A Survey on Post-training of Multimodal Large Language Models

Haonan Zhang[1,2], Pengpeng Zeng[1*], Libin Cao[1], Wenrui Lai[1], Jinlong Li[4,5], Duo Peng[1], Yi Bin[1], Xuanhan Wang[1], Ji Zhang[3], Jingkuan Song[1], Nicu Sebe[4], Yuchuan Wu[2], Yongbin Li[2], Heng Tao Shen[1], Jieping Ye[2]

[1]School of Computer Science and Technology, Tongji University, Shanghai, China.
[2]Qwen-Character Team, Alibaba Group.
[3]School of Computing and Artificial Intelligence, Southwest Jiaotong University, Sichuan, China.
[4]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy.
[5]Department of Computer Science, ETH Zurich, 8092, Zürich, Switzerland.

*Corresponding author(s). E-mail(s): is.pengpengzeng@gmail.com;

Abstract

The technical evolution of Multimodal Large Language Models (MLLMs) has profoundly influenced the AI community, reshaping how intelligent systems understand, reason, and interact in both the digital and physical worlds. Through large-scale multimodal pretraining, MLLMs lay the foundation for general perception and alignment, while it remains unclear how such models can be transformed into human-grounded behaviors. Multimodal Post-training (MMPoT), which refines pretrained MLLMs aligned with human intent and real-world task demands, has become a decisive paradigm for advancing multimodal general intelligence. To this end, both academia and industry have devoted increasing efforts to this direction, pushing the boundaries of multimodal general intelligence at an unprecedented pace. In this survey, we systematically review existing MMPoT research through a behavior-shaping perspective. Guided by this view, we first introduce a unified framework that organizes current MLLMs post-training methods into five foundational behaviors: instruction following, preference calibration, reason enhancement, domain adaptation, and scalable learning. Moreover, we revisit benchmarks and evaluation protocols to critically assess the capabilities and limitations of post-trained MLLMs. To conclude, we provide a synthesized outlook on broader open questions, and highlight future research directions toward more general and reliable multimodal capabilities. An updated paper list is available at: https://github.com/zchoi/Awesome-post-training-for-MLLMs.

Keywords: Multimodal Large Language Models, Post-training, Human Alignment.

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) have emerged as a promising paradigm for advancing general-purpose artificial intelligence (Liu et al. 2023; Dai et al. 2023). By integrating vision, language, and other modalities within a unified framework, these models enable

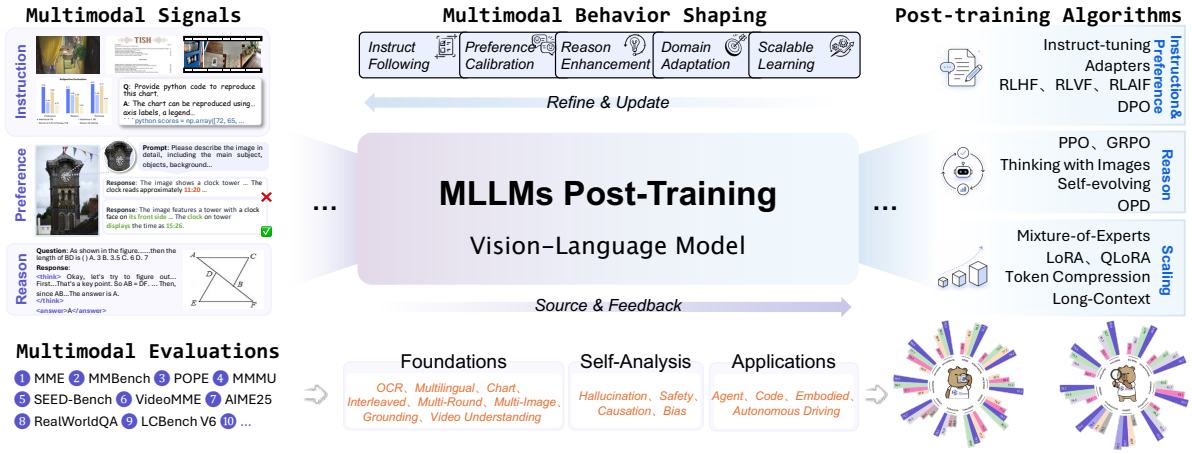

Fig. 1: **Overview of multimodal behavior shaping for MLLMs post-training.** Post-training algorithms can be viewed as behavior-shaping mechanisms that steer pretrained MLLMs toward desired behaviors, while multimodal data and benchmarks provide learning signals and evaluative feedback for iterative refinement.

a broad spectrum of applications, driving practical advances in vision-language assistants (Zhang et al. 2025b; Liu et al. 2025b), autonomous driving (Jia et al. 2023; Xu et al. 2024b), and embodied intelligence (Driess et al. 2023; Kim et al. 2024).

With Large Language Models (LLMs) (Bai et al. 2023; Touvron et al. 2023) serving as their core reasoning and generation engine, MLLMs extend language-centric capabilities to multimodal inputs through multimodal pretraining (Bai et al. 2023; Dai et al. 2023; Liu et al. 2023). This constitutes the prevailing paradigm for building modern MLLMs, in which a powerful LLM backbone is coupled with modality encoders and cross-modal connectors, and trained on large-scale corpora spanning images, videos, text, audio, and other modalities. By optimizing objectives such as next-token prediction, multimodal pretraining establishes basic cross-modal representations and alignment, supporting progress in various downstream tasks such as visual question answering (Khayatkhoei et al. 2025; Marino et al. 2019), captioning (Sarto et al. 2025; Chen et al. 2025a), document understanding (Wang et al. 2025a; Zhang et al. 2025c), and visual reasoning (Dong et al. 2025; Yang et al. 2026).

Despite these advances, most pretrained MLLMs primarily learn statistical patterns from large-scale data. Such pretraining alone does not ensure that model behaviors are aligned with human needs, ethical principles, or real-world requirements. To bridge this gap, recent research has increasingly shifted toward ***multimodal post-training*** (MMPoT), a stage after pretraining that transforms the learned capabilities of MLLMs into reliable and controllable behaviors. This emerging direction has attracted rapidly growing research interest, spanning diverse application scenarios (Zhang et al. 2025a; Saab et al. 2024; Zhang et al. 2026b) and optimization goals (Wang et al. 2022; Li et al. 2026a). Prevalent multimodal post-training approaches typically follow a standard paradigm, *i.e.*, multimodal supervised fine-tuning (SFT). It tunes pretrained MLLMs on instruction-formatted data (*i.e.*, multimodal inputs paired with user instructions and target responses), learning to respond to multimodal instructions and interact with users in a task-oriented manner. This simple yet influential scheme has driven the success of early MLLMs post-training methods (Zhu et al. 2024; Dai et al. 2023; Liu et al. 2023, 2024a). However, SFT mainly relies on imitation learning, its training objective cannot fully judge whether a response is preferred, faithful to evidence, safe, or robust under complex reasoning. Reinforcement Learning in MLLMs has made great progress in closing this gap between model's output and human intention. It learns from feedback lists or preference pairs, encouraging the model to optimize model behavior that better matches human or AI judgments.

This evolution spans from early reinforcement-based post-training approaches such as proximal policy optimization (Stephan et al. 2024; Luo et al. 2024a) to more scalable strategies like direct preference optimization (Rafailov et al. 2023; Shao et al. 2024; Yu et al. 2026). Inspired by the success of reasoning-oriented models such as o1 (Jaech et al. 2024) and DeepSeek-R1 (Guo et al. 2025b) in the language domain, increasing research attention has been devoted to eliciting and enhancing the reasoning capabilities of MLLMs. Compared to early reinforcement learning methods, this scheme obviates the need for human-labeled trajectories via carefully designed reward mechanisms (*e.g.,* ORM and PRM) and RL strategies (*e.g.,* GRPO (Guo et al. 2025b)), achieving impressive performance in complex problems.

Prior surveys have examined specific aspects of MLLMs post-training (Yu et al. 2025b; Zhou et al. 2025a; Zhang et al. 2026c), yet a unified perspective remains lacking. An application-oriented survey (Yu et al. 2025b) focuses on human alignment in specific domains, whereas technique-oriented surveys (Zhou et al. 2025a; Zhang et al. 2026c) concentrate on individual methodological paradigms. Our analysis reveals that existing MLLMs post-training methods differ substantially in their feedback sources, optimization objectives, updated model components, and evaluation protocols. Together, these variations expose a fragmented methodological landscape: methods may pursue similar behavioral goals using distinct feedback signals or apply similar training recipes to optimize different capabilities. These observations motivate the following fundamental question:

> *What common process underlies MLLMs post-training, and how does it steer pretrained MLLMs toward desired multimodal behaviors?*

As the scope of MLLMs post-training continues to broaden, this survey addresses this question by advocating a behavior-shaping perspective: instead of treating multimodal post-training as isolated recipes, we examine them through the full behavior-shaping loop as shown in Fig. 1, *i.e.*, **1) *Refine & Update***: what algorithms are applied and which capabilities it aims to improve. **2) *Source & Feedback***: where supervision comes from and how the resulting behavior is evaluated. By synthesizing these perspectives, we seek to delineate the evolving methodological landscape of MLLMs post-training and shed light on the emerging frontiers that may shape future research.

The main contributions are summarized as follows:

- We conceptualize MLLMs post-training as a process of multimodal behavior shaping, offering a unified perspective on how pretrained MLLMs acquire reliable and versatile behavioral capabilities.
- We present a taxonomy that organizes MLLMs post-training methods into five major families: multimodal instruction following, preference calibration, reasoning enhancement, domain adaptation, and scalable training.
- We systematically examine the datasets and evaluations used in MLLMs post-training, revealing how benchmarks and metrics define and measure desirable multimodal behavior.
- We identify key challenges and outline promising research directions, offering a roadmap for advancing MLLMs post-training toward dependable multimodal intelligence.

Scope. This survey focuses on post-training for Multimodal Large Language Models (MLLMs), examining how pretrained models are adapted for human interaction and task-oriented use. Our scope centers on multimodal understanding over images, videos, audio, and text, while specialized sensor modalities such as infrared and LiDAR are beyond the main focus. Some methods span both pre-training and post-training, we include them when their post-training components are relevant to our analysis. Downstream domains, such as autonomous driving and healthcare, are discussed only to illustrate specific design choices; the primary focus remains on underlying post-training mechanisms.

Roadmap. The remainder of this survey is organized as follows. Sec. 2 introduces the basic definition of MLLMs post-training within our behavior-shaping framework. Moreover, Sec. 3–7 review the five major families of post-training methods. Further, Sec. 8 surveys the mainstream benchmarks and evaluation protocols used to assess post-trained MLLMs. Finally, Sec. 9 discusses the key challenges facing MLLMs post-training and outlines future research directions.

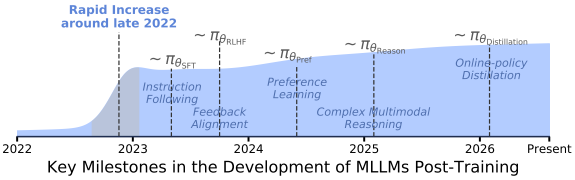

Fig. 2: **Key milestones of MLLMs post-training.** It has rapidly become a central mechanism for endowing pretrained MLLMs with the ability to exhibit more aligned and reliable behaviors, marking significant progress in multimodal intelligence.

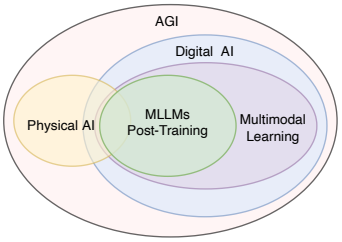

Fig. 3: **A Venn diagram showing the interrelationships among key AI fields.** MLLMs post-training connects multimodal learning with digital AI and physical AI, representing a core step in the progression towards Artificial General Intelligence (AGI).

2 Overview of Post-training for MLLMs

This section provides a brief introduction of MLLMs post-training, covering its basic definition (Sec. 2.1), motivation (Sec. 2.2), and position within key AI fields (Sec. 2.3).

2.1 What Is MLLMs post-training?

MLLMs post-training refers to targeted adaptation and alignment procedures applied to pretrained MLLMs. Unlike pre-training, which builds broad multimodal representations, post-training uses task- and instruction-level supervision to align behavior with human intent, strengthen cross-modal reasoning, and improve downstream reliability.

Formally, let x_m denote multimodal inputs (*e.g.,* images, videos, audio, documents, screenshots, or environment observations), x_t denote textual instructions and dialogue history, and y denote outputs such as responses, tool calls, grounding predictions, or actions. Post-training optimizes:

$$\pi_\theta(y \mid x_m, x_t). \tag{1}$$

Given supervision or feedback f (*e.g.,* demonstrations, preferences, rewards, teacher outputs, or grounding annotations), the general objective is:

$$\begin{aligned} \max_{\theta} \quad & \mathbb{E}_{(x_m, x_t, f)\sim\mathcal{D}}\left[U(\pi_\theta; x_m, x_t, f)\right] \\ s.t. \quad & C(\pi_\theta) \le \tau. \end{aligned} \tag{2}$$

where U captures helpfulness, faithfulness, reasoning success, safety, or task reward, while C encodes constraints such as latency, memory, annotation cost, privacy risk, and safety violations. This abstraction shows that post-training methods differ mainly in their supervision, feedback, and constraint design, which determine how models acquire, generalize, and maintain desired behaviors across diverse tasks, domains, and real-world deployment settings.

