.

By prioritizing high-quality data and innovative optimization frameworks, the field is moving towards more effective and scalable models that can also better unlock the reasoning potential of MLLMs.

Insight from LLM Alignment

Aligning LLMs has been a critical focus in recent research, offering valuable insights that can inform the development of MLLMs. By examining lessons learned from existing LLM alignment strategies, we can uncover key principles that may enhance MLLM community.

(1) *Improving Training Efficiency.* Current MLLM alignment methods rely on the DPO loss function. However, since DPO requires loading both the policy model and reference model simultaneously, the training speed is significantly reduced. Could reference-free approaches like SimPO [141] be leveraged to further enhance training efficiency? This approach might accelerate the training process while reducing dependence on reference models. Further investigation into the specific role and impact of reference models in MLLM alignment is critical, as it could inform both efficiency improvements and optimized model design.

(2) *Mitigating Overoptimization/Reward Hacking.* In LLM alignment using DPO [9] or RLHF, overoptimization remains a key challenge [142], [143], where performance measured by the learned proxy reward model improves while true quality stagnates or deteriorates. Specifically, when training data exhibits significant quality disparities (e.g., excessive bias toward specific task types), models may overfit these task-specific patterns and underperform in real-world scenarios. Mitigation strategies include: using balanced training datasets to ensure diversity and representativeness,

preventing narrow optimization; mplementing early stopping when validation performance plateaus; incorporating regularization techniques to reduce overreliance on training data and enhance generalization.

6.7 MLLM as Agents

LMs combine the powerful reasoning abilities of LLMs with the capability to understand and process data from multiple modalities, including images, texts, and audio. enables them to extract knowledge from various types of information sources and perform integrated analysis, making them highly advantageous in handling complex real-world tasks [144], [145], [146], [147], [148]. , the multimodal understanding capability of MLLMs provides a solid foundation for their application in complex tasks. example, in the field of autonomous driving, MLLMs can process data from various modalities such as camera images, vehicle sensors, and traffic signals, allowing them to accurately perceive the surrounding environment [149], [150]. nd, MLLMs inherit the powerful reasoning abilities of LLMs [120], enabling them to make precise decisions after perceiving the environment. n faced with complex tasks, MLLMs can break down the task and develop detailed execution plans to solve the problem efficiently. nstance, in industrial robotics, MLLMs can plan precise robot action sequences based on product assembly requirements, effectively directing the robot to complete various operations [151], [152]. thermore, thanks to their powerful multimodal perception capabilities and the generalization abilities of LLMs across different domains, MLLMs exhibit excellent scalability and versatility, allowing them to be easily transferred to other tasks without the need for task-specific parameter adjustments.

vever, there are still several pending issues that need to be addressed in order to transform MLLMs into highly effective agents. *ulti-agent collaboration*. rent frameworks for multi-agent collaboration primarily focus on text-based agents, which have been shown to significantly expand the cognitive boundaries of individual agents. v-ever, MLLM-based multi-agent systems still lack mature solutions, evident in the absence of tailored multimodal communication and information-sharing mechanisms [153], as well as customized multimodal memory mechanisms. *Robustness*. robustness of MLLM-based agents in open environments has not been systematically validated. example, [154] demonstrated that adding adversarial perturbations to images in a web navigation environment can significantly hijack agents, forcing them to execute targeted adversarial goals. sequent research should introduce adversarial robustness testing and safeguarding techniques for MLLM agents. *Security*. introduction of more complex components in MLLM agents increases the diversity of security risks. se risks span from multimodal environmental perception to reasoning and memory, with the potential for malicious attacks [155], leading to privacy breaches and hijacking. ure work should comprehensively explore diverse security protection mechanisms for MLLM agents to mitigate these risks.

