Self-Improvement in Multimodal Large Language Models: A Survey

Shijian Deng[1] **Kai Wang**[2] **Tianyu Yang**[3] **Harsh Singh**[4] **Yapeng Tian**[1]
[1]The University of Texas at Dallas [2]University of Toronto
[3]University of Notre Dame [4]Mohamed bin Zayed University of Artificial Intelligence
{shijian.deng,yapeng.tian}@utdallas.edu kaikai.wang@mail.utoronto.ca
tyang4@nd.edu harsh.singh@mbzuai.ac.ae

Abstract

Recent advancements in self-improvement for Large Language Models (LLMs) have efficiently enhanced model capabilities without significantly increasing costs, particularly in terms of human effort. While this area is still relatively young, its extension to the multimodal domain holds immense potential for leveraging diverse data sources and developing more general self-improving models. This survey is the first to provide a comprehensive overview of self-improvement in Multimodal LLMs (MLLMs). We provide a structured overview of the current literature and discuss methods from three perspectives: 1) data collection, 2) data organization, and 3) model optimization, to facilitate the further development of self-improvement in MLLMs. We also include commonly used evaluations and downstream applications. Finally, we conclude by outlining open challenges and future research directions.

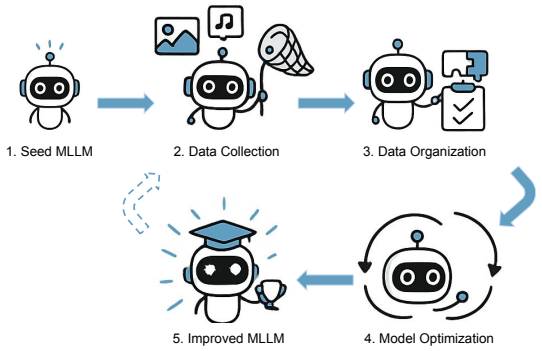

Figure 1: An illustration of self-improvement in Multimodal Large Language Models. The process involves selecting a seed MLLM to generate new data, organizing that data into a dataset, and finally obtaining an improved model through training. This process can be iterative to achieve recursive self-improvement.

1 Introduction

Self-improvement aims to enable models to collect and organize data required to build a better generation of themselves, which offers a path to overcome the costly scaling issues and potential performance ceilings of static training paradigms. In Multi-Modal Large Language Models (MLLMs), self-improvement seeks to use MLLMs themselves to obtain their own training data, resulting in improved MLLMs. Recent research (Favero et al., 2024; Deng et al., 2024b; Amirloo et al., 2024) show that this approach can significantly reduce hallucinations and improve performance on general tasks with relatively low cost. Significant progress has been made in this direction. Some current studies (Zhou et al., 2024a) partially leverage self-improvement by combining it with external tools or peer models, while others (Yu et al., 2024b) explore approaches that rely solely on a single model to handle all processes, toward full self-improvement. Although previous work (Tao et al., 2024) has summarized the self-improvement in text-only LLMs and other surveys study the general scope of MLLMs (Yin et al., 2024; Zhang et al., 2024a) or specific issues such as hallucinations (Bai et al., 2024), there is no comprehensive survey that focuses on these self-improvement methods for MLLMs. To fill this gap, we dedicate this paper to providing a comprehensive review of this area and identifying the challenges that need to be addressed.

Compared to self-improvement in LLMs (Huang et al., 2022; Tao et al., 2024), self-improvement in MLLMs faces unique challenges, such as the inclusion of multiple modalities. This can introduce modality alignment problems, which are known to cause issues like hallucination in MLLMs (Li et al., 2023b). Additionally, MLLMs often cannot generate all the training data it needed by themselves, as most current models (Liu et al., 2024a; Bai et al., 2023) are unable to generate images directly.

Despite these challenges, there is growing in-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1987–2006

terest in leveraging self-improvement in MLLMs to build models more effectively and efficiently. Promising results have already been achieved in this area. This paper aims to summarize previous works, compare methods, and provide clearer guidance for future research directions in this field.

In this survey, we follow the structure outlined below: First, we provide an overview of the field. Next, we introduce the most commonly used seed models that serve as starting points for self-improvement. For the detailed methodology, we divide the discussion into three parts as shown in Fig. 2: data collection, data organization, and model optimization. We list current approaches and discuss their differences. We also collect evaluation methods commonly used to measure performance gains from self-improvement, compiling benchmark results for a comprehensive comparison. Additionally, we discuss downstream applications, to highlight the real-world impact of this paradigm. Finally, we identify the challenges in this field, which also represent potential future directions, and conclude the survey.

With this work, we aim to establish a clearer pathway for developing the next generation of MLLMs with better self-improvement mechanisms, moving beyond random exploration with biases. We hope to attract more researchers to explore this promising direction.

2 Overview

In this section, we first formally define self-improvement in multimodal large language models (MLLMs) in the context of this paper, and then compare it to similar concepts that have been used in MLLMs research. Afterwards, we summarize representative works in this domain to provide a general overview of the existing methods.

2.1 Definition

There are many similar terms to Self-Improvement, such as Self-Evolution, Self-Training, Self-Consistency, Self-Correction, Self-Reflection, and Self-Refinement, which have also been mentioned in previous MLLM research. There is a trend where the boundaries between these concepts are becoming blurred, and they may become more interchangeable in the future, depending heavily on the context. However, we clearly distinguish two paradigms. In this paper, we define self-improvement shown in Fig. 1 as updating the model

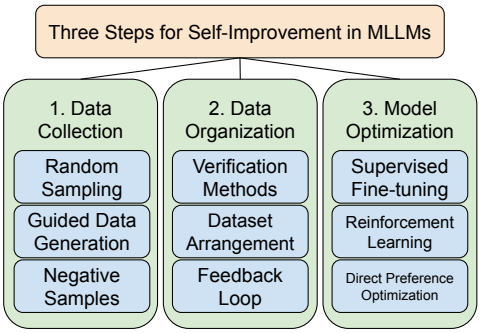

Figure 2: An overview of three steps for self-improvement in MLLMs. Each step can involve different methods based on requirements. For the full taxonomy please check Fig. 3.

from m_0 to m_1, as opposed to self-refinement, which involves updating responses in context from r_0 to r_1. Formally, we express these concepts as follows:

Self-Improvement (Model Update through Training): $m_1 = I(m_0, D)$, where $I(\cdot)$ denotes the self-improvement operator that upgrades the entire model by training on self-curated multimodal dataset D.

Self-Refinement (Response Update in Context): $r_1 = R(r_0, c)$, where $R(\cdot)$ represents the self-refinement operator that refines the initial response r_0 based on the context c, which can be seen as a type of test-time scaling (or test time self-improvement). It is worth noting that some refined responses may have the potential to be incorporated into training data and thus contribute to further self-improvement.[1]

A typical self-improvement process in MLLMs involves three modules: data collection, data organization, and model optimization as demonstrated in Fig. 2, which follows the structure of a general model-building process but focuses on automating the model development process using models rather than relying heavily on human intervention. While these commonly used modules are widely involved in the self-improvement of MLLMs, it is important to note that their life cycle does not necessarily conclude once an improved model is obtained. The iterative loop can persist, using the newly improved model as the seed for the next stage of self-improvement as demonstrated in Fig. 1. This life cycle can be highly dynamic, particularly in online settings, where data collection is directly in-

[1]Here, we do not consider storing newly acquired skills during inference in memory as an analogy for parameter tuning.

fluenced by the optimization design. This design may incorporate or encourage the model to explore more diverse or constrained data generation in subsequent rounds.