2.2 Why MLLMs Post-training: Progressive Behavior Shaping

From a behavior-shaping perspective, the evolution of MLLMs post-training reflects a gradual shift from capability activation to policy refinement, as illustrated in Fig. 2. Early supervised fine-tuning (Dai et al. 2023; Liu et al. 2023) teaches pretrained multimodal models to follow vision-language instructions, activating latent perceptual knowledge for task-oriented behavior. Subsequent methods based on feedback alignment (Sun et al. 2024a; Yu et al. 2024) and preference learning (Ouali et al. 2024) further refine model policies toward human-preferred responses, mitigating hallucinations and improving reliability. More recent advances in complex multimodal reasoning (Huang et al. 2025) and online policy distillation (Hou et al. 2026; Li et al. 2026b) extend this process from response-level correction to higher-level reasoning and decision making, enabling more consistent reasoning, interaction, and adaptation. Overall, post-training has emerged as a central mechanism for transforming pretrained MLLMs with broad multimodal representations into systems that exhibit aligned, reliable, and controllable behaviors.

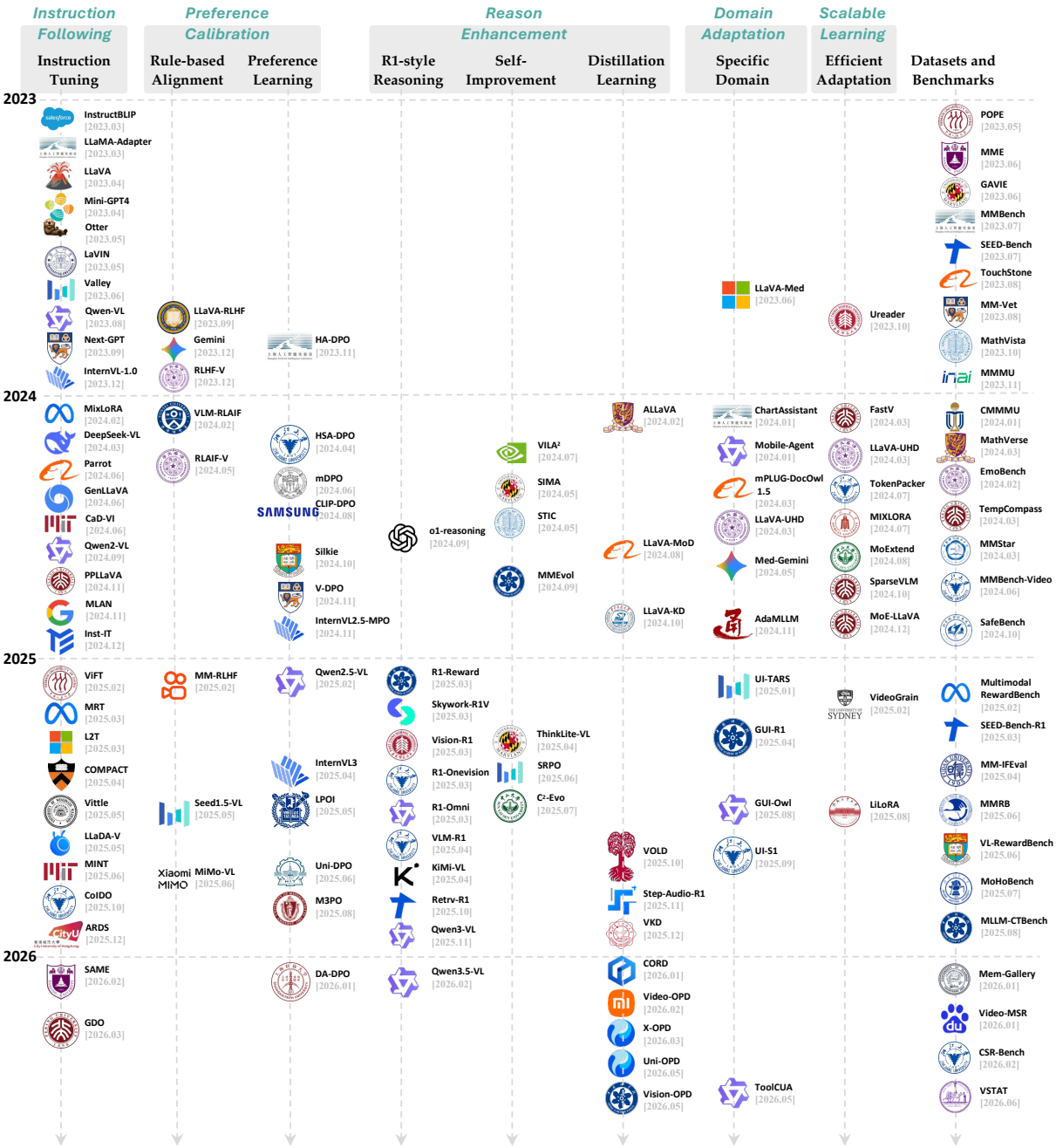

Fig. 4: A timeline of MLLMs post-training research.

2.3 Post-training MLLMs as the Next Frontier

The rapid progress of foundation models has made multimodal intelligence a key pathway toward AGI. However, most MLLMs remain primarily digital systems: they can perceive and reason over multimodal inputs, but still lack reliable interaction, tool use, and action in open environments. Post-training is therefore a critical bridge. By leveraging instruction data, feedback signals, preference objectives, reasoning traces, and agentic experience, it equips MLLMs with aligned, goal-directed, and decision-making capabilities. In this

Table 1: **Representative multimodal instruction following methods in MMPoT.** PD = pre-training data, ID = instruction-tuning data. I, T, V, A, and 3D mean image, text, video, audio, and 3D point cloud modalities, respectively. *: For methods with multiple visual encoders or backbone configurations, we report the best-performing or most representative variant.

Method	Base LLM	Visual Encoder	# Params.	Modality	Training Data Scale	Source
LLaVA (Liu et al. 2023)	Vicuna	CLIP ViT-L/14	~7.3B / ~13.3B	I+T→T	595K PD + 158K ID	Paper / GitHub / Project
MiniGPT-4 (Zhu et al. 2024)	Vicuna-7B / 13B	EVA-CLIP ViT-G/14	~7B / 13B	I+T→T	~5M PD+ 3.5K ID	Paper / GitHub / Project
InstructBLIP (Dai et al. 2023)	Flan-T5 / Vicuna	EVA-CLIP ViT-g	3B–13B variants	I+T→T	129M PD	Paper / GitHub
mPLUG-Owl (Ye et al. 2023b)	LLaMA-7B	CLIP ViT-L/14	~7B	I+T→T	392k ID	Paper / GitHub
LLaMA-Adapter V2 (Gao et al. 2023)	LLaMA-7B / 13B	CLIP ViT	7B / 13B	I+T→T	567K PD + 132K ID	Paper / GitHub
LLaVA-1.5 (Liu et al. 2024a)	Vicuna-7B / 13B	CLIP ViT-L/14-336	~7B / ~13B	I+T→T	558K PD + 665K ID	Paper / GitHub
CogVLM (Wang et al. 2024f)	Vicuna-7B / LLaMA	EVA-CLIP ViT	~7B	I+T→T	1.54B PD	Paper / GitHub
LLaVA-NeXT (Liu et al. 2024b)	Vicuna / Mistral / Yi	CLIP ViT-L/14-336	7B / 13B / 34B variants	I+T→T	558K PD + 760K ID	GitHub / Project
Qwen2-VL-Instruct (Wang et al. 2024e)	Qwen2	ViT	2B / 7B / 72B	I+T+V→T	-	Paper / GitHub
LLaVA-OneVision (Li et al. 2024a)	Qwen2	SigLIP	0.5B / 7B / 72B	I+T+V→T	~4.8M PD + 4.8M ID	Paper / Project
InternVL2.5 (Chen et al. 2024d)	InternLM2.5 / Qwen variants	InternViT	1B–78B variants	I+T+V→T	~120B tokens + 16.3M ID	Paper / GitHub / Project
Qwen2.5-VL-Instruct	Qwen2.5	ViT	3B / 7B / 72B	I+T+V→T	~4.1T tokens + ~2M ID	Paper / GitHub
VILA (Lin et al. 2024)	LLaMA-2 / Vicuna variants	CLIP ViT-L/14	2.7B / 7B / 13B / 40B	I+T+V→T	~50M PD	Paper / GitHub
MobileVLM (Chu et al. 2023)	MobileLLaMA-1.4B / 2.7B	CLIP ViT-L/14	1.7B / 3B	I+T→T	558K PD + 665K ID	Paper / GitHub
DeepSeek-VL (Lu et al. 2024a)	DeepSeek-LLM	SigLIP+SAM	1.3B / 7B	I+T→T	-	Paper / GitHub
Cambrian-1 (Tong et al. 2024)	LLaMA-3 / Vicuna variants	ViT / ConvNeXT series	8B / 13B / 34B variants	I+T→T	~7M ID	Paper / GitHub / Project
MiniCPM-V 2.6 (Yao et al. 2024)	Qwen2-7B	SigLIP-400M	~8B	I+T+V→T	570M PD + ~8.8M ID	Paper / GitHub / Model
Molmo (Deitke et al. 2025)	OLMo	ViT	1B / 7B / 72B variants	I+T→T	712K PD + ~31.6M ID	Paper / GitHub / Project
Pixtral-12B (Agrawal et al. 2024)	Mistral-based decoder	Pixtral-ViT	~12B	I+T→T	-	Paper / Project / Model
Otter (Li et al. 2025b)	OpenFlamingo / LLaMA-style LM	CLIP ViT-L/14	~9B	I+T+V→T	2.8M ID	Paper / GitHub
LaVIN (Luo et al. 2023a)	LLaMA-7B / 13B	CLIP ViT	7B / 13B	I+T→T	210K ID	Paper / GitHub
Valley (Luo et al. 2026)	Stable-Vicuna	CLIP ViT-L/14	7B	I+T+V→T	1.297M PD + 234K ID	Paper / GitHub
NExT-GPT (Wu et al. 2023)	Vicuna	ImageBind	7B	I+T+V+A→I+T+V+A	15K PD + 5K ID	Paper / GitHub / Project
InternVL-1.0 (Chen et al. 2024e)	InternLM / Vicuna variants	InternViT	6B–26B variants	I+T→T	6.01B PD + ~4M ID	Paper / GitHub
GenLLaVA (Hernandez et al. 2024)	Mistral-7B	SigLIP	7B	I+T→I+T	558K PD + ~2.08M ID	Paper
MLAN (Tu et al. 2025)	LLaVA-style MLLMs	CLIP ViT-L/14	7B variants	I+T→T	558K PD + 186K ID	Paper
Inst-IT (Peng et al. 2026)	Qwen2-7B*	SigLIP-SO400*	~7B	I+T+V→T	558K PD + 243K ID	Paper / GitHub
Vittle (Oh et al. 2026)	LLaVA-style MLLMs	CLIP ViT-L/14-336px	Backbone-dependent	I+T→T	558K PD + 665K ID	Paper
LLaDA-V (You et al. 2026)	LLaDA	SigLIP 2-so400m-patch14-384	8B	I+T→T	558K PD + 12M ID	Paper / Project
LLaVA-NeXT-Interleave (Li et al. 2024a)	Vicuna / Mistral / Qwen variants	CLIP ViT-L/14-336	7B / 13B	I+T+V→T	1.18M interleaved ID	Paper / Project
Mantis-Idefics2 (Jiang et al. 2024)	LLaVA / Idefics-style backbones	SiGLIP	7B / 8B	I+T→T	143M PD + 721K ID	Paper / Project
ImageBind-LLM (Han et al. 2023)	LLaMA	-	7B	I+T+V+A+3D→T	940M PD + 205.5K ID	Paper / GitHub
ShareGPT4V (Chen et al. 2024a)	Vicuna-7B	CLIP ViT-L/14	7B	I+T→T	1.2M PD + 665K ID	Paper / GitHub / Project

sense, post-training is not merely a downstream optimization stage, but a bridge linking multimodal learning, digital AI, and physical AI, as illustrated in Fig. 3. It moves MLLMs beyond passive understanding toward robust interaction, grounded reasoning, and safety-aware action. A chronological overview of MLLMs post-training research is provided in Fig. 4.

3 MMPoT of Instruction Following

Multimodal instruction tuning is an early and foundational post-training technique for enabling instruction-following ability in MLLMs. It adapts pretrained multimodal models using vision-language instruction-response pairs, thereby ***transforming raw multimodal representations into task-oriented distributions for instruction-following behavior***. As the first post-training step, it establishes the interface between user intent, multimodal input, and model response. This simple yet effective paradigm underpins early MLLMs such as InstructBLIP (Dai et al. 2023), LLaVA (Liu et al. 2023), and their variants. Representative methods are summarized in Tab. 1.