7 Conclusion

field of MLLM alignment is developing rapidly. his paper, we conduct a systematic and comprehensive survey of existing research on MLLM alignment, focusing on four crucial questions: t application scenarios can be covered, how to construct datasets, how to evaluate algorithms, and where the direction of the next alignment algorithm lies. ar as we know, this paper is the first systematic survey dedicated to MLLM alignment. hope that this survey will facilitate further research in this area.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, 2020. 1

[2] C. Fu, Y.-F. Zhang, S. Yin, B. Li, X. Fang, S. Zhao, H. Duan, X. Sun, Z. Liu, L. Wang *et al.*, "Mme-survey: A comprehensive survey on evaluation of multimodal llms," *arXiv*, 2024. 1, 2

[3] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv*, 2024. 1, 2

[4] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini *et al.*, "Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models," *arXiv*, 2024. 1, 2

[5] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *arXiv*, 2024. 1, 2

[6] W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, and W. Ping, "Nvlm: Open frontier-class multimodal llms," *arXiv*, 2024. 1, 2

[7] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet *et al.*, "Pixtral 12b," *arXiv*, 2024. 1, 2

[8] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, Y. Li, Z. Long, H. Gao, K. Li *et al.*, "Vita-1.5: Towards gpt-4o level real-time vision and speech interaction," *arXiv*, 2025. 1, 2

[9] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *NeurIPS*, 2023. 1, 11

[10] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K. Y. Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.-C. Zhu, Y. Guo, and W. Gao, "Ai alignment: A comprehensive survey," *arXiv*, 2024. 2

[11] K. Kumar, T. Ashraf, O. Thawakar, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, P. H. S. Torr, S. Khan, and F. S. Khan, "Llm post-training: A deep dive into reasoning large language models," *arXiv*, 2025. 2

[12] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell, "Aligning large multimodal models with factually augmented rlhf," *ACL*, 2023. 3, 4, 7

[13] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, and T.-S. Chua, "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," *CVPR*, 2024. 3, 4, 5, 7

[14] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," *AAAI*, 2024. 3, 4, 8, 9

[15] Z. Zhao, B. Wang, L. Ouyang, X. Dong, J. Wang, and C. He, "Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization," *arXiv*, 2024. 3, 4, 8

[16] F. Wang, W. Zhou, J. Y. Huang, N. Xu, S. Zhang, H. Poon, and M. Chen, "mdpo: Conditional preference optimization for multimodal large language models," *EMNLP*, 2024. 3, 4

[17] T. Yu, H. Zhang, Q. Li, Q. Xu, Y. Yao, D. Chen, X. Lu, G. Cui, Y. Dang, T. He, X. Feng, J. Song, B. Zheng, Z. Liu, T.-S. Chua, and M. Sun, "Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness," *CVPR*, 2024. 3, 4, 9

[18] L. Xue, M. Shu, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S. Ryoo, S. Kendre, J. Zhang, C. Qin, S. Zhang, C.-C. Chen, N. Yu, J. Tan, T. M. Awalgaonkar, S. Heinecke, H. Wang, Y. Choi, L. Schmidt, Z. Chen, S. Savarese, J. C. Niebles, C. Xiong, and R. Xu, "xgen-mm (blip-3): A family of open large multimodal models," *arXiv*, 2024. 3, 4

[19] J. Fu, S. Huangfu, H. Fei, X. Shen, B. Hooi, X. Qiu, and S.-K. Ng, "Chip: Cross-modal hierarchical direct preference optimization for multimodal llms," *ICLR*, 2025. 3, 4, 5

[20] Y. Fu, R. Xie, X. Sun, Z. Kang, and X. Li, "Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization," *arXiv*, 2024. 3, 4, 5

[21] J. Lu, J. Wu, J. Li, X. Jia, S. Wang, Y. Zhang, J. Fang, X. Wang, and X. He, "Dama: Data- and model-aware alignment of multi-modal llms," *arXiv*, 2025. 2, 3, 4

[22] Z. Yang, X. Luo, D. Han, Y. Xu, and D. Li, "Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key," *CVPR*, 2025. 3, 4

[23] S. Xing, Y. Wang, P. Li, R. Bai, Y. Wang, C. Qian, H. Yao, and Z. Tu, "Re-align: Aligning vision language models via retrieval-augmented direct preference optimization," *arXiv*, 2025. 3, 4