We conceptualize self-improvement in MLLMs as a spectrum of methods aiming to reduce human workload and maximize automation in improving model performance. Some methods target full autonomy, while others are limited to guided or assisted self-improvement, as long as they do not fully rely on human effort. Most papers in our survey do not leverage stronger external models. However, external models can be treated as tools that the seed model calls or uses. Under this formulation, we believe such approaches fall within the spectrum of self-improvement, albeit at the less independent end due to their reliance on external tools. To illustrate this, we add Tab. 1 comparing different levels of self-improvement in MLLMs, detailing what they automate and their limitations, allowing all discussed methods to fit organically within this spectrum.

2.2 Related and Representative Works

Improvement without human supervision in MLLMs encompasses various strategies aimed at enhancing model performance through internal mechanisms. These approaches can be broadly categorized into Self-Refinement, Peer-Improvement, Self-Improvement for image LLMs, and extensions for Video LLMs and agents.

2.2.1 Self-Refinement and Peer-Improvement

Early methods like **Woodpecker** (Yin et al., 2023) and **VCD** (Leng et al., 2024) focus on reducing hallucinations within generated content through training-free techniques. Due to the big gap between proprietary models and early open-weight models, **LLaVA** (Liu et al., 2024a) and **HA-DPO** (Zhao et al., 2023) leverage GPT-4 to help build or refine multimodal capabilities, avoiding human supervision from scratch.

2.2.2 Self-Improvement in Image Large Language Models

Self-improvement strategies aim to enhance model abilities fundamentally by modifying model weights and reducing dependency on external models. Recent methods include on-the-fly enhancement of instruction-tuning data **VIGC** (Wang et al., 2024a), shifting from answering generation to self-questioning **SQ-LLaVA** (Sun et al., 2025a), and synergy-driven cycles that interleave describing and locating objects **SC-Tune** (Yue et al., 2024b). Others reduce hallucinations by converting training-free interventions into trainable ones**M3ID** (Favero et al., 2024), enabling interpretability in decision-making without extra annotations like **LLaVA-ASD** (Deng et al., 2024a), leveraging data augmentation to construct preference pairs like **SeVa** (Zhu et al., 2024), and applying step-wise self-rewarding **CSR** (Zhou et al., 2024b). Some approaches rely on internal checks, such as visual metrics for preference tuning **SIMA** (Wang et al., 2024b) or using the model's own encoder for fine-grained alignment **FiSAO** (Cui et al., 2024).

2.2.3 Extensions to Video

i-SRT (Ahn et al., 2024a) applies self-improvement in video large language models, addressing the issue of self-generated preferences that are linguistically plausible but not grounded in the visual content of the associated video. **Video-STaR** (Zohar et al., 2024) adapts the STaR approach for the video domain, enabling the use of any labeled video dataset (such as Kinetics-700) for video instruction tuning.

2.2.4 Multimodal Agents

When augmenting MLLMs as agents and allowing them to act or even interact with each other, self-improvement enhances model performance across various tasks, including learning through self-play in image identification (Konyushkova et al., 2025) or improving decision-making in games such as Blackjack and ALFWorld (Zhai et al., 2025).

3 Seed Models

A seed model does not need to be exceptionally strong, but it must clear a small set of capability floors that the self-improvement loop relies on. If these floors are missing, the model tends to generate low-quality data and the loop either stalls or collapses.

Capability floors. Some skills are costly to "retrofit" purely from self-improvement and therefore should be present in the seed:

- Basic visual grounding
- Robust text-in-the-wild / layout handling
- Temporal aggregation for video
- Coherent reasoning traces (for reflection)

Level	Who does the heavy-lifting?	Typical technique / example
L0 – No self-improvement	Humans do all data collection and curation	InstructGPT-style SFT (Ouyang et al., 2022)
L1 – Human-guided improvement	Model generates responses, while humans choose preferred data	RLHF-V, human-guided reward modelling (Yu et al., 2024a)
L2 – Peer improvement	External models (e.g. GPT-4-V) supply data; minimal direct human toil	Distillation (Liu et al., 2024a),
L3 – Hybrid self-improvement	Model collects its own data, but queries external augmentations or verifiers	Hybrid approaches (Zhou et al., 2024a)
L4 – Conditional self-improvement	Target model runs its own data loop except images come from existing datasets	RLAIF-V (Yu et al., 2024b) with self-reward
L5 – High self-improvement	Model generates and curates both images and text without external data sources	UniRL (Mao et al., 2025)

Table 1: Levels of multimodal self-improvement

Common choices. Several commonly used MLLMs have been adopted as seed models in self-improvement research:

- **LLaVA** (Liu et al., 2024a): As one of the earliest popular MLLMs, LLaVA has been widely used in MLLM self-improvement research due to its representativeness. The most commonly used versions are LLaVA-1.5 (7B and 13B). Some works, such as STIC and BDHS, utilize LLaVA-1.6.
- **Qwen-VL** (Bai et al., 2023): Built on top of Qwen-LM, this model uses a three-stage training pipeline: Pretraining, Multi-task Pretraining, and Supervised Fine-tuning, to optimize its performance.
- **InstructBLIP** (Dai et al., 2023): InstructBLIP introduces an instruction-aware Query Transformer that extracts informative features tailored to given instructions. It is trained on 13 datasets converted into an instruction-tuning format.
- **MiniGPT4** (Zhu et al., 2023): An early effort to reproduce an open-source GPT-4, MiniGPT4 aligns a frozen visual encoder with a frozen advanced LLM (Vicuna) using a single projection layer.
- **Video-LLaVA** (Lin et al., 2023): It is commonly used as a seed model in video models. As its name implies, Video-LLaVA is similar to LLaVA but also fine-tuned on video datasets. It is designed for both image and video comprehension tasks.

Beyond these commonly used seed models, some works train their own seed models from scratch using a pretrained LLM to maintain more control over the entire process and address specific needs.

4 Data Collection

Effective data collection is crucial for enabling MLLMs to acquire and refine specific abilities. In conventional machine learning approaches, data collection typically relies on extensive human labor. This labor-intensive process, while effective, can be both time-consuming and costly, and is often limited by the availability and scalability of human resources.

In the context of self-improvement for MLLMs, a shift towards autonomous data collection is both desirable and increasingly feasible, thereby reducing the dependency on human intervention. This approach not only enhances efficiency but also enables continuous and scalable learning. We compare advantages and disadvantages of these methods in Tab. 2.

4.1 Random Sampling

The most straightforward method for autonomous data generation is random sampling (Zhao et al., 2023), where the model generates data by sampling from its existing knowledge base without specific guidance. Although random sampling is simple to implement and can produce a diverse set of data, it has notable inefficiencies such as the generation of redundant or irrelevant data, which can waste computational resources and time.

4.2 Guided Data Generation

To address the inefficiencies of random sampling, guided data generation techniques have been developed (Cheng et al., 2024). These methods employ predefined pipelines with carefully designed prompts to steer the model towards generating desired and high-quality responses. One prominent technique is Chain-of-Thought (CoT), which encourages the model to generate intermediate reasoning steps before producing a final answer. In order to further improve sample efficiency, some approaches adopt search-based methods such as beam search and Monte Carlo Tree Search (MCTS) and its variants (Yao et al., 2024).