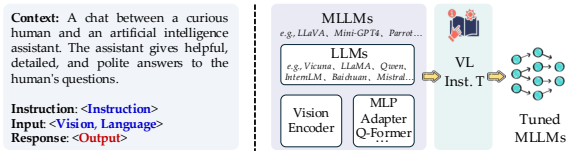

Fig. 5: **A basic view of behavior shaping for instruction following. Left**: A prompt template that organizes multimodal instruction data. We take the built-in Jinja template from HuggingFace `llava-hf/llava-1.5-7b-hf` as an example, which works for both training and inference. **Right**: The instruction-tuning paradigm used to train MLLMs. VL Inst. T: Vision-Language Instruction Tuning.

3.1 Preliminary

Given a pretrained MLLM with parameters θ_0, multimodal instruction tuning adapts the model using instruction-formatted visual-language examples. Each training instance can be represented as:

$$(x_m, x_u, y) \in \mathcal{D}_{\text{inst}}, \tag{3}$$

where x_m denotes the multimodal input (*e.g.*, image, video, and speech), x_u denotes the user instruction, and $y = (y_1, \ldots, y_T)$ is the target

response. The model generates the response conditioned on both multimodal input and instruction:

$$p_\theta(y \mid x_m, x_u) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, x_m, x_u). \quad (4)$$

The instruction-tuned model is obtained by minimizing the supervised next-token prediction loss:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(x_m, x_u, y)\sim\mathcal{D}_{\text{inst}}} \left[-\sum_{t=1}^{T} \log p_\theta \left(y_t \mid y_{<t}, x_m, x_u \right) \right]. \quad (5)$$

Depending on the training recipe, θ may include only the connector, selected adapter modules, the LLM backbone, the modality encoder, or all trainable components. A basic diagram of the instruction tuning pipline is illustrated in Fig. 5.

3.2 Visual Instruction Tuning

Visual instruction tuning adapts pretrained multimodal models to follow natural-language instructions grounded in images or videos. By learning from vision-language instruction-response pairs, it aligns user intent, visual context, and model responses within a unified generative interface. This process activates latent multimodal knowledge as task-oriented instruction-following behavior, enabling MLLMs to address diverse visual tasks with reduced reliance on task-specific heads.

•**Visual Adaptation to LLMs.** Early visual instruction tuning primarily focused on connecting pretrained visual encoders with LLMs to enable image-grounded instruction following. LLaVA (Liu et al. 2023) introduced a widely adopted two-stage recipe that first aligns CLIP (Radford et al. 2021) visual features with the embedding space of Vicuna (Chiang et al. 2023), and then fine-tunes the model on visual instruction data for open-ended multimodal dialogue. Subsequent works explored lighter or more modular variants: MiniGPT-4 (Zhu et al. 2024) uses a lightweight projection module to connect vision features to a mostly frozen LLM, mPLUG-Owl (Ye et al. 2023b) adopts a modular image-conditioned design, MultiModal-GPT (Gong et al. 2023) studies parameter-efficient tuning, and InstructBLIP (Dai et al. 2023) uses an instruction-aware Q-Former (Li et al. 2022). Together, these works established the standard paradigm of bridging vision and language backbones, aligning their representations, and adapting the resulting model with instruction data.

•**Instruction-Aware Visual Injection.** Building on these foundations, subsequent studies improve visual instruction tuning through instruction-aware feature selection, stronger training recipes, and efficient adaptation. LLaVA-1.5 (Liu et al. 2024a) improves the reproducibility and effectiveness of this recipe through a refined data mixture, higher image resolution, and optimized training configurations. LaVIN (Luo et al. 2023a) introduces lightweight adaptation modules for efficient vision-language alignment, while LLaMA-Adapter V2 (Gao et al. 2023) demonstrates that adapter-based tuning can activate multimodal instruction-following capabilities at limited training cost. Together, these works show that effective visual instruction tuning depends not only on cross-modal alignment, but also on how visual information is selected, integrated, and optimized during adaptation.

•**Fine-grained Instruction Tuning.** Another direction extends visual instruction tuning from holistic image understanding to fine-grained grounding and region-aware interaction. GPT4RoI (Zhang et al. 2025e) introduces region-level visual information into instruction tuning. Shikra (Chen et al. 2023) and Ferret (You et al. 2024; Zhang et al. 2024a) combine referring expressions, grounding, and dialogue to support user-specified object or region understanding, while Kosmos-2 (Peng et al. 2024) links language generation with visual regions.

•**Instruction Tuning Beyond Images.** Visual instruction tuning has also been extended beyond single-image understanding to more complex inputs and scenarios. For temporal and multi-image reasoning, Video-LLaMA (Zhang et al. 2023), VideoChat, Video-ChatGPT (Li et al. 2025c), and Valley (Luo et al. 2023b) adapt instruction tuning to videos or multiple images, enabling event understanding, cross-image comparison, and visual information synthesis. LLaVA-OneVision (Li et al. 2024a), Qwen2.5-VL (Bai et al. 2025), and InternVL3 (Zhu et al. 2025) represent this shift from image-chat post-training to full-stack multimodal post-training. Their recipes commonly combine high-quality

Table 2: **Representative multimodal preference calibration methods in MMPoT.** Here, I, V, A, and T denote image, video, audio, and text, respectively. RS: reject sampling, BoN: best-of-N, CFG: Classifier-Free Guidance.

Method	Feedback Granularity	#Params.	Modality	Optimization Paradigm	Source
Multimodal RLHF					
LLaVA-RLHF (Sun et al. 2024a)	Pair-level	7B / 13B	I+T→T	PPO	Paper / GitHub / Project
RLHF-V (Yu et al. 2024)	Span-level	13B	I+T→T	DDPO	Paper / GitHub / Project
VisualPRM (Wang et al. 2025b)	Step-level	8B	I+T→T	PRM + BoN	Paper / Project
Gemini (Team et al. 2023)	Mixed-level	–	I+V+A+T→T	SFT + RLHF	Paper / Project
Seed1.5-VL (Guo et al. 2025a)	Mixed-level	20B	I/V+T→T	RLHF + RLVR + RS	Paper / GitHub / Project
MiMo-VL (Yue et al. 2025b)	Mixed-level	7B	I/V+T→T	MORL	Paper / GitHub
Multimodal RLAIF					
VLM-RLAIF (Ahn et al. 2024)	Pair-level	7B	V+T→T	Context-aware RM + PPO	Paper / GitHub / Project
RLAIF-V (Yu et al. 2025a)	Mixed-level	7B / 12B	I+T→T	Iterative DPO + BoN	Paper / GitHub
Oracle-RLAIF (Shi et al. 2025)	Listwise-level	7B	V+T→T	Oracle ranker + rank-aware GRPO	Paper
Multimodal DPO					
Video-DPO (Zhang et al. 2025d)	Response-level	7B	V+T→T	DPO	Paper / GitHub
HA-DPO (Zhao et al. 2023)	Response-level	7B / 13B	I+T→T	DPO	Paper / GitHub / Project
V-DPO (Xie et al. 2024)	Response-level	7B	I+T→T	CFG + DPO	Paper / GitHub
CLIP-DPO (Ouali et al. 2024)	Response-level	1.7B / 3B / 7B	I+T→T	DPO + CLIP-ranked preferences	Paper
CHAIR-DPO (Compagnoni et al. 2025)	Object-level	7B / 8B	I+T→T	DPO + CHAIR-based preferences	Paper / GitHub
mDPO (Wang et al. 2024b)	Response-level	3B / 7B	I+T→T	conditional and anchored DPO	Paper / GitHub / Project
MoD-DPO (Chaubey et al. 2026)	Modality-level	–	A+V+T→T	DPO	Paper
OmniDPO (Chen et al. 2026)	Modality-level	–	A+V+T→T	DPO	Paper
DA-DPO (Qiu et al. 2026)	Pair-level	7B / 13B	I+T→T	DPO	Paper / GitHub / Project
LPOI (Zadeh et al. 2025)	Listwise-level	–	I+T→T	Anchored DPO	Paper / GitHub / Project

instruction data, higher-resolution visual tokens, task-specific mixtures, and test-time strategies.

3.3 Instruction Data Mixtures

Instruction data mixtures refer to post-training datasets that combine instruction-response examples from multiple tasks, domains, modalities, and supervision sources.

•**Text-Multimodal Balance.** Text-only user-assistant conversations are often mixed with multimodal instruction data to preserve general language and conversational capabilities while extending instruction following to visual and other non-textual inputs. LaVIN (Luo et al. 2023a) constructs minibatches by sampling both language-only and multimodal examples, whereas MultiInstruct (Xu et al. 2023) explores different strategies for combining single-modal and multimodal instructions, including mixed instruction tuning.

•**Multimodal Data Composition.** Beyond balancing text-only and multimodal examples, the internal composition of multimodal data is equally important. LLaVA-1.5 (Liu et al. 2024a) shows that compact, carefully curated mixtures of captioning, VQA, and instruction data can yield strong open-source baselines. ShareGPT4V (Chen et al. 2024a) highlights the value of high-quality captions, while Cambrian-1 (Tong et al. 2024) examines how data composition interacts with

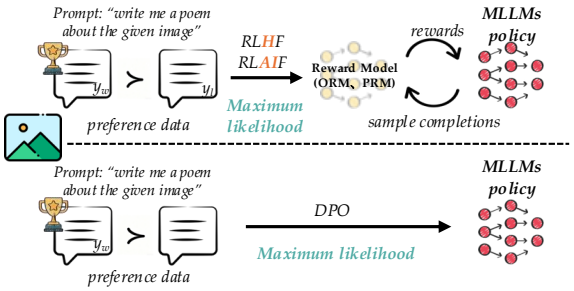

Fig. 6: **A basic view of behavior shaping for preference calibration. Top**: A pipeline of reinforcement learning from human/AI feedback (RLHF and RLAIF). **Bottom**: A pipeline of direct preference optimization (DPO).

model architecture, image resolution, and evaluation performance. Overall, effective instruction mixtures depend not only on scale but also on careful source balancing, filtering, deduplication, and coverage analysis to preserve general language abilities while improving multimodal behavior.

4 MMPoT of Preference Calibration

Following instruction tuning, multimodal preference calibration further refines MLLMs using human- or model-generated feedback to satisfy desired behavioral criteria. It uses preference

pairs, comparative judgments, or reward signals to distinguish desirable responses from weaker alternatives, ***thereby shaping preference-calibrated behavior that favors outputs with fewer hallucinations, stronger visual grounding, and closer alignment with user intent.*** Tab. 2 summarizes representative preference calibration methods.

4.1 Multimodal Reinforcement Learning with Human Feedback

Reinforcement Learning with Human Feedback (RLHF) aligns MLLMs with human preferences using feedback on responses grounded in multimodal inputs. It is typically applied after SFT. Human comparisons or ratings are first used to train a reward model. The learned reward then guides policy optimization, encouraging responses that are helpful, safe, and faithful to multimodal evidence. Fig. 6 (top) illustrates the simplified pipeline of the RLHF.

4.1.1 Reward Mechanisms

In RLHF-based alignment, reward modeling is needed to turn preference supervision into an optimization signal for policy improvement. For multimodal tasks, this signal can be defined at different levels. The commonly used reward modeling strategies mainly fall into two categories: Outcome Reward Mechanisms (ORM) and Process Reward Mechanisms (PRM).

•**Outcome Reward Mechanisms.** Outcome reward mechanisms assign a scalar reward to the final response of a model. This reward typically reflects whether the answer is correct, helpful, and consistent with the visual input. ORM is simple and broadly applicable, since it only requires judging the final output. LLaVA-RLHF (Sun et al. 2024a) extends this formulation by augmenting the reward model with factual signals, such as captions and ground-truth options, so that the reward is less dominated by fluent but visually unsupported answers. RLHF-V (Yu et al. 2024) further moves from coarse response ranking to fine-grained correctional feedback, where hallucinated segments are explicitly identified and corrected. However, it provides limited supervision for how the model reaches the answer, making it less effective when intermediate reasoning errors are hidden by a plausible final response.

•**Process Reward Mechanisms.** To address the coarse supervision of ORM, process reward mechanisms evaluate the intermediate steps that lead to the final response. Instead of only scoring the endpoint, PRM assigns feedback to reasoning traces, subgoals, or step-by-step decisions, encouraging the model to follow a grounded and reliable reasoning path. This is useful for complex multimodal tasks, where correct answers often depend on visual grounding, multi-step inference, and faithful use of evidence. VisualPRM (Wang et al. 2025b), supervise intermediate reasoning steps rather than final answer, enabling finer credit assignment. This trajectory shifts reward modeling from holistic preference prediction toward grounded and step-aware evaluation.