[24] Y.-F. Zhang, T. Yu, H. Tian, C. Fu, P. Li, J. Zeng, W. Xie, Y. Shi, H. Zhang, J. Wu, X. Wang, Y. Hu, B. Wen, F. Yang, Z. Zhang, T. Gao, D. Zhang, L. Wang, R. Jin, and T. Tan, "Mm-rlhf: The next step forward in multimodal llm alignment," *arXiv*, 2025. 2, 3, 5, 7, 8, 10

[25] L. Li, Z. Xie, M. Li, S. Chen, P. Wang, L. Chen, Y. Yang, B. Wang, and L. Kong, "Silkie: Preference distillation for large visual language models," *EMNLP*, 2023. 3, 4, 8

[26] Y. Ouali, A. Bulat, B. Martinez, and G. Tzimiropoulos, "Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms," *arXiv*, 2024. 3, 4, 8

[27] X. Wang, J. Chen, Z. Wang, Y. Zhou, Y. Zhou, H. Yao, T. Zhou, T. Goldstein, P. Bhatia, F. Huang, and C. Xiao, "Enhancing visual-language modality alignment in large vision language models via self-improvement," *NAACL*, 2024. 3, 4, 9

[28] T. Xiong, X. Wang, D. Guo, Q. Ye, H. Fan, Q. Gu, H. Huang, and C. Li, "Llava-critic: Learning to evaluate multimodal models," *CVPR*, 2024. 2, 3, 5

[29] W. Wang, Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, J. Zhu, X. Zhu, L. Lu, Y. Qiao, and J. Dai, "Enhancing the reasoning ability of multimodal large language models via mixed preference optimization," *arXiv*, 2024. 2, 3, 5, 9

[30] T. Luo, A. Cao, G. Lee, J. Johnson, and H. Lee, "vVLM: Exploring visual reasoning in VLMs against language priors," *openreview*, 2024. 3, 5, 9

[31] R. Wijaya, N.-B. Nguyen, and N.-M. Cheung, "Multimodal preference data synthetic alignment with reward model," *arXiv*, 2024. 3, 5

[32] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv*, 2025. 3, 5

[33] X. Zhao, S. Ding, Z. Zhang, H. Huang, M. Cao, W. Wang, J. Wang, X. Fang, W. Wang, G. Zhai, H. Duan, H. Yang, and K. Chen, "Omnialign-v: Towards enhanced alignment of mllms with human preference," *arXiv*, 2025. 3, 5, 8, 10

[34] Y. Peng, G. Zhang, X. Geng, and X. Yang, "Lmm-r1," https://github.com/TideDra/lmm-r1, MSRA, 2025. 3, 5

[35] X. Wang and P. Peng, "Open-r1-video," https://github.com/Wang-Xiaodong1899/Open-R1-Video, Peking University, 2025. 3, 5

[36] H. Shen, Z. Zhang, Q. Zhang, R. Xu, and T. Zhao, "Vlm-r1: A stable and generalizable r1-style large vision-language model," https://github.com/om-ai-lab/VLM-R1, Zhejiang University, 2025. 3, 5

[37] Y. Zheng, J. Lu, S. Wang, Z. Feng, D. Kuang, and Y. Xiong, "Easyr1: An efficient, scalable, multi-modality rl training framework," https://github.com/hiyouga/EasyR1, Beihang University, 2025. 3, 5

[38] L. Chen, L. Li, H. Zhao, Y. Song, and Vinci, "R1-v: Reinforcing super generalization ability in vision-language models with less than $3," https://github.com/Deep-Agent/R1-V, Peking University, 2025. 2, 3, 5

[39] F. Meng, L. Du, Z. Liu, Z. Zhou, Q. Lu, D. Fu, B. Shi, W. Wang, J. He, K. Zhang, P. Luo, Y. Qiao, Q. Zhang, and W. Shao, "Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning," https://github.com/ModalMinds/MM-EUREKA, Shanghai AI Laboratory, 2025. 3, 5