4.3 Negative Samples

Negative samples are essential for refining the model's ability to distinguish between correct and

Table 2: Comparison of Data Collection Methods

Method	Benefits	Drawbacks
Random Sampling (Zhao et al., 2023; Yu et al., 2024b)	Easy to use; works for any MLLM	May not be efficient; difficult to obtain samples with desired features
Prompt-Guided Generation (Wang et al., 2024a; Fang et al.)	Highly controllable; can generate almost any type of response	Requires significant human effort; difficult to scale
Chain of Thought (Zhai et al., 2025; Zohar et al., 2024)	Can generate long responses for reasoning tasks	Sometimes produces redundant or irrelevant reasoning steps
Input Injection (Zhou et al., 2024a; Zhu et al., 2024)	Can generate negative examples	Minor distortions may sometimes produce better examples than undistorted ones
Sourcing from Multiple MLLMs (Li et al., 2023a; Xiong et al., 2024)	Ensures diversity in generated outputs	Requires additional effort to manage different models

incorrect responses, thereby enhancing its overall accuracy and reliability. Various strategies have been explored to generate negative samples autonomously. **Poorly Designed Prompts** (Deng et al., 2024b): Crafting ambiguous or misleading prompts can lead the model to generate suboptimal or incorrect responses. **Distorted Images** (Zhou et al., 2024a): Introducing visual distortions or noise into images challenges the model's visual comprehension capabilities. **Attention Masking** (Amirloo et al., 2024): Manipulating the attention mechanism during the decoding process can result in responses that focus on irrelevant parts of the input. Additionally, the generation of negative samples can be finely controlled by altering the decoding path (Deng et al., 2025b), which produces responses that are less grounded in the visual context to the desired level, serving as effective negative examples for training.

Some methods utilize peer models for data generation (distillation), but implementing the same pipeline with the seed model itself may theoretically produce similar effects.

5 Data Organization

The data collected by MLLMs may not be directly suitable for feeding back into the models without further processing. To ensure the efficacy of self-improvement, a thorough verification and processing step is essential before leveraging the newly obtained data. The quality of this organization process is paramount, as it directly determines the robustness and reliability of the self-improvement mechanism in MLLMs. We compare these methods in Tab. 3.

5.1 Verification Methods

The verification process can be a critical step during data organization and is usually implemented using either predefined rules or sophisticated models. Each method has its own advantages and limitations, which are discussed below.

5.1.1 Rule-Based Verification

Rule-based organization involves applying predefined criteria to assess the quality and correctness of the generated data. This approach is straightforward and computationally efficient but may lack flexibility in handling diverse data scenarios. **Majority Voting (Ensembling or Consensus):** The simplest approach compares multiple generated responses and selects the one with the highest frequency. While easy to implement, it may not always yield the best quality data, as the most frequent response might still contain inaccuracies or lack diversity. **Ground Truth Alignment** (He et al., 2024a): For datasets with established ground truths, the verification can involve cross-referencing the model's output with the correct answers. For instance, in terms of the tasks requiring bounding boxes, an Intersection over Union (IoU) threshold can determine the acceptance of generated content (Yue et al., 2024b). If the IoU score exceeds the predefined threshold, the content is deemed acceptable; otherwise, it can be discarded or flagged for further review. Alternatively, IoU can also be used as a reward function during RL training (Liu et al., 2025a).

5.1.2 Model-Based Verification

Model-based organization leverages additional models to assess the quality of generated data. This method can provide more nuanced evaluations and modifications but may introduce additional computational overhead. **Peer Model Evaluation:** Utilizing separate peer models to judge the quality of outputs can reduce bias and improve the reliability of the verification process. These models can provide independent evaluations, enhancing the overall robustness of data verification. **Self-Critic Mechanism** (Wang et al., 2024b): The MLLM itself can generate the rewards that evaluate the correctness and relevance of the data at various levels-token (Cui et al., 2024), sentence (Zhou et al., 2024b), or output. This allows for more detailed assessments compared to rule-based methods.

Table 3: Comparison of Data Organization Methods

Method	Benefits	Drawbacks
Pre-assigned Labels (Zhou et al., 2024a)	No extra effort required after data collection	Cannot handle complex cases
Rule-Based Organization (Yue et al., 2024b)	Highly explainable	Not robust enough for novel samples
Self-Evaluation (Ahn et al., 2024b)	No additional reliance on external tools	Can suffer from model bias or hallucinations
Judgment by External Verifiers (Sun et al., 2024a)	Well-defined verifiers are highly robust	Some verifiers may incur significantly higher costs
Feedback from Environment (Zhai et al., 2025)	Robust and requires minimal additional effort	Many cases may be difficult to implement

5.1.3 Verification from the Environment

MLLMs used as agents that interact with their environment can also leverage environmental feedback for verification. The environment can be either the real world (Guo et al., 2025a; Chen et al., 2025d) or simulated environments, such as games (Zhai et al., 2025; Konyushkova et al., 2025).

5.2 Dataset Arrangement

Arranging the newly-collected and verified dataset can be an important step in the self-improvement (Wang et al., 2024a). Based on the goal of self-improvement, the new dataset can be formulated from previously generated and processed data: e.g., fixing answers, rewriting rationales, inserting missing evidence, or normalizing formats. Depending on the goal, the arranged set can be derived from prior model outputs by:

- *editing/refinement* of outputs or rationales (e.g., generator- corrector workflows and reflective self-correction) (Wang et al., 2024a; Zhang et al., 2024b; He et al., 2024a),
- *topic-aware overwriting* where errors are corrected within semantic clusters (He et al., 2024b),
- *curriculum or subset scheduling* when the emphasis is ordering/pruning rather than rewriting (Deng et al., 2025a).

When rated scores from judges are available, the same pool can be *reformatted* into preference-learning pairs/lists for DPO or into continuous rewards for RL (e.g., critic/reward-based organization) (Xiong et al., 2024).

5.3 Feedback Loop

The data collection-organization pipeline in self-improving MLLMs is not necessarily unidirectional. Verified or re-scored outputs can be fed back to the generation model to modify prompts, constraints, or exemplars, creating a closed feedback loop that iteratively enhances data quality and model performance.

Iterative Refinement (data-centric): In each round, the current model generates training candidates, the organization step verifies, filters, or transforms them into a curated set, and the model is updated on this set before repeating. This improves context quality and reduces noise over successive iterations (Liu et al., 2024c; Deng et al., 2025c; Luo et al., 2024a).

Recursive Improvement (model-centric): The loop also supports upgrading the teacher/critic/peer panel (or the policy itself), so that the next round of data is produced and/or filtered by a strictly stronger model, enabling co-evolution of data and capability (Tan et al., 2024; Mao et al., 2025; Hong et al., 2025; Chen et al., 2025c).

6 Model Optimization

After obtaining the filtered dataset, the next step is to update the parameters of the seed model. Several training methods have been employed in self-improvement for MLLMs, including supervised fine-tuning, reinforcement learning, and direct preference optimization. As discussed in the paper DeepSeekMath (Shao et al., 2024), all these methods are actually connected. We compare advantages and disadvantages of these methods in Tab. 4.

Table 4: Comparison of Model Optimization Methods

Method	Benefits	Drawbacks
SFT (Wang et al., 2024a; Luo et al., 2024a; Xiong et al., 2024)	Highly efficient when using existing high-quality datasets	Requires human effort or high-cost strong models
PPO (Yue et al., 2024b; Zhai et al., 2025)	A classic online RL method, easy to deploy	The reward model may be difficult to obtain
GRPO (Chen et al., 2025b)	More efficient than PPO since no value model is needed	Involves a trade-off between efficiency and the number of groups
RFT (Liu et al., 2024c)	Can be used in an offline manner	All negative samples are discarded
DPO (Li et al., 2023a; Ouali et al., 2025; Luo et al., 2024b)	Can leverage both positive and negative samples	May experience distribution shift issues after extensive training

6.1 Supervised Fine-tuning

Instruction tuning, or supervised fine-tuning (SFT), has become a widely adopted post-training method to enable LLMs and MLLMs to follow instructions and solve a broader range of general tasks. In

SFT, the model is trained to minimize the discrepancy between its predictions and the ground truth responses provided in the dataset.