4.1.2 Policy Learning of RLHF

In RLHF, an MLLM first generates candidate responses for each multimodal instruction. Human annotators evaluate these responses based on rubric such as helpfulness and correctness, producing pairwise preferences or scalar ratings. These annotations train a reward model r_ϕ to approximate human judgments. The reward may be assigned to the final response by an ORM or to intermediate reasoning steps by a PRM. The policy π_θ is optimized against the learned reward, typically under a KL-divergence constraint that limits deviation from the reference policy:

$$\begin{aligned}\mathcal{L}_{\text{RLHF}} = &- \mathbb{E}_{y\sim\pi_\theta(\cdot|x_t,x_v)}\left[r_\phi(x_t,x_v,y)\right] \\ &+ \beta D_{\text{KL}}\left(\pi_\theta(\cdot \mid x_t,x_v)\|\pi_{\text{ref}}(\cdot \mid x_t,x_v)\right).\end{aligned} \tag{6}$$

where π_θ is the policy being optimized, π_{ref} is the fixed reference policy, usually the SFT model, and (x_t, x_v) denotes the textual instruction and visual input. The response y is sampled from π_θ, and $r_\phi(x_t, x_v, y)$ is the reward assigned by the reward model. The KL term keeps π_θ close to π_{ref}, with β controlling its strength.

This paradigm has inspired broad explorations. As a well-known pioneering method, LLaVA-RLHF (Sun et al. 2024a) adapts KL-regularized RLHF to MLLMs and uses factual rewards to mitigate reward hacking and hallucination. RLHF-V (Yu et al. 2024) instead applies dense preference optimization over corrected response spans, enabling more targeted

and data-efficient updates. At the frontier scale, MM-RLHF (Zhang et al. 2025g) further introduces 120K fine-grained human preference pairs, critique-based reward modeling, and dynamic reward scaling. Recent methods increasingly combine reward sources: Seed1.5-VL (Guo et al. 2025a) pairs RLHF with verifier-based rewards, while MiMo-VL (Xiaomi 2025) applies mixed on-policy RL with diverse rewards to improve visual understanding and multimodal reasoning.

4.2 Multimodal Reinforcement Learning with AI Feedback

Reinforcement Learning from AI Feedback (RLAIF) extends RLHF by supplementing or replacing human judgments with AI-generated feedback. AI evaluators score or rank candidate responses, and the resulting signals guide policy optimization. Unlike purely supervised post-training, RLAIF creates an iterative evaluation-and-refinement loop, offering a scalable approach to multimodal behavior shaping. RLAIF follows the same pipeline as RLHF, as shown in Fig. 6 (top).

4.2.1 RLHF vs. RLAIF

A central limitation of RLHF is the high cost and limited scalability of collecting reliable human preferences. The burden is greater in multimodal settings, where annotators must inspect visual evidence, identify hallucinations, assess safety risks, and evaluate complex reasoning. This process is time-consuming and can produce inconsistent judgments. RLAIF reduces this dependence by employing capable AI models as evaluators or critics to generate preference labels, scalar scores, or textual critiques for MLLM responses.

4.2.2 RLAIF Training Pipeline

Like RLHF, RLAIF commonly uses preference-based optimization but obtains feedback from AI evaluators rather than relying solely on human annotators. VLM-RLAIF (Ahn et al. 2024) uses an AI evaluator to score or compare candidate responses and then optimizes the model toward more helpful and visually faithful outputs. RLAIF-V (Yu et al. 2025a) further targets visual alignment by assessing whether responses are supported by visual evidence, particularly in cases of hallucination or weak grounding. However, reward-model-based alignment can suffer from unstable optimization and reward hacking, while also requiring a separate reward model. To address these limitations, recent methods increasingly adopt direct preference objectives, such as DPO, which optimize pairwise preferences without explicit reward modeling or PPO. The next section reviews this line of work.

4.3 Multimodal Direct Preference Optimization

In this section, we introduce the definition of Multimodal Direct Preference Optimization (DPO) and examine its major variants in MLLMs post-training. The bottom panel of Fig. 6 presents the DPO optimization scheme.

4.3.1 Preliminary

Although RLHF and RLAIF are effective, applying them to multimodal models remains challenging. Training a reliable reward model over high-dimensional vision-language inputs is costly and prone to bias. Direct Preference Optimization (DPO) (Rafailov et al. 2023), originally introduced for preference alignment in NLP, offers a simpler alternative. Rather than training an explicit reward model and then optimizing the policy through reinforcement learning, DPO learns directly from paired preference examples. In MLLMs post-training, let $x = (x_m, x_t)$ denote a multimodal instruction and (y_w, y_l) denote two candidate responses, where y_w is preferred over y_l. DPO then optimizes:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)}\left[\log \sigma\left(\beta\Big(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\right)\right],$$

where π_θ is the policy model, π_{ref} is a reference model, and β controls the strength of the preference optimization. This formulation preserves the core objective of preference-based alignment while avoiding explicit reward modeling and reinforcement learning updates, making it more stable, efficient, and easier to scale to multimodal settings.

4.3.2 The Evolution of Multimodal DPO

•**Response-level Multimodal DPO.** The simplest multimodal extension of DPO conditions preference optimization on images or videos while treating the full response as the target. Examples include Silkie (Li et al. 2023b), which builds VLFeedback with GPT-4V evaluations, video-language DPO methods that use detailed captions as textual proxies for video content (Zhang et al. 2025d), and ISR-DPO (Ahn et al. 2025), which iteratively generates and evaluates responses to ground preferences in informative video regions.
•**Hallucination-aware DPO.** Since hallucination is a central issue in MLLMs, visual-grounding DPO methods construct preference data to favor grounded responses over hallucinated ones. HA-DPO (Zhao et al. 2023) pairs non-hallucinatory responses with hallucinated alternatives, while V-DPO (Xie et al. 2024), CLIP-DPO (Ouali et al. 2024), and CHAIR-DPO (Compagnoni et al. 2025) introduce visual contrast, image-response consistency, or object-aware signals to strengthen visual grounding and reduce hallucination.
•**Modality-conditioned DPO.** A key limitation of multimodal DPO is unconditional preference learning, where models may rely on language-only cues while ignoring visual or auditory evidence. Modality-conditioned methods address this by grounding preferences in multimodal inputs. mDPO (Wang et al. 2024b) adds image-conditioned optimization with a reward anchor, MoD-DPO (Chaubey et al. 2026) separates modality-specific preference signals with modality-aware regularization, and OmniDPO (Chen et al. 2026) uses both textual and multimodal preference pairs to improve audio-video interaction understanding and sensitivity to multimodal evidence.
•**Hard-negative DPO.** Response-level preference labels are often too coarse for multimodal alignment, where errors may occur at specific tokens, objects, regions, or subtle visual distinctions. Hard-negative and fine-grained DPO methods provide more targeted supervision by emphasizing difficult pairs and visually sensitive outputs. DA-DPO (Qiu et al. 2026) reweights examples by pair difficulty, UE-DPO (Zhang et al. 2026a) uses token-level uncertainty to correct visually unsupported tokens, and OPA-DPO (Yang et al. 2025c) constructs more preference data to reduce distribution mismatch with the current model.
•**Listwise Multimodal DPO.** Standard DPO learns from binary chosen–rejected pairs, while listwise multimodal preference optimization ranks multiple candidates to provide richer supervision. LPOI (Zadeh et al. 2025) builds ranked visual preference samples by masking critical objects and interpolating affected regions, forming a progression from limited visibility to complete object evidence.

5 MMPoT of Reason Enhancement

Multimodal preference calibration improves human-preferred responses but remains insufficient for reasoning-intensive tasks such as multi-step inference, mathematical reasoning, and long-horizon decision making. Inspired by o1-style reasoning models (Jaech et al. 2024) and DeepSeek-Math/R1 (Shao et al. 2024; Guo et al. 2025b), recent work extends reward-based reinforcement learning to MLLMs. This section reviews how reasoning-oriented supervision shapes ***grounded, structured, and multi-step inference across modalities***, with representative methods summarized in Tab. 3.

5.1 R1-based Multimodal Reasoning

R1-based multimodal reasoning extends R1-style post-training to multimodal inputs, using verifiable rewards to elicit structured reasoning grounded in multimodal evidence. Its general paradigm is illustrated in Fig. 7 (left).

5.1.1 From LLM-R1 to MLLM-R1

R1-style post-training (Jaech et al. 2024; Shao et al. 2024; Guo et al. 2025b) marks a shift in LLM reasoning optimization: from imitating annotated chain-of-thought (CoT) traces to learning reasoning behaviors through verifiable feedback. MLLM-R1 extends this principle by conditioning the reasoning process on multimodal evidence rather than text alone. Let $x = (x_m, x_t)$ denote a multimodal instruction, where x_m represents the multimodal input and x_t represents the textual instruction. An R1-style MLLM is then trained to

Table 3: **Representative multimodal reason enhancement methods in MMPoT.** Here, I, V, A, and T denote image, video, audio, and text, respectively.

Method	Base Model	# Param.	Modality	SFT	Algorithm	Source
R1-based Multimodal Reasoning						
VLM-R1 (Shen et al. 2025)	Qwen2.5-VL	∼4B / 7B	I+T→T	✗	GRPO	Paper / GitHub
Visual-RFT (Liu et al. 2025c)	Qwen2-VL	2B	I+T→T	✗	GRPO	Paper
Vision-R1 (Huang et al. 2025)	Qwen2.5-VL	7B / 32B / 72B	I+T→T	✓	GRPO	Paper / GitHub
Video-R1 (Feng et al. 2026)	Qwen2.5-VL-Instruct	7B	I/V+T→T	✓	T-GRPO	Paper / GitHub
VisualPRM (Wang et al. 2025b)	InternVL-style MLLMs	8B	I+T→T	✗	–	Paper / Project
VisualThinker-R1-Zero (Zhou et al. 2025b)	Qwen2-VL	2B	I+T→T	✗	GRPO	Paper / GitHub
MM-Eureka (Meng et al. 2025)	Qwen2.5-VL-Instruct	7B / 32B	I+T→T	✗	GRPO	Paper / GitHub
R1-Onevision (Yang et al. 2025b)	Qwen2.5-VL	3B / 7B	I+T→T	✓	GRPO	Paper
R1-Omni (Zhao et al. 2025)	HumanOmni	∼0.5B	A+V+T→T	✓	GRPO	Paper
Retrv-R1 (Zhu et al. 2026)	Qwen2.5-VL	3B / 7B	I/T→I/T	✓	GRPO	Paper
Thinking with Images						
GRIT (Fan et al. 2026)	Qwen2.5-VL / InternVL3	3B / 2B	I+T→T	✗	GRPO-GR	Paper
Point-RFT (Ni et al. 2026)	Qwen2.5-VL	7B	I+T→T	✓	GRPO	Paper
OpenThinkIMG (Su et al. 2025)	Qwen2-VL	2B	I+T→T	✓	V-ToolRL	Paper
VisionReasoner (Liu et al. 2025a)	Qwen2.5-VL	7B	I+T→T	✗	GRPO	Paper
DeepEyes (Zheng et al. 2025)	Qwen2.5-VL	7B	I+T→T	✗	GRPO	Paper / GitHub
VTool-R1 (Wu et al. 2025)	Qwen2.5-VL	3B / 7B / 32B	I+T→T	✗	GRPO	Paper / GitHub
LanteRn (Viveiros et al. 2026)	Qwen2.5-VL-Instruct	3B	I+T→T	✓	Latent-aware GRPO	Paper
Multimodal Self-Evolving						
VIGC (Wang et al. 2024a)	Vicuna+ViT-G/14	–	I+T→T	✗	–	Paper / Project
MindGYM (Xu et al. 2026)	Qwen2.5-VL / InternVL3	7B / 14B / 32B / 38B	I/T→T	✗	–	Paper
SRPO (Wan et al. 2026)	Qwen2.5-VL	7B / 32B	I+T→T	✓	GRPO	Paper
LLaVA-Critic (Xiong et al. 2025)	LLaVA-OneVision	7B / 72B	I+T→T	✗	–	Paper
MM-UPT (Wei et al. 2026a)	Qwen2.5-VL	7B	I+T→T	✓	GRPO	Paper / GitHub
LLaVA-Critic-R1 (Wang et al. 2025c)	Qwen2.5-VL / ThinkLite-VL	7B	I/V+T→T	✗	GRPO	Paper
Multimodal Distillation						
LLaVA-KD (Cai et al. 2025)	Qwen-series	1B / 2B	I+T→T	✗	–	Paper / GitHub
LLAVADI (Xu et al. 2024a)	LLaVA-v1.5 / MobileLLaMA	13B / ∼2.7B	I+T→T	✗	–	Paper
LLaVA-MoD (Shu et al. 2025)	Qwen-1.5	7B / 2B	I+T→T	✗	–	Paper / GitHub
Video-OPD (Li et al. 2026b)	Qwen3-VL-Instruct	32B / 8B	V+T→T	✗	OPD	Paper
X-OPD (Cao et al. 2026)	Qwen3-Omni / Qwen3	∼3B	A+T→T	✗	OPD	Paper
Uni-OPD (Hou et al. 2026)	Qwen3-VL-Instruct	4B	I+T→T	✗	OPD	Paper
Vision-OPD (Yuan et al. 2026)	Qwen3.5-VL	4B / 9B	I+T→T	✗	OPD	Paper
VA-OPD (Liu et al. 2026)	Qwen3-VL	2B / 4B / 8B / 32B	I+T→T	✗	OPD	Paper

generate a structured reasoning response:

$$y = \texttt{<think>}\ r\ \texttt{</think><answer>}\ a\ \texttt{</answer>}, \tag{7}$$

where r is the intermediate reasoning trajectory and a is the final answer. The model defines:

$$\pi_\theta(y \mid x_m, x_t) = \pi_\theta(r, a \mid x_m, x_t). \tag{8}$$

A common training pipeline first performs cold-start supervised tuning on formatted reasoning data and then applies reinforcement learning with verifiable rewards:

$$\mathcal{L}_{\mathrm{R1}} = -\mathbb{E}_{y\sim\pi_\theta(\cdot|x_m,x_t)}\left[\mathcal{A}(x,y)\right] + \lambda\mathcal{R}(\pi_\theta, \pi_{\mathrm{ref}}), \tag{9}$$

where $\mathcal{A}(x, y)$ denotes a reward-derived learning signal, such as outcome reward, group-relative advantage, or verifier score, and $\mathcal{R}(\pi_\theta, \pi_{\mathrm{ref}})$ is an optional regularization term that constrains the updated policy. Below, we outline how to adapt R1-style optimization to multimodal reasoning.