[40] H. Zhou, X. Li, R. Wang, M. Cheng, T. Zhou, and C.-J. Hsieh, "R1-zero's "aha moment" in visual reasoning on a 2b non-sft model," *arXiv*, 2025. 3, 5

[41] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, Y. Hu, and S. Lin, "Vision-r1: Incentivizing reasoning capability in multimodal large language models," *arXiv*, 2025. 3, 5

[42] J. Zhao, X. Wei, and L. Bo, "R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning," *arXiv*, 2025. 3, 5

[43] S. Leng, J. Wang, J. Li, H. Zhang, Z. Hu, B. Zhang, H. Zhang, Y. Jiang, X. Li, D. Zhao, F. Wang, Y. Rong, A. Sun, and S. Lu, "Mmr1: Advancing the frontiers of multimodal reasoning," https://github.com/LengSicong/MMR1, Nanyang Technological University, 2025. 3, 5

[44] Z. Liu, Y. Zang, X. Dong, P. Zhang, Y. Cao, H. Duan, C. He, Y. Xiong, D. Lin, and J. Wang, "Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models," *ICLR*, 2024. 3, 5, 9

[45] H. Jia, C. Jiang, H. Xu, W. Ye, M. Dong, M. Yan, J. Zhang, F. Huang, and S. Zhang, "Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization," *arXiv*, 2024. 3, 5, 9

[46] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv*, 2024. 3, 5

[47] R. Liu, C. Li, H. Tang, Y. Ge, Y. Shan, H. Lu, and J. Yang, "PPLLaVA: Varied video sequence understanding with prompt guidance," *openreview*, 2024. 3, 6

[48] C. Tang, Y. Li, Y. Yang, J. Zhuang, G. Sun, W. Li, Z. MA, and C. Zhang, "Enhancing multimodal LLM for detailed and accurate video captioning using multi-round preference optimization," *openreview*, 2024. 3, 6, 8

[49] S. Eom, J. Shim, E. Yoon, H. S. Yoon, H. Ko, M. A. Hasegawa-Johnson, and C. D. Yoo, "SQuba: Speech mamba language model with querying-attention for efficient summarization," *openreview*, 2024. 3, 6, 9

[50] H. Chen, W. Zhao, Y. Li, W. Li, Z. Li, N. Zhu, T. Zhong, Y. Wang, Y. Shang, L. Guo, J. Han, T. Liu, J. Liu, and T. Zhang, "3d-CT-GPT++: Enhancing 3d radiology report generation with direct preference optimization and large vision-language models," *openreview*, 2024. 3, 7

[51] R. Zhang, X. Wei, D. Jiang, Z. Guo, S. Li, Y. Zhang, C. Tong, J. Liu, A. Zhou, B. Wei, S. Zhang, P. Gao, C. Li, and H. Li, "Mavis: Mathematical visual instruction tuning with an automatic data engine," *arXiv*, 2024. 3, 7, 8

[52] K. Jiao, Z. Fang, J. Liu, B. Li, Z. Qiao, Y. Xu, Y. Zhu, X. Liu, J. Wang, and X. Li, "InteractiveCOT: Aligning dynamic chain-of-thought planning for embodied decision-making," *openreview*, 2024. 3, 7, 8

[53] D. Li, T. Cai, T. Tang, W. Chai, K. R. Driggs-Campbell, H. Wang, and G. Wang, "Homiebot: an adaptive system for embodied mobile manipulation in open environments," *openreview*, 2024. 3, 7, 8

[54] C. Liu, G. Tianyi, Y. Liu, and L. Xu, "AdPO: Enhancing the adversarial robustness of large vision-language models with preference optimization," *openreview*, 2024. 3, 7, 9

[55] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales, "Safety fine-tuning at (almost) no cost: A baseline for vision large language models," *ICML*, 2024. 3, 7, 9

[56] A. Afzali, B. khodabandeh, A. Rasekh, M. JafariNodeh, S. K. Ranjbar, and S. Gottschalk, "Aligning visual contrastive learning models via preference optimization," in *ICLR*, 2025. 3, 7