Formally, given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the input and y_i the corresponding target output, the objective is to minimize the cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} y_{i,t} \log p(y_{i,t}|x_i, y_{i,<t}; \theta) \quad (1)$$

This loss function encourages the model to generate outputs that closely match the ground truth. In the context of self-improvement for MLLMs, it enables the new model to better align with the desired improvement goals in generated output.

6.2 Reinforcement Learning

Reinforcement learning (RL) methods have been used to improve MLLMs without human demonstration data, particularly for preference alignment and reasoning tasks. It aims to generate outputs that receive high rewards. The objective is then expressed as:

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_{(x,y)\sim D_{\pi_\theta}} r(x, y) \quad (2)$$

Methods such as Proximal Policy Optimization (PPO) have been initially employed in RLHF for MLLMs (Sun et al., 2023). More recently, GRPO (Shao et al., 2024) has emerged as an efficient alternative to PPO for training MLLMs (Chen et al., 2025b), as it does not require a value model.

6.3 Direct Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al., 2024) is a reinforcement-learning-free offline alternative for preference learning, which has become the de facto standard in preference optimization for MLLMs. Unlike SFT, which can only leverage positive data, it can also take advantage of negative data. It formulates the optimization problem as follows:

Given a pair of outputs (y^+, y^-) where y^+ is preferred over y^-, the objective is to maximize the likelihood of preferred outputs while minimizing the likelihood of dispreferred outputs. The DPO loss can be expressed as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[\log \sigma(s(y_i^+) - s(y_i^-))\right] \quad (3)$$

This objective encourages the model to assign higher scores to preferred outputs compared to dispreferred ones.

6.4 Other Enhanced Variants

Some works adjust the classic method by adding additional components (Xiao et al., 2024), such as penalty terms for specific designs. For example, incorporating regularization terms can help maintain model stability and prevent overfitting to the preference data.

6.5 Alternative Ways of Using Negative Samples

It is worth noting that preference learning is not the only way to utilize negative data samples. Combining negative samples with self-reflection and correction using a CoT approach can further enhance model performance. This involves generating detailed reasoning steps that allow the model to identify and correct its own errors, thereby improving the quality of the outputs.

6.6 Curriculum

Multi-stage training with different optimization methods has become common practice in MLLM training. Some research shows that certain training stages may hurt the model (Zhou et al., 2025), while other studies find that certain performance gains can be more easily obtained by combining different stages of optimization (Huang et al., 2025b).

7 Dataset and Evaluation

There are some datasets released for improving MLLMs but no benchmarks specifically designed for self-improvement in MLLMs. Most of the time, researchers use existing MLLM benchmarks and report performance gains compared to the seed model and other SOTA models. Some attempts, such as LLM-Evolve (You et al., 2024) aim to build a new type of benchmark; however, this particular benchmark operates in a non-parametric setting.

7.1 Dataset

Some datasets aim to improve MLLMs, such as **VLFeedback (Li et al., 2024b).** VLFeedback is the first large-scale AI-annotated vision-language feedback dataset. It contains over 82K multi-modal instructions and comprehensive rationales generated by models. Additional datasets have been contributed by the community, including those used in **RLAIF-V** and **Open-R1-Multimodal**. These AI-created datasets have demonstrated their usefulness in improving various MLLMs. However, they remain limited to specific tasks and offer only

incremental improvements. The challenge of building a more general dataset capable of supporting a wider range of tasks remains an open question for the research community.

Several new datasets have emerged for the purpose of self-improvement for MLLMs. For instance, the DeepPerception Dataset (Ma et al., 2025) aims to enhance the cognitive visual perception capabilities of MLLMs for the task of knowledge-intensive visual grounding (KVG); it comprises high-quality, knowledge-aligned training samples generated through an automated data synthesis pipeline. The OmniAlign-V-DPO Dataset (Zhao et al., 2025b) leverages the answers from the OmniAlign-V SFT dataset as positive examples. To create the necessary preference pairs for DPO, negative samples are generated using another MLLM, LLaVANext-InternLM-7B, through a process called rejection sampling. The Vision-Prefer Dataset (Wu et al., 2024) is another high-quality and fine-grained preference dataset created for aligning text-to-image generative models. It aggregates feedback from AI annotators, specifically utilizing the capabilities of GPT-4V to evaluate generated images based on defined criteria. The LLaVA-Critic dataset (Xiong et al., 2024), comprising 113,000 evaluation instruction samples across 46,000 images, was generated using a GPT-assisted pipeline, with GPT-4o providing judgment scores and reasons for evaluating MLLM responses. We summarize these datasets in the following table to better demonstrate their differences.

7.2 Benchmarks

Evaluating the self-improvement of MLLMs can leverage current popular MLLM benchmarks. These benchmarks can be broadly categorized as follows:

7.2.1 General Knowledge

Benchmarks in this category assess the model's ability to understand and reason across multiple disciplines using multimodal inputs. Notable benchmarks include **MMMU** (Yue et al., 2024c) and **MMStar** (Chen et al., 2024a), which focus on comprehensive multimodal understanding across various academic and professional domains.

7.2.2 Reasoning

These benchmarks evaluate higher-order cognitive abilities and commonsense reasoning within multimodal contexts. Examples such as **Mathvista** (Lu et al., 2023) and **VCR** (Zellers et al., 2019) are designed to test mathematical reasoning and commonsense understanding through visual inputs.

7.2.3 Hallucination

Detecting and mitigating hallucinations in generated content is crucial for reliable MLLMs. Benchmarks like **CHAIR** (Rohrbach et al., 2018), **POPE** (Li et al., 2023b), and **AMBER** (Wang et al., 2023) provide metrics and evaluation frameworks to assess the accuracy and relevance of model outputs against visual inputs.

7.2.4 Medical

Medical benchmarks focus on the model's capability to understand and reason with medical images and related queries. Datasets such as **VQA-RAD** (Lau et al., 2018), **SLAKE** (Liu et al., 2021), and **PathVQA** (He et al., 2020) are designed to evaluate the model's proficiency in medical image analysis and question-answering tasks.

7.2.5 Video QA

Assessing MLLMs' understanding of dynamic visual content is addressed by video-based benchmarks. Notable datasets include **MSVD-QA** (Xu et al., 2017), **MSRVTT-QA** (Xu et al., 2017), **TGIF-QA** (Jang et al., 2017), and **ActivityNet-QA** (Yu et al., 2019), which provide question-answer pairs based on video clips to test temporal and contextual reasoning.

7.2.6 Judging Abilities

Evaluating the model's capability to act as a judge involves assessing various aspects such as alignment, safety, and bias. Benchmarks like **MJ-Bench** (Chen et al., 2024b) are designed to measure these attributes, ensuring that the model's evaluations are reliable and consistent. Meanwhile, **AutoBench-V** (Bao et al., 2024) attempts to enable the MLLM itself to propose and construct new benchmarks.