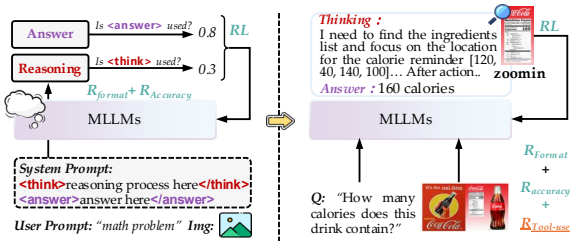

Fig. 7: **A basic view of behavior shaping in R1-style reasoning and thinking with images for reason enhancement.** The former (**Left**) reasons with pure natural language while the latter (**Right**) mixes explicit visual clues (*e.g.*, `bbox, point, and Seg.`) via a chain-of-thought manner.

5.1.2 R1 Training Paradigm for MLLMs

R1-style training for MLLMs broadly follows two paradigms: *R1-Zero* and *R1*. R1-Zero directly applies rule-based reinforcement learning to elicit

reasoning behavior. R1 first establishes a cold-start reasoning policy and then refines it through Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al. 2024), often using Group Relative Policy Optimization (GRPO) (Shao et al. 2024). The key distinction is whether multimodal reasoning is elicited solely by reward signals or bootstrapped from curated reasoning data.

•**R1-Zero-style Training.** R1-Zero-style methods follow DeepSeek-R1-Zero (Guo et al. 2025b) by omitting supervised CoT warm-up and directly optimizing a base or instruction-tuned MLLM with verifiable rewards:

$$\pi_{\theta_0} \xrightarrow[\nabla_\theta \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right]]{\text{RLVR/GRPO}} \pi_{\theta_{\text{zero}}}. \tag{10}$$

The reward function is typically rule-based and outcome-verifiable, without relying on human-written CoT traces. For a multimodal problem $x = (x_t, x_v)$ and a generated response y, the reward can be decomposed into answer correctness and format validity:

$$r(x,y) = \lambda_{\text{ans}} r_{\text{ans}}(x,y) + \lambda_{\text{fmt}} r_{\text{fmt}}(y), \tag{11}$$

where $r_{\text{fmt}}(y)$ checks whether the response follows the required reasoning or answer format, and $r_{\text{ans}}(x,y)$ verifies the final answer:

$$r_{\text{ans}}(x,y) = \begin{cases} 1, & \text{Ans}(y) = a^\star(x), \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Here, $a^\star(x)$ denotes the ground-truth answer for input x, and $\text{Ans}(y)$ extracts the final answer from the model response. Since $x = (x_m, x_t)$ combines a multimodal input with a textual instruction, the answer reward encourages responses that are consistent with multimodal evidence rather than language priors alone.

This reward schema underpins early R1-Zero-style explorations for MLLMs. VisualThinker-R1-Zero (Zhou et al. 2025b) demonstrates that a small vision-language model without SFT warm-up can develop longer reasoning traces and self-reflection through rule-based RL. MM-Eureka (Meng et al. 2025) extends this idea to broader multimodal reasoning and reports visual "aha moments," characterized by longer responses, higher accuracy rewards, and emergent reflection. ThinkLite-VL (Wang et al. 2026) performs pure reinforcement fine-tuning (RFT) without distillation and uses MCTS-guided difficulty filtering to select challenging but solvable visual reasoning samples.

•**R1-style Training.** Though effective, R1-Zero-style training can be unstable when the initial policy lacks reliable reasoning behavior. To reduce noisy exploration and sparse rewards, later methods add a cold-start SFT stage before RL refinement, *i.e.*, a more stable two-stage recipe: cold-start reasoning initialization followed by reinforcement learning:

$$\pi_{\theta_0} \xrightarrow{\text{SFT}(\mathcal{D}_{\text{reason}})} \pi_{\theta_{\text{sft}}} \xrightarrow[\nabla_\theta \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right]]{\text{RLVR/GRPO}} \pi_{\theta_{\text{r1}}}. \tag{13}$$

Along this direction, Vision-R1 (Huang et al. 2025) and Skywork-R1V (Peng et al. 2025) combine cold-start multimodal CoT supervision with GRPO to bootstrap visual reasoning. Vision-R1 uses format and outcome rewards, while Skywork-R1V further incorporates iterative SFT and adaptive-length CoT distillation. R1-Onevision (Yang et al. 2025b) and Retrv-R1 (Zhu et al. 2026) adapt R1-style optimization through task-specific reasoning structures. The former formalizes visual evidence into reasoning-friendly representations, whereas the latter combines retrieval-specific CoT activation with curriculum-guided RL. R1-Omni (Zhao et al. 2025) and VLM-R1 (Shen et al. 2025) emphasize task-specific reward design. R1-Omni jointly rewards emotion prediction and explanation quality, while VLM-R1 targets stable optimization for visual grounding and reasoning.

5.2 Thinking with Images

Textual Chain-of-Thought (CoT) improves multimodal reasoning through stepwise language reasoning, but often treats images as static context. Thinking with images addresses this limitation by making visual information an active reasoning medium, allowing MLLMs to inspect, manipulate, and integrate images as intermediate evidence throughout reasoning, as shown in Fig. 7 (right).

•**Visual Evidence Grounding.** A direct way to think with images is to make each reasoning step traceable to visual evidence. For example, GRIT (Fan et al. 2026) links reasoning steps

to bounding-box regions through a "look–think–look" process. Point-RFT (Ni et al. 2026) uses point-level references for document and chart reasoning, showing that grounded rationales outperform text-only CoT under reinforcement fine-tuning. VisionReasoner (Liu et al. 2025a) unifies detection, segmentation, counting, and VQA within a reasoning-integrated perception framework.

•**Visual Tool Use and Manipulation.** Beyond explicit references, another direction gives models actions over visual inputs. OpenThinkIMG (Su et al. 2025) provides standardized interfaces for detection, segmentation, OCR, cropping, and drawing, and trains adaptive tool invocation through V-ToolRL. DeepEyes (Zheng et al. 2025) and its follow-up variants (Hong et al. 2025) allow VLMs to zoom into and ground relevant visual regions, learning effective inspection strategies through end-to-end RL. VTool-R1 (Wu et al. 2025) interleaves textual reasoning with visual tool operations and uses outcome rewards to elicit strategic tool use for structured chart and table reasoning.

•**Latent Visual Reasoning.** A more implicit direction keeps visual thinking inside hidden states instead of exposing every step as boxes, points, or tool calls. Latent Visual Reasoning (Li et al. 2025a; Xu et al. 2025b) generates latent states that reconstruct query-relevant visual tokens, enabling reasoning in the visual embedding space. LanteRn (Viveiros et al. 2026) combines language with compact latent visual representations, using SFT for grounding and RL for task-level utility. Although primarily designed for visual generation, GoT-R1 (Duan et al. 2025) further suggests that RL can induce semantic-spatial planning beyond fixed templates.

5.3 Self-Evolution for Multimodal Reasoning

Multimodal self-evolving methods improve reasoning through iterative self-improvement. The model generates candidate solutions, evaluates or revises its own reasoning traces, and uses the refined outputs as new training signals. This helps MLLMs strengthen visual-textual reasoning, correct unsupported steps, and improve problem-solving behavior with less reliance on human annotations. The overall workflow of self-evolving is shown in Fig. 8 (left).

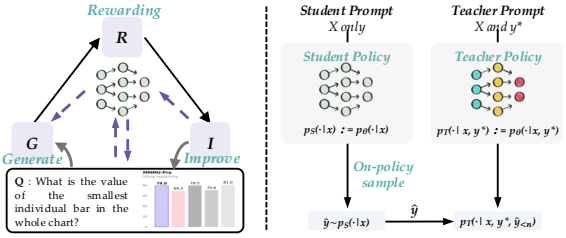

Fig. 8: A basic view of behavior shaping in self-evolving and distillation for reason enhancement.

5.3.1 Self-Generated Data Learning

Self-generated data learning uses model-produced instructions, tasks, and examples to expand training coverage while reducing reliance on curated datasets. VIGC (Wang et al. 2024a) generates visual instructions and iteratively corrects low-quality outputs to reduce hallucinated supervision. MM-Instruct (Liu et al. 2024c) constructs large-scale data by pairing generated instructions with images and producing coherent answers. MindGYM (Xu et al. 2026) organizes self-generated challenging questions into a reasoning curriculum.

5.3.2 Reflection and Critique-based Learning

Beyond generating data, another direction teaches models to inspect and revise their own reasoning. R3V (Cheng et al. 2025) bootstraps positive and negative CoT rationales and trains models to refine flawed reasoning through reflection. SRPO (Wan et al. 2026) introduces reflection-aware reinforcement learning, encouraging concise self-reflection before final answers. LLaVA-Critic and MMEvol (Xiong et al. 2025; Luo et al. 2025b) provide a multimodal critic for evaluating responses across diverse tasks, supporting scalable self-critique. V-Reflection (Zhou et al. 2026) further introduces a "think-then-look" mechanism, prompting models to re-examine visual evidence during reasoning.

Table 4: **Representative multimodal domain adaptation methods in MMPoT.** Doc. = document, EHR = electronic health record, HRV = high-resolution vision, Med. = medical, Inst. = instruction, Traj. = trajectory, RS = remote sensing.

Method	Domain	Input	Adapted Cap.	Source
Mobile-Agent (Wang et al. 2024c)	GUI	Screenshot, Inst.	Perception+Action	Paper / GitHub
GUI-R1 (Luo et al. 2025a)	GUI	Screenshot, Traj.	Reason+Action	Paper / GitHub
mPLUG-DocOwl1.5 (Hu et al. 2024)	Doc.	Doc., Chart	Layout+Reason	Paper / GitHub
LLaVA-UHD (Guo et al. 2024)	HRV	HRV	Perception	Paper / GitHub
Med-Gemini (Saab et al. 2024)	Med.	Med., EHR	Clinical Reason	Paper
AdaMLLM (Cheng et al. 2024a)	Med., Food, RS	Image	Efficiency	Paper

5.3.3 Verifier-guided Self-Improvement

A stronger form of self-evolution closes the training loop through automatic verification. MM-UPT (Wei et al. 2026a) performs unsupervised GRPO using majority-vote self-rewards instead of manually annotated rewards. SelfJudge (Wu et al. 2026) evaluates sampled responses on unlabeled data to support unsupervised reasoning improvement. AGILE (Zeng et al. 2025) converts executable code and environment outcomes into verifiable rewards for visual puzzles, while Jigsaw-R1 (Wang et al. 2025d) derives rule-based rewards from jigsaw tasks. LLaVA-Critic-R1 (Wang et al. 2025c) reformulates critic data as verifiable signals, allowing critique and generation to improve each other.

5.4 Efficient Reasoning

Efficient reasoning seeks to preserve multimodal reasoning capabilities while reducing post-training and deployment costs. Existing methods mainly follow two routes: offline knowledge distillation and on-policy distillation. The main stages of distillation are outlined in Fig. 8 (right).

5.4.1 Knowledge Distillation

Knowledge distillation transfers capabilities from a strong teacher to a smaller or more efficient student. It reduces inference and deployment costs while preserving multimodal reasoning capabilities.

•**Behavior Distillation.** Behavior distillation transfers teacher responses, logits, features, or cross-modal relations to compact students. LLaVA-KD (Cai et al. 2025) compresses large MLLMs through output-distribution and relational distillation. LLAVADI (Xu et al. 2024a) compares different distillation targets and highlights the importance of token-level alignment and feature transfer.

Fig. 9: **A basic view of behavior shaping for domain adaptation.** Ins. T: Instruction-tuning.

•**Feedback Distillation.** Feedback distillation transfers preference or quality signals beyond direct response imitation. Silkie (Li et al. 2023b) distills AI-generated multimodal preferences through DPO to improve helpfulness, visual faithfulness, and safety. LLaVA-MoD (Shu et al. 2025) combines mimic and preference distillation to train a compact MoE-based student.