[57] Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y. LeCun, Y. Ma, and S. Levine, "Fine-tuning large vision-language models as decision-making agents via reinforcement learning," *NeurIPS*, 2024. 3, 7

[58] L. Team, "The llama 3 herd of models," *arXiv*, 2024. 3, 8

[59] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," *ICLR*, 2024. 3, 8

[60] B.-K. Lee, S. Chung, C. W. Kim, B. Park, and Y. M. Ro, "Phantom of latent for large language and vision models," *arXiv*, 2024. 3, 8

[61] Y.-F. Zhang, H. Zhang, H. Tian, C. Fu, S. Zhang, J. Wu, F. Li, K. Wang, Q. Wen, Z. Zhang, L. Wang, R. Jin, and T. Tan, "Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?" *ICLR*, 2025. 3, 9

[62] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, and F. Zhao, "Are we on the right way for evaluating large vision-language models?" *arXiv*, 2024. 3, 9

[63] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" *ECCV*, 2023. 3, 9

[64] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, J. Lei, Q. Lu, R. Chen, P. Xu, R. Zhang, H. Zhang, P. Gao, Y. Wang, Y. Qiao, P. Luo, K. Zhang, and W. Shao, "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," *ICML*, 2024. 3, 9

[65] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna, "Blink: Multimodal large language models can see but not perceive," *ECCV*, 2024. 3, 9

[66] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," *ICLR*, 2023. 3, 9

[67] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang, "Sqa3d: Situated question answering in 3d scenes," in *ICLR*, 2023. 3, 9

[68] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *CVPR*, 2024. 3, 9

[69] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024. 3, 9

[70] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, "Mantis: Interleaved multi-image instruction tuning," *TMLR*, 2024. 3, 9

[71] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," *EMNLP*, 2019. 3, 9

[72] Y. Wang, Y. Wang, D. Zhao, C. Xie, and Z. Zheng, "Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models," *arXiv*, 2024. 3, 9

[73] H. Qiu, W. Hu, Z.-Y. Dou, and N. Peng, "Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models," *ACL*, 2024. 3, 9

[74] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *EMNLP*, 2023. 3, 9

[75] J. Wang, Y. Zhou, G. Xu, P. Shi, C. Zhao, H. Xu, Q. Ye, M. Yan, J. Zhang, J. Zhu *et al.*, "Evaluation and analysis of hallucination in large vision-language models," *arXiv*, 2023. 3, 9

[76] A. Ben-Kish, M. Yanuka, M. Alper, R. Giryes, and H. Averbuch-Elor, "Mocha: Multi-objective reinforcement mitigating caption hallucinations," *EMNLP*, 2023. 3, 9

[77] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," in *ICLR*, 2023. 3, 9

[78] J. Wang, Y. Wang, G. Xu, J. Zhang, Y. Gu, H. Jia, M. Yan, J. Zhang, and J. Sang, "An llm-free multi-dimensional benchmark for mllms hallucination evaluation," *arXiv*, 2023. 3, 9

[79] X. Wang, Y. Zhou, X. Liu, H. Lu, Y. Xu, F. He, J. Yoon, T. Lu, G. Bertasius, M. Bansal *et al.*, "Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences," *ACL*, 2024. 3, 9

[80] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell, "Aligning large multimodal models with factually augmented rlhf," *ACL*, 2023. 3, 9

[81] K.-i. Lee, M. Kim, S. Yoon, M. Kim, D. Lee, H. Koh, and K. Jung, "Vlind-bench: Measuring language priors in large vision-language models," *NAACL*, 2024. 3, 9

[82] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob *et al.*, "Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *CVPR*, 2024. 3, 9

[83] W. Huang, H. Liu, M. Guo, and N. Z. Gong, "Visual hallucinations of multi-modal large language models," *ACL*, 2024. 3, 9

[84] X. Chen, C. Wang, Y. Xue, N. Zhang, X. Yang, Q. Li, Y. Shen, J. Gu, and H. Chen, "Unified hallucination detection for multimodal large language models," *ACL*, 2024. 3, 9