7.3 Meta-Analysis Across Benchmarks

Using the compiled results, we observed the following robust patterns:

Method-Task Match. Rule-/verification-based RL (e.g., with step-wise or outcome checks) drives the largest absolute gains on verifiable tasks (visual math, programmatic reasoning, constrained captioning), while preference/AI-feedback data most reliably lowers hallucination metrics (e.g.,

Table 5: Comparison of Vision-Language Feedback Datasets

Dataset Name	Feedback Type	Key Benefits	Limitations
DeepPerception	Implicit (Task-based RL)	Focuses on perception-cognition synergy; automated data synthesis.	PPO usage is implied, not explicitly stated; size of synthesized data not clear.
VLFeedback	AI-generated (GPT-4V)	Large scale; diverse instructions; safety-focused; generated by strong MLLM.	Potential biases from GPT-4V.
OmniAlign-V-DPO	AI-generated (LLaVANext)	DPO-specific format; uses rejection sampling for negative samples.	Quality depends on base SFT data and negative sampling strategy.
VisionPrefer	AI-generated (GPT-4V)	Fine-grained preferences across multiple aspects.	Primarily for reward model training, DPO usage for validation.
LLaVA-Critic	AI-generated (GPT-4o)	Reliable evaluation scores; generates reward signals for DPO; diverse tasks covered.	Quality of reward signals depends on LLaVA-Critic's performance.

POPE/AMBER) and improves general helpfulness/faithfulness.

Seed Strength Matters. Relative improvement Δ_{seed} typically shrinks as seed models get stronger; however, strong seeds show more stable gains across benchmarks. For identical pipelines (e.g., STIC-style), better seeds consistently yield higher end performance.

Cross-Benchmark Inconsistency. Methods that boost compositional reasoning can regress on perception-heavy tasks (fine-grained recognition, OCR, attribute binding), and vice versa. Pairwise rank correlations between benchmarks are often modest; gains on one suite do not guarantee gains on others.

Persistent Bottlenecks. We observe recurring difficulties in:

- Fine-grained spatial reasoning (counting under occlusion, relative positions)
- Multi-image/multi-hop consistency
- Long-horizon video temporal grounding
- Diagram/chart/plan understanding
- Robustness under noisy OCR or layout-heavy documents
- Hallucination recurs in open-world scenes unless visual evidence is tightly verified.

Judge/Reward Leakage. When the same or closely related judges curate and evaluate (e.g., GPT-4V-like feedback used both for data construction and testing), scores inflate. Separation of curation and evaluation signals is critical for credible claims.

Efficiency Analysis. We discuss the efficiency of self-improvement methods in MLLMs from a computational cost perspective, considering factors like memory use during each stage and the data generation scale. First, regarding data sampling: random sampling often has the highest cost since it normally has a high rejection rate. Prompt-guided generation helps address this issue by giving more guidance, thereby reducing the search space of possible responses. Using negative samples further enables the usage of all generated data; even samples considered low-scoring can be used as negative samples, thus avoiding waste. For verification methods, the rule-based method generally has the lowest cost, since checking whether generated content satisfies rules is typically straightforward. Model-based verification can handle very complex scenarios but has the highest cost. Verifying the outcome in the real environment can have the highest cost due to simulation complexity but may yield the highest feedback quality, especially for the most difficult verifications.

8 Conclusion

In this paper, we presented a comprehensive and structured survey of self-improvement in multimodal large language models (MLLMs). We defined the concept of self-improvement as used in this survey and clarified its differences from other related concepts. We discussed and compared representative works in this domain, highlighting their similarities and differences from three perspectives: 1) data collection, 2) data organization, and 3) model optimization. Further, we summarized commonly used evaluations and applications. Finally, we identified current challenges and potential opportunities for future research. We hope this survey serves as a valuable guide for researchers interested in exploring and developing new self-improvement methods for MLLMs.

Limitations

Due to space limitations, this paper primarily focuses on a macro-level description and analysis of self-improvement within the current scope of MLLMs. Given the rapid evolution of the field, some of the most recent developments and new directions may not be included. Since we focus on the MLLM domain, we did not review work that involves only LLMs or agents; however, some methods may potentially be adapted to MLLMs as well. Despite these limitations, we believe this work, as the first survey in the area of self-improvement in MLLMs, provides a valuable overview of current research.

References

Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024a. i-srt: Aligning large multimodal models for videos by iterative self-retrospective judgment. *arXiv preprint arXiv:2406.11280*.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024b. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. 2024. Self-training: A survey. *Neurocomputing*, page 128904.

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. 2024. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*.

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

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuying Chen, Mohamed Elhoseiny, and Xiangliang Zhang. 2024. Autobench-v: Can large vision-language models benchmark themselves? *arXiv preprint arXiv:2410.21259*.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

Kejia Chen, Jiawen Zhang, Jiacong Hu, Jiazhen Yang, Jian Lou, Zunlei Feng, and Mingli Song. 2025a. Shape: Self-improved visual preference alignment by iteratively generating holistic winner. *arXiv preprint arXiv:2503.04858*.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025b. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

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

Xiuwei Chen, Wentao Hu, Hanhui Li, Jun Zhou, Zisheng Chen, Meng Cao, Yihan Zeng, Kui Zhang, Yu-Jie Yuan, Jianhua Han, et al. 2025c. C2-evo: Co-evolving multimodal data and model for self-improving reasoning. *arXiv preprint arXiv:2507.16518*.

Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. 2025d. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *arXiv preprint arXiv:2502.05450*.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. 2024b. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*.

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

Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. 2024. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. *arXiv preprint arXiv:2410.14148*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. 2025a. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*.

Shijian Deng, Erin E Kosloski, Siddhi Patel, Zeke A Barnett, Yiyang Nan, Alexander Kaplan, Sisira Aarukapalli, William T Doan, Matthew Wang, Harsh Singh, et al. 2024a. Hear me, see me, understand me: Audio-visual autism behavior recognition. *IEEE Transactions on Multimedia*.

Shijian Deng, Wentian Zhao, Yu-Jhe Li, Kun Wan, Daniel Miranda, Ajinkya Kale, and Yapeng Tian. 2025b. Efficient self-improvement in multimodal large language models: A model-level judge-free approach. In *Second Conference on Language Modeling*.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025c. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024b. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, et al. 2025. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. *arXiv preprint arXiv:2508.07407*.

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jan Kautz, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. Vilaˆ 2: Vlm augmented vlm with self-improvement.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025a. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025b. Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. 2025. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. 2025a. Improving vision-language-action model with online reinforcement learning. *arXiv preprint arXiv:2501.16664*.

Zilu Guo, Hongbin Lin, Zhihao Yuan, Chaoda Zheng, Pengshuo Qiu, Dongzhi Jiang, Renrui Zhang, Chun-Mei Feng, and Zhen Li. 2025b. Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. *arXiv preprint arXiv:2503.10529*.

Jiayi He, Hehai Lin, Qingyun Wang, Yi Fung, and Heng Ji. 2024a. Self-correction is more than refinement: A learning framework for visual and language reasoning tasks. *arXiv preprint arXiv:2410.04055*.

Lehan He, Zeren Chen, Zhelun Shi, Tianyu Yu, Jing Shao, and Lu Sheng. 2024b. A topic-level self-correctional approach to mitigate hallucinations in mllms. *arXiv preprint arXiv:2411.17265*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. 2025. Reinforcing multimodal understanding and generation with dual self-rewards. *arXiv preprint arXiv:2506.07963*.

Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. 2025a. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *arXiv preprint arXiv:2503.18135*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025b. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Weiyang Jin, Baihan Yang, Huan-ang Gao, Jingwei Zhao, Kangliang Chen, and Hao Zhao. Spa: Enhancing 3d multimodal llms with mask-based streamlining preference alignment.