5.4.2 On-Policy Distillation

On-policy distillation (OPD) combines policy optimization with teacher supervision using rollouts from the current student policy. Compared with offline distillation, OPD better matches supervision to student behavior. It also converts sparse sequence-level feedback into dense token-level guidance. Video-OPD (Li et al. 2026b) applies OPD to temporal video grounding, using token-level feedback and prioritizing informative samples. X-OPD (Cao et al. 2026) extends

Table 5: **Representative multimodal LoRA-based methods in MMPoT.** † indicates that the GPU count was inferred from the official training script rather than explicitly reported in the paper.

Method	Base	Trainable Params.	Rank	# GPUs	Source
LLaVA-MoLE (Chen et al. 2024c)	Vicuna-7B-v1.5	0.3B/7B	32	64*A100	Paper/GitHub
MixLoRA (Shen et al. 2024b)	Vicuna-7B v1.3	∼10M/7B	4	4*A100	Paper/GitHub
MokA (Wei et al. 2026b)	LLaMA2	∼0.1B/7B	4	8†	Paper/GitHub
LiLoRA (Che et al. 2026)	LLaVA-1.5-7B	∼0.25B/7B	128 / 64	4†	Paper/GitHub

Table 6: **Representative multimodal MoE-based methods in MMPoT.** Exp. = experts; Act. = activated experts; FFN = feed-forward network. * For Qwen3-Omni, the expert configuration is inferred from the same-scale Qwen3-30B-A3B setting, as its report identifies the model as 30B-A3B without specifying the detailed expert configuration.

Method	#Exp./Act.	Act. Params	#Params.	Tuned	Source
MoE-LLaVA (Lin et al. 2026)	4/2	3.6B	5.3B	FFN-MoE	Paper/GitHub
MoExtend (Zhong et al. 2024)	-/-	13B	-	New Exp.	Paper/GitHub
Qwen3-Omni (Xu et al. 2025a)	128*/8*	3B	30B	T–T MoE	Paper/GitHub
Qwen3-VL (Team 2025)	-/-	3B / 22B	30B / 235B	VL MoE	Paper/GitHub
MiniMax-01 (MiniMax 2025)	32/-	45.9B	456B	MoE LLM	Paper/GitHub
Seed1.5-VL (Guo et al. 2025a)	-/-	20B	-	MoE LLM	Paper/GitHub

OPD to speech LLMs and uses text-based teachers to align speech and textual reasoning. Uni-OPD (Hou et al. 2026) generalizes the paradigm to LLMs and MLLMs through balanced student data and improved teacher feedback. Vision-OPD (Yuan et al. 2026) performs self-distillation from crop-based to full-image policies, while VA-OPD (Liu et al. 2026) prioritizes visually important tokens for more effective supervision.

6 MMPoT of Domain Adaptation

Domain adaptation specializes pretrained MLLMs for settings with distinct data distributions, task protocols, and reliability requirements, as shown in Fig. 9. Unlike general instruction tuning or alignment, it targets behavior specialization: models must handle domain-specific evidence formats, output structures, action spaces, and evaluation criteria. Representative methods (Tab. 4) adapt along different paths. GUI agents such as Mobile-Agent (Wang et al. 2024c) integrate visual perception, planning, and stepwise actions, while GUI-R1 (Luo et al. 2025a) uses R1-style RL with rule-based rewards for cross-platform execution. Document and high-resolution settings are addressed by mPLUG-DocOwl1.5 (Hu et al. 2024) and LLaVA-UHD (Guo et al. 2024), which improve structure awareness and visual granularity. In medicine, Med-Gemini (Saab et al. 2024) supports clinical reasoning and specialized modalities, while AdaMLLM (Cheng et al. 2024a) adapts inference under domain and resource constraints. Overall, domain adaptation requires joint consideration of data distribution, visual detail, task interfaces, and reliability.

7 MMPoT of Scalable Training

This section examines how MLLM behavior can be shaped at scale. As MLLMs grow in model size, modality coverage, and application scope, the challenge is no longer merely to improve their capabilities, but to induce reliable behavior without repeatedly updating the full model or relying on costly supervision.

7.1 Parameter-Efficient Post-Training

7.1.1 Low-Rank Adaptation

Low-rank adaptation (LoRA) is a widely used parameter-efficient strategy for MLLMs post-training, as shown in Fig. 10 (left). Instead of updating the full pretrained backbone, LoRA

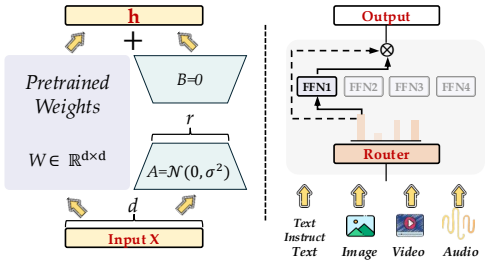

Fig. 10: **A basic view of behavior shaping in parameter-efficient post-training for scalable learning.**

freezes the original weights and inserts trainable low-rank update matrices into selected layers. During post-training, only these lightweight LoRA parameters are optimized, while the base model remains fixed. This makes LoRA especially suitable for adapting MLLMs to new instruction data, video tasks, or domain-specific datasets with limited computational cost. Building on this paradigm, MLLMs post-training has extended LoRA in several directions in Tab. 5. LLaVA-LoRA (Liu et al. 2023) applies LoRA to visual instruction tuning for efficient adaptation of LLaVA-style models. LLaVA-MoLE (Chen et al. 2024c) and MixLoRA (Shen et al. 2024b) augment LoRA with mixture-based routing to mitigate interference across heterogeneous instruction data. More recent methods, including MokA (Wei et al. 2026b) and LiLoRA (Che et al. 2026), tailor low-rank updates to multimodal and continual adaptation. Together, these studies position LoRA as both a parameter-efficient technique and a scalable mechanism for behavior shaping in MLLMs post-training.

7.1.2 Mixture-of-Experts Adaptation

Mixture-of-Experts (MoE) adaptation improves efficiency by routing each token or task to a subset of model parameters, as illustrated in the right panel of Fig. 10. This sparse activation is well-suited to MLLMs, where different modalities, domains, and reasoning patterns may benefit from specialized experts. Existing methods follow two main directions. As demonstrated in Tab. 6, MoE-LLaVA (Lin et al. 2026) and large-scale systems such as Qwen3-Omni (Xu et al. 2025a), Qwen3-VL (Team 2025), MiniMax-01 (MiniMax 2025), and Seed1.5-VL (Guo et al. 2025a) use sparse expert backbones to expand model capacity while limiting active computation. MoExtend (Zhong et al. 2024) instead adds new experts to pre-trained MoE models for new modalities or tasks while preserving the original backbone. Overall, MoE-based MLLMs post-training scales behavior shaping by selectively activating or extending specialized experts rather than uniformly updating dense parameters.

7.2 Compute-Efficient Post-Training

Compute-efficient post-training reduces the cost of adapting and deploying MLLMs while preserving task-relevant information. Existing methods (Tab. 7) mainly target high-resolution visual processing and excessive visual tokens. The mechanism underlying compute-efficiency is visualized in Fig. 11.

7.2.1 Efficient Visual Processing

Visual encoding is a major bottleneck because high-resolution inputs produce many visual tokens. Efficient methods preserve fine details while controlling training and inference costs. LLaVA-UHD (Guo et al. 2024) uses image modularization and token compression, while AdaMLLM/AdaLLaVA (Xu et al. 2025c) dynamically adjusts inference computation. InternVL2 (Chen et al. 2024e), UReader (Ye et al. 2023a), and Monkey (Li et al. 2024d) improve high-resolution processing through dynamic tiling, adaptive cropping, and patch-wise processing, respectively.

7.2.2 Token Compression

Token compression lowers prefill and attention costs by removing redundant visual tokens while retaining task-relevant evidence. FastV (Chen et al. 2024b) prunes tokens after early LLM layers, while VisionZip (Yang et al. 2025a) selects informative image and video tokens. SparseVLM (Zhang et al. 2024c) uses text-guided sparsification and token recycling, whereas TRIM (Song et al. 2025) relies on CLIP-based relevance. TokenPacker (Li et al. 2025d) adopts coarse-to-fine compression to preserve local details.

Table 7: **Representative multimodal compute-efficient methods in MMPoT.** EVP = efficient visual processing, TC = token compression, LCO = long-context optimization, HR = high resolution, OCR = optical character recognition.

Type	Method	Input	Granularity	Core Technique	Main Goal	Source
EVP	LLaVA-UHD (Guo et al. 2024)	HR image	Region / Token	Image modularization, token compression, spatial schema	Fine-grained HR perception	Paper/GitHub
	AdaMLLM/AdaLLaVA (Xu et al. 2025c)	Image	Instance / Budget	Dynamic inference reconfiguration	Accuracy–latency trade-off	Paper/GitHub
	InternVL2 (Chen et al. 2024e)	HR image	Tile	Dynamic image tiling	Dense visual perception	Project/GitHub
TC	FastV (Chen et al. 2024b)	Image / Video	Layer / Token	Attention-guided visual token pruning	Reduce prefilling and attention cost	Paper/GitHub
	VisionZip (Yang et al. 2025a)	Image / Video	Token	Informative token selection	Remove visual redundancy	Paper/GitHub
	SparseVLM (Zhang et al. 2024c)	Image / Video	Token / Layer	Text-guided sparsification, token recycling	Reduce visual FLOPs	Paper/GitHub
	TokenPacker (Li et al. 2025d)	Image	Projector / Token	Coarse-to-fine token packing	Compress while preserving details	Paper/GitHub
LCO	LongVU (Shen et al. 2024a)	Video	Frame / Token	Spatiotemporal adaptive compression	Long-video compression	Paper/GitHub
	LongVILA (Chen et al. 2025b)	Video	Sequence / System	Long-context extension, SFT, sequence parallelism	Scalable long-video training	Paper/GitHub
	LongVA (Zhang et al. 2024b)	Video	Context	Language-to-vision context transfer	Long-context video understanding	Paper/GitHub
	VideoChat-Flash (Li et al. 2024c)	Video	Hierarchical token	HiCo, short-to-long training	Extremely long-video modeling	Paper/GitHub

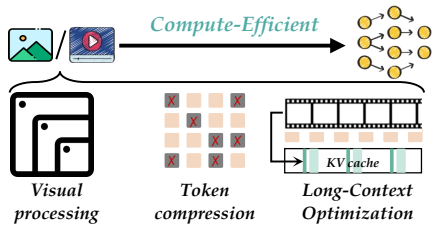

Fig. 11: **A basic view of behavior shaping in compute-efficient post-training for scalable learning.**

7.2.3 Long-Context Optimization

Long-context optimization reduces the cost of post-training MLLMs for long videos, multi-image inputs, and extended multimodal conversations. It manages spatial redundancy within frames and temporal redundancy across frames through compression, memory, sampling, and sequence parallelism. LongVU (Shen et al. 2024a) adaptively removes redundant frames and query-irrelevant spatial tokens. LongVILA (Chen et al. 2025b) combines context extension, long-video SFT, and multimodal sequence parallelism. LongVA (Zhang et al. 2024b) transfers extended context capabilities from language to vision, enabling long-video understanding without additional video training. VideoGrid-style methods arrange frames into temporal grids for denser coverage under fixed context budgets. VideoChat-Flash (Li et al. 2024c) combines hierarchical token compression with short-to-long training for long videos.

8 MLLMs Post-Training Benchmarks

This section reviews the datasets and benchmarks used for MLLMs post-training and evaluation. The discussion is organized into two parts. The first categorizes representative datasets, while the second reviews widely used metrics for general evaluation. A schematic overview is illustrated in Fig. 12.

8.1 Datasets and Benchmarks

Datasets and benchmarks play a central role in MLLMs post-training by defining which behaviors are learned, calibrated, and evaluated. As summarized in Tab. 8, existing resources can be grouped by their behavior-shaping objectives: instruction following, preference alignment, reasoning enhancement, and domain adaptation.

•**Instruction Following.** Instruction-following resources establish basic MLLM interaction by teaching models to interpret multimodal prompts and produce task-appropriate responses. LLaVA-Instruct-150K (Liu et al. 2023) and ShareGPT4V (Chen et al. 2024a) provide large-scale supervision for visual dialogue and general task completion. SEED-Bench (Li et al. 2023a), MM-Vet (Yu et al. 2023), MMBench (Liu et al. 2024f), and MME (Fu et al. 2026) evaluate generalization across perception, cognition, OCR, spatial reasoning, and bilingual settings. MM-IFInstruct (Ding et al. 2025) and VC-IFInstruct (He et al. 2026) further target fine-grained compliance with explicit visual and textual constraints.

•**Preference Calibration.** Alignment resources provide feedback for training and criteria for evaluating faithfulness, safety, and human preference. POPE (Li et al. 2023c), MMHal-Bench (Sun et al. 2024b), AMBER (Wang et al. 2023), and HallusionBench (Guan et al. 2024) assess visual grounding and hallucination. VLGuard (Zong et al.