[85] J. Liu, Y. Fu, R. Xie, R. Xie, X. Sun, F. Lian, Z. Kang, and X. Li, "Phd: A prompted visual hallucination evaluation dataset," *CVPR*, 2024. 3, 9

[86] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *AAAI*, 2019. 3, 9

[87] M. Wu, J. Ji, O. Huang, J. Li, Y. Wu, X. Sun, and R. Ji, "Evaluating and analyzing relationship hallucinations in lvlms," *ICML*, 2024. 3, 9

[88] N. Raihan, A. Anastasopoulos, and M. Zampieri, "mhumaneval – a multilingual benchmark to evaluate large language models for code generation," *arXiv*, 2025. 3, 9

[89] C. Cui, Y. Zhou, X. Yang, S. Wu, L. Zhang, J. Zou, and H. Yao, "Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges," *arXiv*, 2023. 3, 9

[90] B. Yan, J. Zhang, Z. Yuan, S. Shan, and X. Chen, "Evaluating the quality of hallucination benchmarks for large vision-language models," *arXiv*, 2024. 3, 9

[91] Q. Guo, S. Pang, X. Jia, and Q. Guo, "Efficiently adversarial examples generation for visual-language models under targeted transfer scenarios using diffusion models," *arXiv*, 2024. 3, 9

[92] M. Li, L. Li, Y. Yin, M. Ahmed, Z. Liu, and Q. Liu, "Red teaming visual language models," *ACL*, 2024. 3, 9

[93] Y. Zhang, Y. Huang, Y. Sun, C. Liu, Z. Zhao, Z. Fang, Y. Wang, H. Chen, X. Yang, X. Wei, H. Su, Y. Dong, and J. Zhu, "Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models," *NeurIPS*, 2024. 3, 9

[94] H. Tu, C. Cui, Z. Wang, Y. Zhou, B. Zhao, J. Han, W. Zhou, H. Yao, and C. Xie, "How many unicorns are in this image? a safety evaluation benchmark for vision llms," *arXiv*, 2023. 3, 10

[95] X. Li, H. Zhou, R. Wang, T. Zhou, M. Cheng, and C.-J. Hsieh, "Mossbench: Is your multimodal language model oversensitive to safe queries?" *arXiv*, 2024. 3, 10

[96] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai *et al.*, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," *ICLR*, 2023. 3, 10

[97] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv*, 2024. 3, 10

[98] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum, "Livebench: A challenging, contamination-free llm benchmark," *ICLR*, 2024. 3, 10

[99] P. Padlewski, M. Bain, M. Henderson, Z. Zhu, N. Relan, H. Pham, D. Ong, K. Aleksiev, A. Ormazabal, S. Phua, E. Yeo, E. Lamprecht, Q. Liu, Y. Wang, E. Chen, D. Fu, L. Li, C. Zheng, C. de Masson d'Autume, D. Yogatama, M. Artetxe, and Y. Tay, "Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models," *arXiv*, 2024. 3, 10

[100] S. Gureja, L. J. V. Miranda, S. B. Islam, R. Maheshwary, D. Sharma, G. Winata, N. Lambert, S. Ruder, S. Hooker, and M. Fadaee, "M-rewardbench: Evaluating reward models in multilingual settings," *arXiv*, 2024. 3, 10

[101] L. Li, Y. Wei, Z. Xie, X. Yang, Y. Song, P. Wang, C. An, T. Liu, S. Li, B. Y. Lin, L. Kong, and Q. Liu, "Vlrewardbench: A challenging benchmark for vision-language generative reward models," *arXiv*, 2024. 3, 10

[102] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi, "Rewardbench: Evaluating reward models for language modeling," *arXiv*, 2024. 3, 10

[103] Z. Chen, Y. Du, Z. Wen, Y. Zhou, C. Cui, Z. Weng, H. Tu, C. Wang, Z. Tong, Q. Huang, C. Chen, Q. Ye, Z. Zhu, Y. Zhang, J. Zhou, Z. Zhao, R. Rafailov, C. Finn, and H. Yao, "Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation?" *arXiv*, 2024. 3, 10