Ksenia Konyushkova, Christos Kaplanis, Serkan Cabi, and Misha Denil. 2025. Vision-language model dialog games for self-improvement. *arXiv preprint arXiv:2502.02740*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Boyu Li, Haobin Jiang, Ziluo Ding, Xinrun Xu, Haoran Li, Dongbin Zhao, and Zongqing Lu. 2024a. Selu: Self-learning embodied mllms in unknown environments. *arXiv preprint arXiv:2410.03303*.

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

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024b. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. *arXiv preprint arXiv:2410.09421*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. SceMQA: A scientific college entrance level multimodal question answering benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 109–119, Bangkok, Thailand. Association for Computational Linguistics.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Sannyuya Liu, Jintian Feng, Zongkai Yang, Yawei Luo, Qian Wan, Xiaoxuan Shen, and Jianwen Sun. 2024b. Comet:"cone of experience" enhanced large multimodal model for mathematical problem generation. *Science China Information Sciences*, 67(12):1–2.

Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. 2024c. Diving into self-evolving training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025a. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*.

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 preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. 2024a. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024b. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. 2025. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*.

Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. 2025. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning. *arXiv preprint arXiv:2505.23380*.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2025. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pages 395–413. Springer.

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

Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. 2025. Skywork r1v: pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*.

Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. 2024. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*.

Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. 2024. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. 2024. Silmm: Self-improving large multimodal models for compositional text-to-image generation. *arXiv preprint arXiv:2412.05818*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. 2024a. Stllava-med: Self-training large language and vision assistant for medical question-answering. *arXiv preprint arXiv:2406.19973*.

Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. 2024b. Sq-llava: Self-questioning for large vision-language assistant. In *European Conference on Computer Vision*, pages 156–172. Springer.

Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. 2025a. Sq-llava: Self-questioning for large vision-language assistant. In *European Conference on Computer Vision*, pages 156–172. Springer.

Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. 2025b. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *arXiv preprint arXiv:2505.18536*.

Zeyi Sun, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Tong Wu, Dahua Lin, and Jiaqi Wang. 2025c. Seagent: Self-evolving computer use agent with autonomous learning from experience. *arXiv preprint arXiv:2508.04700*.

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

Wentao Tan, Qiong Cao, Yibing Zhan, Chao Xue, and Changxing Ding. 2024. Beyond human data: Aligning multimodal large language models by iterative self-evolution. *arXiv preprint arXiv:2412.15650*.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

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

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025a. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Wanfu Wang, Qipeng Huang, Guangquan Xue, Xiaobo Liang, and Juntao Li. 2025b. Learning active perception via self-evolving preference optimization for gui grounding. *arXiv preprint arXiv:2509.04243*.

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

Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, et al. 2025. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025a. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*.

Weijia Wu, Chen Gao, Joya Chen, Kevin Qinghong Lin, Qingwei Meng, Yiming Zhang, Yuke Qiu, Hong Zhou, and Mike Zheng Shou. 2025b. Reinforcement learning in vision: A survey. *arXiv preprint arXiv:2508.08189*.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Multimodal large language model is a human-aligned annotator for text-to-image generation. *arXiv preprint arXiv:2404.15100*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

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. *arXiv preprint arXiv:2410.02712*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Llm-evolve: Evaluation for llm's evolving capability on benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16937–16942.

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

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

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Junpeng Yue, Xinru Xu, Börje F Karlsson, and Zongqing Lu. 2024a. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. *arXiv preprint arXiv:2410.03450*.

Tongtian Yue, Jie Cheng, Longteng Guo, Xingyuan Dai, Zijia Zhao, Xingjian He, Gang Xiong, Yisheng Lv, and Jing Liu. 2024b. Sc-tune: Unleashing self-consistent referential comprehension in large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13073–13083.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024c. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. 2025. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems*, 37:110935–110971.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Vision-r1: Evolving human-free alignment in large

vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024b. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2407.11422*.

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Yifan Sun, Haocheng Feng, et al. 2024c. Automated multi-level preference for mllms. *arXiv preprint arXiv:2405.11165*.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. 2024d. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025a. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. 2025b. Omnialign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*.

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

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, et al. Anyprefer: An automatic framework for preference data synthesis. In *Neurips Safe Generative AI Workshop 2024*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 291–300.

Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. 2024. Video-star: Self-training enables video instruction tuning with any supervision. *arXiv preprint arXiv:2407.06189*.

A Full Taxonomy

For space reasons, the main paper provides only an overview of our taxonomy of self-improvement in MLLMs. This appendix presents the complete hierarchy covering Data Collection (§4), Data Organization (§5), and Model Optimization (§6), and annotates each branch with representative works. See Fig. 3 for the full diagram.

B Applications

Self-improvement can be particularly useful for applications that lack sufficient related instruction data. Models can autonomously generate the required data and conduct self-improvement to acquire new skills for downstream tasks.

B.1 Math & Science

Tasks in fields like math and many other sciences require advanced reasoning sometimes including multimodal reasoning to address. However, the underlying reasoning data is not abundant, since humans seldom write down all the details of their reasoning steps, let alone reasoning that occurs via unconscious pathways. Self-improvement frameworks combined with peer-improvement have enabled MLLMs to autonomously generate and refine multimodal reasoning content, significantly reducing reliance on human-annotated data. For example, methods like MAVIS (Zhang et al., 2024d) and COMET (Liu et al., 2024b) enhance mathematical reasoning by generating problems and visual explanations through structured prompts and alignment techniques. Similarly, frameworks like G-LLaVA (Gao et al., 2023) integrate geometry-specific tasks with generated datasets, achieving state-of-the-art performance on benchmarks like ScienceQA (Lu et al., 2022), SceMQA (Liang et al., 2024) and PHYSICS (Feng et al., 2025b).

B.2 Control

Self-improvement in MLLMs can be applied to real-world applications such as control. Recent work (Zhou et al.) proposes an automatic framework for preference data synthesis and employs an MLLM with an image segmentation model as a tool, judged by GPT-4o, to improve object segmentation and trajectory generation. The proposed method achieved a 15.50% improvement in four visuo-motor control tasks.

B.3 Healthcare

Exciting advancements, such as STLLaVA-Med (Sun et al., 2024b), have introduced the Self-Training Large Language and Vision Assistant for medical applications. This innovative approach focuses on training a policy model (an MLLM) to auto-generate medical visual instruction data, improving data efficiency through Direct Preference Optimization (DPO). Notably, a more robust and larger model (e.g., GPT-4o) serves as a biomedical expert, guiding the DPO fine-tuning process on the auto-generated data to effectively align the policy model with human preferences. This method achieves impressive zero-shot performance on three major medical VQA benchmarks: VQA-RAD, SLAKE, and PathVQA, while using only 9% of the available medical data. Additionally, LLaVA-ASD (Deng et al., 2024a) has explored using self-improvement approaches to enable MLLMs not only to assist in screening but also to provide explanations for their decision-making processes. This advancement offers a more explainable AI-assisted screening approach, enhancing transparency and user trust.

B.4 Personalization

With self-improvement approaches, users can easily personalize MLLMs (Pi et al., 2024; Pham et al., 2024) using automated pipelines to construct datasets and train models for their own use, requiring minimal additional effort.