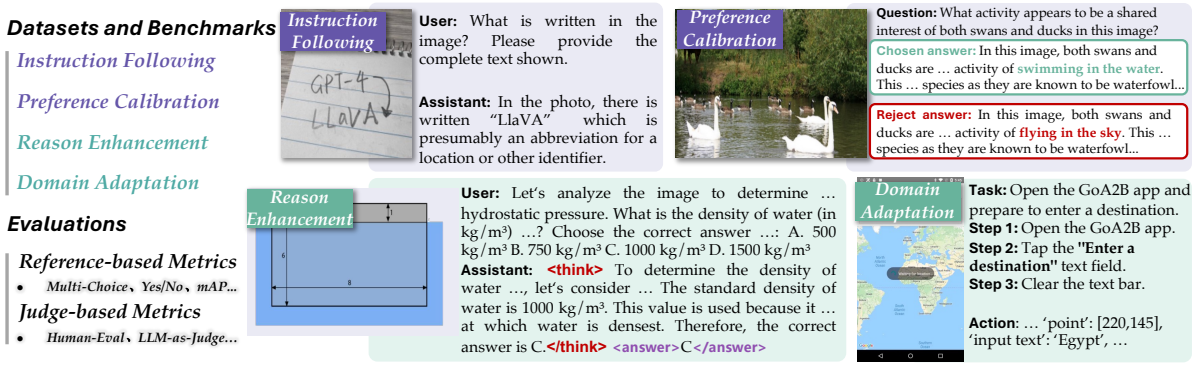

Fig. 12: Overview of the evaluation system of MMPoT.

2024), Lingua-SafetyBench (Shi et al. 2026), MM-SafetyBench (Liu et al. 2024d), FigStep (Gong et al. 2025), and JailBreakV-28K (Luo et al. 2024b) examine unsafe instructions, cross-modal attacks, and visual jailbreaks. LLaVA-RLHF (Sun et al. 2024a), RLHF-V (Yu et al. 2024), and SPA-VL (Zhang et al. 2025f) provide preference and feedback data for alignment optimization.

•**Reason Enhancement.** Reasoning resources support and evaluate multi-step inference over visual and textual evidence. ScienceQA (Lu et al. 2022), GQA (Hudson and Manning 2019), AI2D (Kembhavi et al. 2016), and CLEVR (Johnson et al. 2017) cover scientific, compositional, and structured visual reasoning. More challenging benchmarks, including MMMU (Yue et al. 2024), MathVista (Lu et al. 2024b), MathVision (Wang et al. 2024d), OlympiadBench (He et al. 2024), PuzzleBench (Zhang et al. 2025i), MME-CoT (Jiang et al. 2025), and MME-Reasoning (Yuan et al. 2025), emphasize expert knowledge, mathematical reasoning, *etc.*

•**Domain Adaptation.** Domain-specific resources extend evaluation from general visual dialogue to specialized inputs, evidence types, and task protocols. DocVQA (Mathew et al. 2021), OCRBench (Liu et al. 2024g), and ChartQA (Masry et al. 2022) assess document layouts, text recognition, and tables. Resources for medicine, GUI interaction, video, and high-resolution vision further evaluate adaptation to specialized evidence and action requirements. Compared with general instruction benchmarks, they place greater emphasis on robustness and deployability in specialized settings.

8.2 Evaluation Metrics

Evaluation metrics convert model outputs into comparable scores. Existing MLLMs benchmarks generally use reference-based or judge-based metrics, depending on whether outputs admit deterministic evaluation.

8.2.1 Reference-based Metrics

Reference-based metrics compare predictions with ground-truth answers and are suited to closed-ended or objectively verifiable tasks. Common choices include exact match for short-answer QA, accuracy for multiple-choice and yes/no questions, F1 for classification, and mAP or IoU for visual grounding. MME (Fu et al. 2026), MMBench (Liu et al. 2024f), SEED-Bench (Li et al. 2023a), and MMMU (Yue et al. 2024) mainly use accuracy-based scoring, whereas grounding tasks rely on localization metrics. Free-form generation is also evaluated using reference-similarity metrics such as BLEU, CIDEr, and ANLS. These metrics are common in captioning, VQA, and document understanding benchmarks, including ChartQA (Masry et al. 2022) and DocVQA (Mathew et al. 2021). However, they may overlook semantic correctness and reasoning quality. More robust designs include circular evaluation in MMBench (Liu et al. 2024f), log-probability scoring in MMMU (Yue et al. 2024), and reasoning-aware evaluation in MathVista (Lu et al. 2024b) and MME-CoT (Jiang et al. 2025).

8.2.2 Judge-based Metrics

Judge-based metrics are used for open-ended tasks with multiple valid responses. Human experts

Table 8: **Representative datasets and benchmarks for MMPoT.** Each entry is annotated by its primary role: T for datasets mainly used in training, E for benchmarks designed for evaluation, and T+E for resources containing both training and evaluation splits. The "Origin" column identifies the primary organization using common abbreviations.

Name	Year	Scale	Role	Origin	Source
Instruction Following					
LLaVA-Instruct-150K (Liu et al. 2023)	2023	∼80K images / 158K instruction samples	T	UW–Madison, Columbia	Paper / GitHub / Data
SEED-Bench (Li et al. 2023a)	2023	∼19K samples	E	Tencent AI Lab, ARC Lab	Paper / GitHub / Data
MM-Vet (Yu et al. 2023)	2023	200 images / 218 samples	E	NUS, Microsoft Azure AI	Paper / GitHub / Data
MMBench (Liu et al. 2024f)	2024	3,217 data samples	E	SHLAB, ZJU	Paper / GitHub / Data
ShareGPT4V (Chen et al. 2024a)	2024	1,346K samples	T	USTC, SHLAB	Paper / GitHub / Data
MM-IFInstruct (Ding et al. 2025)	2025	23K samples	T	FDU, SII	Paper / GitHub / Data
MME (Fu et al. 2026)	2026	1,187 images / 2,374 samples	E	SKL-NST, CASIA	Paper / GitHub / Data
VC-IFInstruct (He et al. 2026)	2026	10k samples	T	HKUST, PSU	Paper
Others: VQAv2 (Goyal et al. 2017), VizWiz (Gurari et al. 2018), MME-RealWorld (Zhang et al. 2025h), VideoMME (Fu et al. 2025), MMInstruct (Liu et al. 2024e)					
Preference Calibration					
POPE (Li et al. 2023c)	2023	∼9K question samples	E	RUC, Meituan Group	Paper / GitHub / Data
MMHal-Bench (Sun et al. 2024b)	2023	∼96 image-question pairs	E	UCB, MIT-IBM AI Lab	Paper / GitHub / Data
HallusionBench (Guan et al. 2024)	2024	∼346 images / ∼1.1K question samples	E	UMD	Paper / GitHub / Data
LLaVA-RLHF (Sun et al. 2024a)	2024	10K image-based conversations	E	UCB, MIT-IBM AI Lab	Paper / GitHub / Data
RLHF-V (Yu et al. 2024)	2024	1.4K image-based conversations	T	THU, NUS	Paper / GitHub / Data
Lingua-SafetyBench (Shi et al. 2026)	2026	∼100K image-question pairs	E	HKU, ZJU	Paper / GitHub / Data
Others: CHAIR (Rohrbach et al. 2018), MM-SafetyBench (Liu et al. 2024d), FigStep (Gong et al. 2025), JailBreakV-28K (Luo et al. 2024b)					
Reason Enhancement					
ScienceQA (Lu et al. 2022)	2022	∼10.3K image-question pairs / ∼21.2K questions	T+E	UCLA	Paper / GitHub / Data
MMMU (Yue et al. 2024)	2024	∼11.5K questions	E	Waterloo, OSU, CMU	Paper / GitHub / Data
MathVista (Lu et al. 2024b)	2024	6,141 samples	E	UCLA, MSR	Paper / GitHub / Data
PuzzleBench (Zhang et al. 2025i)	2025	11,840 samples	E	SJTU	Paper
MME-CoT (Jiang et al. 2025)	2025	1,130 questions	E	CUHK, ByteDance, NEU	Paper / GitHub / Data
MME-Reasoning (Yuan et al. 2025)	2025	1,188 questions	E	Fudan, CUHK, Shanghai AI Lab	Paper / GitHub / Data
AI2D (Kembhavi et al. 2016)	2016	∼5,000 diagrams / ∼15,000 questions and answers	T+E	AI2, UW	Paper / Data
Others: CLEVR (Johnson et al. 2017), OlympiadBench (He et al. 2024), MathVision (Wang et al. 2024d), MMMU-Pro (Yue et al. 2025a), PuzzleVQA (Chia et al. 2024)					
Domain Adaptation					
DocVQA (Mathew et al. 2021)	2020	∼12.8K images + ∼50K questions	T+E	UB, CVC	Paper / Data
Mind2Web (Deng et al. 2023)	2023	∼137 websites + ∼2.4K tasks + ∼137K annotated steps	T+E	OSU, CMU	Paper / GitHub / Data
OCRBench (Liu et al. 2024g)	2023	∼1K questions	E	SAIL	Paper / GitHub
ChartX (Xia et al. 2025)	2024	∼48K images / 6K validation+test samples	T+E	Shanghai AI Lab, SJTU	Paper / GitHub
ScreenSpot (Cheng et al. 2024b)	2024	∼610 screenshots + ∼1.3K GUI instructions	E	CUHK, SAIL	Paper / GitHub / Data
Others: AndroidControl (Li et al. 2024b), GUI-Odyssey (Lu et al. 2025), OmniAct (Kapoor et al. 2024), ChartQA (Masry et al. 2022)					

or capable LLM/MLLM judges typically assess responses in two ways. Score-based evaluation assigns numerical ratings for helpfulness, relevance, correctness, and detail. MM-Vet (Yu et al. 2023) and MMHal-Bench (Sun et al. 2024b) use this approach for open-ended multimodal responses. Comparison-based evaluation ranks candidate responses and reports win rates or Elo-style scores. It is commonly used in preference and alignment settings, such as LLaVA-RLHF (Sun et al. 2024a), RLHF-V (Yu et al. 2024), and VLFeedback/Silkie (Li et al. 2023b). Although judge-based metrics better capture subjective quality and user preference, their reliability depends on judge capability, prompt design, and evaluation consistency.

9 Future Directions

In this section, we discuss key directions for advancing MLLMs post-training beyond current paradigms. Our discussion is organized around three complementary themes: grounding behavior shaping, reliability-aware evaluation, and scaling for generalization.

9.1 Grounded Behavior Shaping

•**Native Multimodal Post-Training.** Most current post-training pipelines still organize multimodal behavior around language supervision, where visual, video, or audio inputs are eventually aligned to text-style responses. A future direction is to develop native multimodal post-training signals that preserve modality-specific structures, such as spatial layout, temporal continuity, acoustic cues, and cross-modal correspondence. This would allow MLLMs to learn behaviors directly from multimodal evidence rather than treating non-text signals as auxiliary context.

•**From Digital Understanding to Physical Interaction.** Current MLLMs are mainly optimized for understanding digital content such as images, videos, documents, and screens. However,

future multimodal agents must interact with the physical world, where perception is continuous, actions change the environment, and feedback is delayed or uncertain. Post-training should therefore connect visual understanding with action-oriented reasoning, enabling models to move from describing multimodal inputs to making grounded decisions in dynamic environments.

9.2 Reliability-aware Evaluation

•**Trustworthy Evaluation.** Existing evaluations often emphasize benchmark accuracy, but high scores do not necessarily imply reliable behavior. Future evaluation should diagnose whether post-trained MLLMs are visually grounded, calibrated, consistent, safe, and robust under distribution shifts. This requires metrics that expose hallucination, overconfidence, shortcut reasoning, and unstable responses, rather than only rewarding final-answer correctness.

•**Complex Real-world Scenarios.** Many benchmarks are still built from static images, short videos, or closed-form QA tasks. Real applications involve ambiguous evidence, long-horizon context, changing environments, and interaction constraints. Future benchmarks should therefore simulate more complex real-world scenarios, where models must maintain state, handle uncertainty, and adapt their behavior across continuous multimodal inputs.

9.3 Scaling for Generalization

•**Post-training Scaling toward Generalist MLLMs.** Post-training scaling should go beyond simply increasing data volume or training compute, and instead improve behavioral transfer across tasks, domains, modalities, and interaction settings. A key challenge is to determine how diverse supervision signals, model capacity, and optimization strategies jointly contribute to generalization, rather than merely fitting a larger collection of post-training examples. The ultimate goal is to develop generalist post-trained MLLMs whose learned behaviors can be composed, reused, and adapted beyond the datasets, task formats, and reward functions encountered during optimization.

•**Streaming Understanding of a Continuously Unfolding World.** MLLMs must eventually operate in a world that evolves continuously rather than within a fixed collection of benchmark samples. This requires post-training mechanisms that support persistent memory, temporal event abstraction, selective retrieval, online feedback integration, and continual adaptation. Models should be able to accumulate multimodal experiences over extended interactions, update their knowledge and behaviors when the environment changes, and distinguish transient observations from durable information. A sustainable MLLM should therefore absorb new experiences over time while preserving previously acquired capabilities, maintaining behavioral consistency, and avoiding catastrophic forgetting or uncontrolled model drift.