[104] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun, "Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark," *ICML*, 2024. 3, 10

[105] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica, "From crowdsourced data to high-

quality benchmarks: Arena-hard and benchbuilder pipeline," *CoRR*, 2024. 3, 10
[106] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-controlled alpacaeval: A simple way to debias automatic evaluators," *COLM*, 2024. 3, 10
[107] X. Liu, X. Lei, S. Wang, Y. Huang, Z. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W. L. Tam, X. Zhang, L. Sun, X. Gu, H. Wang, J. Zhang, M. Huang, Y. Dong, and J. Tang, "Alignbench: Benchmarking chinese alignment of large language models," *ACL*, 2024. 3
[108] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv*, 2023. 2
[109] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," *arXiv*, 2023. 2
[110] C. Fu, H. Lin, Z. Long, Y. Shen, M. Zhao, Y. Zhang, S. Dong, X. Wang, D. Yin, L. Ma *et al.*, "Vita: Towards open-source interactive omni multimodal llm," *arXiv*, 2024. 2
[111] T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma, "Sft memorizes, rl generalizes: A comparative study of foundation model post-training," *arXiv*, 2025. 2
[112] Y.-F. Zhang, W. Yu, Q. Wen, X. Wang, Z. Zhang, L. Wang, R. Jin, and T. Tan, "Debiasing multimodal large language models," *arXiv*, 2024. 2
[113] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales, "Safety fine-tuning at (almost) no cost: A baseline for vision large language models," *ICML*, 2024. 2
[114] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *NeurIPS*, 2023. 4
[115] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *ICLR*, 2021. 4
[116] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *PMLR*, 2021. 4, 8
[117] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li, "Llava-onevision: Easy visual task transfer," *arXiv*, 2024. 5
[118] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CVPR*, 2022. 5
[119] H. Wang, W. Xiong, T. Xie, H. Zhao, and T. Zhang, "Interpretable preferences via multi-objective reward modeling and mixture-of-experts," *EMNLP*, 2024. 5
[120] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv*, 2025. 5, 11, 12
[121] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker, "Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms," *arXiv*, 2024. 5
[122] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv*, 2024. 5
[123] J. Hu, X. Wu, Z. Zhu, Xianyu, W. Wang, D. Zhang, and Y. Cao, "Openrlhf: An easy-to-use, scalable and high-performance rlhf framework," *arXiv*, 2024. 5
[124] OpenAI., "Gpt-4 technical report," 2023. 7, 11
[125] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv*, 2017. 7
[126] G. Team, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv*, 2024. 8
[127] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *ICLR*, 2021. 8
[128] Y. Zhou, C. Cui, R. Rafailov, C. Finn, and H. Yao, "Aligning modalities in vision large language models via preference fine-tuning," *arXiv*, 2024. 9
[129] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *ICLR*, 2017. 9
[130] J. Ji, J. Zhou, H. Lou, B. Chen, D. Hong, X. Wang, W. Chen, K. Wang, R. Pan, J. Li, M. Wang, J. Dai, T. Qiu, H. Xu, D. Li, W. Chen, J. Song, B. Zheng, and Y. Yang, "[redacted] anything: Training all-modality models to follow instructions with language feedback," *arXiv*, 2024. 11
[131] OpenAI., "Introducing openai o1-preview," 2024. [Online]. Available: https://openai.com/index/introducing-openai-o1-preview/ 11
[132] S. Toshniwal, I. Moshkov, S. Narenthiran, D. Gitman, F. Jia, and I. Gitman, "Openmathinstruct-1: A 1.8 million math instruction tuning dataset," *NeurIPS*, 2024. 11