B.5 3D and embodied intelligence

Recent advances in self-improvement for MLLMs also benefit areas such as 3D and embodied intelligence. A notable example is the MLLM-For3D framework (Huang et al., 2025a), which introduces a method for achieving 3D reasoning segmentation without the need for explicitly labeled 3D training data. This framework leverages pre-trained 2D MLLMs to generate multi-view pseudo segmentation masks along with corresponding text embeddings. These 2D masks are then projected into 3D space and aligned with the text embeddings, effectively transferring the 2D model's understanding to the 3D realm. Similarly, PiSA-Engine (Point-Self-Augmented-Engine) (Guo et al., 2025b) has been introduced as a novel approach for generating instruction point-language datasets enriched with 3D spatial semantics. Streamlining Preference Alignment (Jin et al.), a post-training stage designed for

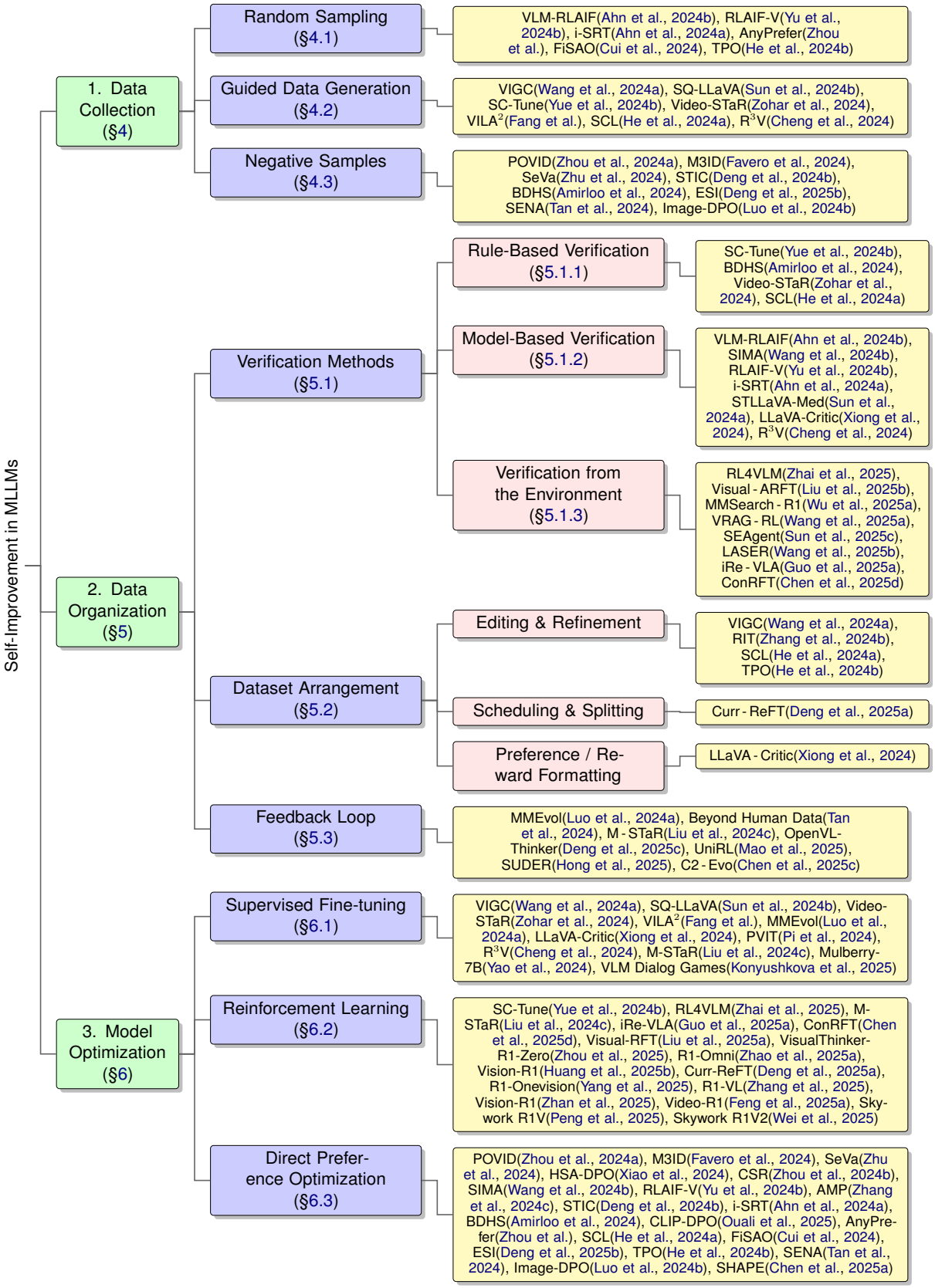

Figure 3: The taxonomy of three steps for self-improvement in MLLMs. Each step can involve different methods based on requirements.

MLLMs equipped with 3D encoders, enhances the ability of MLLMs to understand and reason about 3D spatial relationships, which is fundamental for their effective application in 3D environments.

Self-improvement offers a powerful paradigm for enabling MLLM agents to improve their performance in embodied tasks through interaction with their environment. An example is SELU (Self-Learning in Unknown Environments) (Li et al., 2024a), which allows MLLMs to improve their capabilities in embodied tasks without relying on explicit external human or environmental feedback. SELU adopts an actor-critic framework consisting of two MLLM components: the critic MLLM is responsible for evaluating the outcomes of the actor's actions and for improving its understanding of the environment. Simultaneously, the actor MLLM is improved based on the self-feedback provided by the critic. MART (MLLM As ReTriever) (Yue et al., 2024a) is another example that enhances the performance of embodied agents by utilizing interaction data to fine-tune an MLLM retriever based on preference learning.

C Challenges and Opportunities

Self-improvement in MLLMs presents unique challenges and opportunities compared to text-only LLMs. We expand on these below:

C.1 Uniqueness of Multi-Modality

Many tasks and objectives in MLLMs fundamentally differ from those in LLMs. While LLMs primarily focus on maximizing the likelihood of text sequences, MLLMs must handle objectives incorporating spatial and temporal understanding. For instance, tasks involving images I or videos V require objectives beyond sequential prediction:

- **Spatial Understanding (e.g., Object Detection):** Requires predicting bounding boxes $B = \{b_k\}$ and classes $C = \{c_k\}$. The objective might take the form:

  $$\mathcal{L}_{\text{spatial}} = \sum_k \left(\mathcal{L}_{\text{cls}}(c_k | I; \theta) + \lambda \mathcal{L}_{\text{reg}}(b_k | I; \theta) \right)$$

  where $\mathcal{L}_{\text{cls}}$ is a classification loss and $\mathcal{L}_{\text{reg}}$ is a bounding box regression loss.

- **Temporal Understanding (e.g., Video Action Recognition):** Requires understanding sequences of frames $V = (f_1, \ldots, f_m)$ to predict an action a. The objective could be:

  $$\mathcal{L}_{\text{temporal}} = -\log P(a | V; \theta)$$

Cross-modal alignment and distillation without high-quality data (Liu et al., 2024a) might introduce multimodal hallucination. While text-only LLMs can hallucinate facts, MLLMs can hallucinate content inconsistent with an input image or other modality.

C.2 Better Seed Models and Emerging Modalities

Current self-improvement in MLLMs primarily operates on a limited set of modalities, typically $\mathcal{M}_{\text{current}} = \{\text{Text}, \text{Image}, \text{Video}\}$. The action space $\mathcal{A}$ for self-correction or data generation is often confined to textual outputs. However, significant potential lies in emerging modalities like Audio (A), 3D data (D), and Embodied Actions (Act), extending the modality set to $\mathcal{M}_{\text{emerging}} = \mathcal{M}_{\text{current}} \cup \{A, D, \text{Act}, \ldots\}$.