[133] Z. Liu, Y. Chen, M. Shoeybi, B. Catanzaro, and W. Ping, "Acemath: Advancing frontier math reasoning with post-training and reward modeling," *arXiv*, 2025. 11
[134] DeepSeek-AI, "Deepseek-v3 technical report," *arXiv*, 2024. 11
[135] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang, "Qwen2.5-math technical report: Toward mathematical expert model via self-improvement," *arXiv*, 2024. 11
[136] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "Lima: Less is more for alignment," *NeurIPS*, 2023. 11
[137] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv*, 2023. 11
[138] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto, "s1: Simple test-time scaling," *arXiv*, 2025. 11
[139] G. Cui, L. Yuan, Z. Wang, H. Wang, W. Li, B. He, Y. Fan, T. Yu, Q. Xu, W. Chen, J. Yuan, H. Chen, K. Zhang, X. Lv, S. Wang, Y. Yao, X. Han, H. Peng, Y. Cheng, Z. Liu, M. Sun, B. Zhou, and N. Ding, "Process reinforcement through implicit rewards," *arXiv*, 2025. 11
[140] C. Lyu, S. Gao, Y. Gu, W. Zhang, J. Gao, K. Liu, Z. Wang, S. Li, Q. Zhao, H. Huang, W. Cao, J. Liu, H. Liu, J. Liu, S. Zhang, D. Lin, and K. Chen, "Exploring the limit of outcome reward for learning mathematical reasoning," *arXiv*, 2025. 11
[141] Y. Meng, M. Xia, and D. Chen, "Simpo: Simple preference optimization with a reference-free reward," *NeurIPS*, 2024. 11
[142] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *ICML*, 2023. 11
[143] R. Rafailov, Y. Chittepu, R. Park, H. Sikchi, J. Hejna, W. B. Knox, C. Finn, and S. Niekum, "Scaling laws for reward model overoptimization in direct alignment algorithms," *NeurIPS*, 2024. 11
[144] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui, "The rise and potential of large language model based agents: A survey," *arXiv*, 2023. 12
[145] J. Wang, H. Xu, J. Ye, M. Yan, W. Shen, J. Zhang, F. Huang, and J. Sang, "Mobile-agent: Autonomous multi-modal mobile device agent with visual perception," *arXiv*, 2024. 12
[146] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv*, 2024. 12
[147] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi, K. Ikeuchi, H. Vo, L. Fei-Fei, and J. Gao, "Agent ai: Surveying the horizons of multimodal interaction," *arXiv*, 2024. 12
[148] F. Ma, Y. Zhou, Y. Zhang, S. Wu, Z. Zhang, Z. He, F. Rao, and X. Sun, "Task navigator: Decomposing complex tasks for multimodal large language models," in *CVPR*, 2024. 12
[149] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, H. Tian, L. Lu, X. Zhu, X. Wang, Y. Qiao, and J. Dai, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv*, 2023. 12
[150] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, T. Gao, E. Li, K. Tang, Z. Cao, T. Zhou, A. Liu, X. Yan, S. Mei, J. Cao, Z. Wang, and C. Zheng, "A survey on multimodal large language models for autonomous driving," *arXiv*, 2023. 12
[151] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large

language model for object-centric robotic manipulation," *CVPR*, 2023. 12

[152] J. Liu, C. Li, G. Wang, L. Lee, K. Zhou, S. Chen, C. Xiong, J. Ge, R. Zhang, and S. Zhang, "Self-corrected multimodal large language model for end-to-end robot manipulation," *arXiv*, 2024. 12

[153] T. Ossowski, J. Chen, D. Maqbool, Z. Cai, T. Bradshaw, and J. Hu, "Comma: A communicative multimodal multi-agent benchmark," *arXiv*, 2025. 12

[154] C. H. Wu, R. Shah, J. Y. Koh, R. Salakhutdinov, D. Fried, and A. Raghunathan, "Dissecting adversarial robustness of multimodal lm agents," *ICLR*, 2025. 12

[155] Y. Yang, X. Yang, S. Li, C. Lin, Z. Zhao, C. Shen, and T. Zhang, "Security matrix for multimodal agents on mobile devices: A systematic and proof of concept study," *arXiv*, 2024. 12