Expanding to these domains, particularly embodied AI, drastically increases the complexity and dimensionality of the action space. Self-improvement must transition from generating primarily discrete textual actions $a \in \mathcal{A}_{\text{text}}$ to generating sequences of potentially continuous or high-dimensional actions $a_t \in \mathcal{A}_{\text{embodied}}$ required for interaction within an environment E. The optimization objective shifts towards maximizing expected return in sequential decision-making tasks:

$$\max_{\pi_\theta} \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t) \right]$$

where $\tau = (s_0, a_0, s_1, a_1, \ldots)$ is a trajectory generated by policy π_θ in environment E, s_t is the state (often multimodal), $a_t \in \mathcal{A}_{\text{embodied}}$, R is the reward function, and γ is the discount factor. Works like (Zhai et al., 2025; Guo et al., 2025a; Chen et al., 2025d) are beginning to explore self-improvement in these expanded action and modality spaces.

C.3 Omni I/O

A limitation in MLLM self-improvement is the restricted input/output pipeline. Current models M often follow mappings like $M : (\mathcal{M}_{\text{in}}, T_{\text{prompt}}) \to T_{\text{out}}$, where $\mathcal{M}_{\text{in}}$ might be I or V. Generating the non-textual input data (e.g., images I) often requires external datasets or separate generative models (Luo et al., 2024b). This also limits MLLMs capabilities of self-verification and correction without extra models while forced to do so may compounding hallucinations.

True "Omni I/O" capability implies a model M_{omni} that can handle arbitrary combinations of modalities as both input and output. Let $\mathbb{M}$ be

the set of all relevant modalities. The mapping becomes:

$$M_{\text{omni}} : \{m_i\}_{i=1}^{N_{\text{in}}} \to \{m'_j\}_{j=1}^{N_{\text{out}}}$$

where each $m_i \in \mathbb{M}$ and $m'_j \in \mathbb{M}$. For self-improvement, this means the model should ideally be able to generate its own training data across modalities, such as $M_{\text{omni}} : T \to I$, $M_{\text{omni}} : I \to T$, $M_{\text{omni}} : (I, A) \to (T, V)$, etc., potentially in an interleaved manner. Recent advances like native image generation in GPT-4o/Gemini and open-source efforts like Qwen2.5-Omni (Xu et al., 2025) suggest potential towards this goal, where self-improvement could enhance generation and understanding across text, vision, and audio within a single loop. Some work (Qu et al., 2024) has begun to unify these areas.

C.4 Biases and Robust Verification

After obtaining initial generated data, further verification and organization of this raw data are necessary, as we formulated these as the next steps for conducting self-improvement after collecting data. However, even with these controls, there is still no guarantee that bias and incorrectness can be eliminated. This is a significant challenge and an unsolved problem in self-improvement, as the bias may accumulate and potentially stop further recursive improvement, which presents a good opportunity for future research. The feasibility of self-improvement is intrinsically linked to the ability to reliably *verify* the quality or correctness of the model's outputs. This echoes the computational complexity concept related to P vs NP: generating optimal outputs might be hard, but verifying them should ideally be tractable. We can formalize this with a verification function $V(x, y)$, where x is the input and y is the MLLM's output (which could be multimodal). $V(x, y)$ returns a score or a binary judgment (correct/incorrect, high/low quality).

Self-improvement often relies on optimizing parameters θ based on this verification:

$$\max_{\theta} \mathbb{E}_{(x,y)\sim P(x,y|\theta)}[V(x, y)]$$

or using V implicitly as a reward signal R in reinforcement learning. The core principle is: **Effective self-improvement is contingent upon the existence of an efficient and reliable verification mechanism** V. If the complexity of verification, Complexity(V), is low (e.g., polynomial time), then iterative improvement guided by V becomes practical. As the real world is inherently multimodal, MLLMs could potentially leverage environmental feedback or cross-modal consistency checks as powerful verification signals (Ahn et al., 2024b), potentially making V more robust compared to text-only domains.

C.5 Generalization

Current self-improvement pipelines often focus on specific tasks τ (e.g., reducing hallucinations, improving reasoning on benchmarks) and may exhibit diminishing returns after a finite number k of iterations:

$$\theta_{i+1} = \text{Improve}(\theta_i, \mathcal{D}_i, \tau), \quad i = 0, \ldots, k-1$$

where $\mathcal{D}_i$ is the data used/generated at iteration i. Performance P might plateau, i.e., $P(\theta_k, \tau) \approx P(\theta_{k+1}, \tau)$.

A major future direction is developing a *general* MLLM self-improvement framework capable of recursive enhancement across a universal set of tasks $\mathcal{T}_{\text{univ}}$ without plateauing. The idealized goal is a process:

$$M_{i+1} = \text{SelfImprove}(M_i, \mathcal{T}_{\text{univ}}, \text{WorldKnowledge})$$

such that the model's capabilities $C(M_i)$ monotonically increase across $\mathcal{T}_{\text{univ}}$ as $i \to \infty$:

$$\forall \tau \in \mathcal{T}_{\text{univ}}, \lim_{i\to\infty} P(M_i, \tau) = \text{OptimalPerformance}(\tau)$$

This requires mechanisms that not only refine parameters but potentially adapt the model's architecture, learning algorithms, and knowledge representation recursively, moving beyond narrow, task-specific improvement loops towards universal, open-ended capability growth.

C.6 Scalability

Although we have collected many models and frameworks in this survey, we found that many of these methods are normally conducted on a very small scale. Therefore, the performance gain is not as significant as in many other model developments that simply scale things up. It would be more practical and impactful for the real world deployment if the approaches had satisfactory scalability which would address the data shortage problem and therefore allow the model development to be further scaled up.

C.7 Autonomy

Although current self-improvement MLLM frameworks can reduce the human workload from a data generation and verification perspective, human involvement is still required in many other areas, such as making idea proposals, codebase development, and conducting experiments. To overcome this bottleneck and achieve fully autonomous self-improvement requires agentic-level autonomy. This level of autonomy has the potential to accelerate self-improvement progress by orders of magnitude. Meanwhile, the R&D skills themselves could be further boosted by the improved base MLLMs, such as through better multimodal understanding of the environment. This mutually beneficial self-improvement paradigm can increase effectiveness by removing bottlenecks, eliminating blind spots, and raising the upper bound. Appropriate guardrails designs can become more meaningful in those more autonomous self-improvement approach to mitigate potential risk.

D Related Surveys

There are several surveys on self-improvement/evolution (Tao et al., 2024) and multimodal large language models (Yin et al., 2024). However, to the best of our knowledge, no existing survey specifically addresses self-improvement in multimodal large language models. To fill this gap, we have collected related papers and systematically constructed this survey.

More recent works adjacent to our scope include (i) surveys on reinforcement learning for MLLMs (Sun et al., 2025b; Wu et al., 2025b), which is a specific domain of self-improvement, and (ii) surveys on self-evolving agents that focus on agents rather than MLLMs (Gao et al., 2025; Fang et al., 2025).

Other surveys focus on topics such as self-supervised learning (Gui et al., 2024), self-training (Amini et al., 2024), synthetic data (Bauer et al., 2024), or data augmentation (Feng et al., 2021), which are loosely connected at a high level.

Our survey is the first to focus specifically on self-improvement in MLLMs, collecting a broad range of methods for automating MLLM improvement with less human effort. Concretely, we structure the field into a three-stage pipeline: data collection, data organization, and model optimization to analyze different techniques used in each module. We also formulate unified levels of autonomy for self-improvement in MLLMs to guide future development toward more effective self-improvement methodologies.