10 Conclusion

This survey reviews the rapidly growing literature on MLLMs post-training through a unified behavior-shaping perspective. We examine how post-training methods steer pretrained MLLMs toward reliable, grounded, and task-oriented behavior. We further connect these methods to their supervision sources, feedback signals, benchmarks, and evaluation protocols, providing a structured view of the full behavior-shaping loop. By synthesizing current progress, limitations, and future directions, this survey charts the evolving landscape of MLLMs post-training and points toward dependable multimodal intelligence.

Data Availability. This article surveys existing research on action quality assessment. No new datasets or code were generated during this study. All data and materials discussed are derived from publicly available sources, which are cited throughout the manuscript.

References

Agrawal P, Antoniak S, Hanna EB, et al (2024) Pixtral 12b. arXiv preprint arXiv:241007073

Ahn D, Choi Y, Yu Y, et al (2024) Tuning large multimodal models for videos using reinforcement learning from ai feedback. In: ACL

Ahn D, Choi Y, Kim S, et al (2025) Isr-dpo: Aligning large multimodal models for videos by iterative self-retrospective dpo. In: AAAI

Bai J, Bai S, Chu Y, et al (2023) Qwen technical report. arXiv preprint arXiv:230916609

Bai S, Chen K, Liu X, et al (2025) Qwen2.5-vl technical report. arXiv preprint arXiv:250213923

Cai Y, Zhang J, He H, et al (2025) Llava-kd: A framework of distilling multimodal large language models. In: ICCV, pp 239–249

Cao D, Fu D, Yu H, et al (2026) X-opd: Cross-modal on-policy distillation for capability alignment in speech llms. arXiv preprint arXiv:260324596

Chaubey A, Pang J, Soleymani M (2026) Moddpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. In: CVPR, pp 18284–18294

Che C, Wang Z, Yang P, et al (2026) Lora in lora: Towards parameter-efficient architecture expansion for continual visual instruction tuning. In: AAAI, pp 19978–19986

Chen J, Zhang T, Huang S, et al (2026) Omnidpo: A preference optimization framework to address omni-modal hallucination. In: AAAI, 24, pp 20172–20180

Chen K, Zhang Z, Zeng W, et al (2023) Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:230615195

Chen L, Li J, Dong X, et al (2024a) Sharegpt4v: Improving large multi-modal models with better captions. In: ECCV, pp 370–387

Chen L, Zhao H, Liu T, et al (2024b) An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: ECCV, pp 19–35

Chen S, Jie Z, Ma L (2024c) Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. arXiv preprint arXiv:240116160

Chen X, Shukla SN, Azab M, et al (2025a) Compcap: Improving multimodal large language models with composite captions. In: ICCV

Chen Y, Xue F, Li D, et al (2025b) Longvila: Scaling long-context visual language models for long videos. In: ICLR, pp 18227–18246

Chen Z, Wang W, Cao Y, et al (2024d) Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:241205271

Chen Z, Wu J, Wang W, et al (2024e) Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR

Cheng D, Huang S, Zhu Z, et al (2024a) On domain-adaptive post-training for multimodal large language models. arXiv preprint arXiv:241119930

Cheng K, Sun Q, Chu Y, et al (2024b) Seeclick: Harnessing gui grounding for advanced visual gui agents. In: ACL, pp 9313–9332

Cheng K, YanTao L, Xu F, et al (2025) Vision-language models can self-improve reasoning via reflection. In: ACL, pp 8876–8892

Chia YK, Toh V, Ghosal D, et al (2024) Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In: ACL, Findings, pp 16259–16273

Chiang WL, Li Z, Lin Z, et al (2023) Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna lmsys org (accessed 14 April 2023) p 6

Chu X, Qiao L, Lin X, et al (2023) Mobilevlm: A fast, strong and open vision language assistant for mobile devices. arXiv preprint arXiv:231216886

Compagnoni A, Caffagni D, Moratelli N, et al (2025) Mitigating hallucinations in multimodal llms via object-aware preference optimization. In: BMVC

Dai W, Li J, Li D, et al (2023) Instructblip: Towards general-purpose vision-language models with instruction tuning. NeurIPS

Deitke M, Clark C, Lee S, et al (2025) Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR, pp 91–104

Deng X, Gu Y, Zheng B, et al (2023) Mind2web: Towards a generalist agent for the web. NeurIPS pp 28091–28114

Ding S, Wu S, Zhao X, et al (2025) Mm-ifengine: Towards multimodal instruction following. In: ICCV, pp 1099–1109

Dong Y, Liu Z, Sun HL, et al (2025) Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In: CVPR, pp 9062–9072

Driess D, Xia F, Sajjadi MS, et al (2023) Palm-e: an embodied multimodal language model. In: ICML, pp 8469–8488

Duan C, Fang R, Wang Y, et al (2025) Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:250517022

Fan Y, He X, Yang D, et al (2026) Grit: Teaching mllms to think with images. NeurIPS pp 116522–116543

Feng K, Gong K, Li B, et al (2026) Video-r1: Reinforcing video reasoning in mllms. NeurIPS

Fu C, Dai Y, Luo Y, et al (2025) Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR, pp 24108–24118

Fu C, Chen P, Shen Y, et al (2026) Mme: A comprehensive evaluation benchmark for multimodal large language models. NeurIPS

Gao P, Han J, Zhang R, et al (2023) Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:230415010

Gong T, Lyu C, Zhang S, et al (2023) Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:230504790

Gong Y, Ran D, Liu J, et al (2025) Figstep: Jailbreaking large vision-language models via typographic visual prompts. In: AAAI, pp 23951–23959

Goyal Y, Khot T, Summers-Stay D, et al (2017) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR, pp 6904–6913

Guan T, Liu F, Wu X, et al (2024) Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR

Guo D, Wu F, Zhu F, et al (2025a) Seed1. 5-vl technical report. arXiv preprint arXiv:250507062

Guo D, Yang D, Zhang H, et al (2025b) Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:250112948

Guo Z, Xu R, Yao Y, et al (2024) Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In: ECCV, pp 390–406

Gurari D, Li Q, Stangl AJ, et al (2018) Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR, pp 3608–3617

Han J, Zhang R, Shao W, et al (2023) Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:230903905

He C, Luo R, Bai Y, et al (2024) Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In: ACL, pp 3828–3850

He W, Ju F, Fan Z, et al (2026) Empowering reliable visual-centric instruction following in mllms. In: ACL, pp 9460–9482

Hernandez J, Villegas R, Ordonez V (2024) Generative visual instruction tuning. arXiv preprint arXiv:240611262

Hong J, Zhao C, Zhu C, et al (2025) Deepeyesv2: Toward agentic multimodal model. arXiv preprint arXiv:251105271

Hou W, Peng S, Wang W, et al (2026) Uni-opd: Unifying on-policy distillation with

a dual-perspective recipe. arXiv preprint arXiv:260503677

Hu A, Xu H, Ye J, et al (2024) mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In: EMNLP, pp 3096–3120

Huang W, Jia B, Zhai Z, et al (2025) Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:250306749

Hudson DA, Manning CD (2019) Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR, pp 6700–6709

Jaech A, Kalai A, Lerer A, et al (2024) Openai o1 system card. arXiv preprint arXiv:241216720

Jia F, Mao W, Liu Y, et al (2023) Adriver-i: A general world model for autonomous driving. arXiv preprint arXiv:231113549

Jiang D, He X, Zeng H, et al (2024) Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:240501483

Jiang D, Zhang R, Guo Z, et al (2025) Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:250209621

Johnson J, Hariharan B, Van Der Maaten L, et al (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVP, pp 2901–2910

Kapoor R, Butala YP, Russak M, et al (2024) Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In: ECCV, pp 161–178

Kembhavi A, Salvato M, Kolve E, et al (2016) A diagram is worth a dozen images. In: ECCV, pp 235–251

Khayatkhoei M, Chhikara P, Ilievski F, et al (2025) Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: ICLR

Kim M, Pertsch K, Karamcheti S, et al (2024) Openvla: An open-source vision-language-action model. arXiv preprint arXiv:240609246

Lambert N, Morrison J, Pyatkin V, et al (2024) Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:241115124

Li B, Wang R, Wang G, et al (2023a) Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:230716125

Li B, Zhang Y, Guo D, et al (2024a) Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:240803326

Li B, Sun X, Liu J, et al (2025a) Latent visual reasoning. arXiv preprint arXiv:250924251

Li B, Zhang Y, Chen L, et al (2025b) Otter: A multi-modal model with in-context instruction tuning. TPAMI

Li J, Li D, Xiong C, et al (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML, pp 12888–12900

Li J, Jiang L, Zhang H, et al (2026a) Token reduction via local and global contexts optimization for efficient video large language models. In: CVPR, pp 10451–10461

Li J, Yin H, Xu H, et al (2026b) Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. arXiv preprint arXiv:260202994

Li K, He Y, Wang Y, et al (2025c) Videochat: Chat-centric video understanding. Science China Information Sciences (10):200102

Li L, Xie Z, Li M, et al (2023b) Silkie: Preference distillation for large visual language models. arXiv preprint arXiv:231210665

Li W, Bishop W, Li A, et al (2024b) On the effects of data scale on ui control agents. NeurIPS 37:92130–92154

Li W, Yuan Y, Liu J, et al (2025d) Tokenpacker: Efficient visual projector for multimodal llm. IJCV pp 6794–6812

Li X, Wang Y, Yu J, et al (2024c) Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:250100574

Li Y, Du Y, Zhou K, et al (2023c) Evaluating object hallucination in large vision-language models. In: EMNLP, pp 292–305

Li Z, Yang B, Liu Q, et al (2024d) Monkey: Image resolution and text label are important things for large multi-modal models. In: CVPR, pp 26763–26773

Lin B, Tang Z, Ye Y, et al (2026) Moe-llava: Mixture of experts for large vision-language models. TMM

Lin J, Yin H, Ping W, et al (2024) Vila: On pretraining for visual language models. In: CVPR, pp 26689–26699

Liu H, Li C, Wu Q, et al (2023) Visual instruction tuning. NeurIPS pp 34892–34916

Liu H, Li C, Li Y, et al (2024a) Improved baselines with visual instruction tuning. In: CVPR

Liu H, Li C, Li Y, et al (2024b) Llavanext: Improved reasoning, ocr, and world knowledge

Liu J, Huang X, Zheng J, et al (2024c) Mm-instruct: Generated visual instructions for large multimodal model alignment. arXiv preprint arXiv:240619736

Liu R, Lv X, Li G, et al (2026) Visual-advantage on-policy distillation for vision-language models. arXiv preprint arXiv:260521924

Liu X, Zhu Y, Gu J, et al (2024d) Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In: ECCV, pp 386–403

Liu Y, Cao Y, Gao Z, et al (2024e) Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. Science China Information Sciences p 220103

Liu Y, Duan H, Zhang Y, et al (2024f) Mmbench: Is your multi-modal model an all-around player? In: ECCV, pp 216–233

Liu Y, Li Z, Huang M, et al (2024g) Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences p 220102

Liu Y, Qu T, Zhong Z, et al (2025a) Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. arXiv preprint arXiv:250512081

Liu Y, Zhang Y, Cai J, et al (2025b) Lamra: Large multimodal model as your advanced retrieval assistant. In: CVPR

Liu Z, Sun Z, Zang Y, et al (2025c) Visual-rft: Visual reinforcement fine-tuning. In: ICCV, pp 2034–2044

Lu H, Liu W, Zhang B, et al (2024a) Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:240305525

Lu P, Mishra S, Xia T, et al (2022) Learn to explain: Multimodal reasoning via thought chains for science question answering. NeurIPS pp 2507–2521

Lu P, Bansal H, Xia T, et al (2024b) Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: ICLR

Lu Q, Shao W, Liu Z, et al (2025) Guiodyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. In: ICCV, pp 22404–22414

Luo G, Zhou Y, Ren T, et al (2023a) Cheap and quick: Efficient vision-language instruction tuning for large language models. NeurIPS pp 29615–29627

Luo J, Dong P, Zhai Y, et al (2024a) Rlif: Interactive imitation learning as reinforcement learning. In: ICLR, pp 36329–36351

Luo R, Zhao Z, Yang M, et al (2023b) Valley: Video assistant with large language model enhanced ability. TOMM

Luo R, Wang L, He W, et al (2025a) Gui-r1: A generalist r1-style vision-language action model for gui agents. arXiv preprint arXiv:250410458

Luo R, Zhang H, Chen L, et al (2025b) Mmevol: Empowering multimodal large language models with evol-instruct. In: ACL, Findings, pp 19655–19682

Luo R, Zhao Z, Yang M, et al (2026) Valley: Video assistant with large language model enhanced ability. TOMM

Luo W, Ma S, Liu X, et al (2024b) Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. arXiv preprint arXiv:240403027

Marino K, Rastegari M, Farhadi A, et al (2019) Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR

Masry A, Tan JQ, Joty S, et al (2022) Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: ACL, Findings, pp 2263–2279

Mathew M, Karatzas D, Jawahar C (2021) Docvqa: A dataset for vqa on document images. In: WACV, pp 2200–2209

Meng F, Du L, Liu Z, et al (2025) Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:250307365

MiniMax (2025) Minimax-01: Scaling foundation models with lightning attention.

Ni M, Yang Z, Li L, et al (2026) Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. NeurIPS pp 20538–20559

Oh C, Li J, Im S, et al (2026) Visual instruction bottleneck tuning. NeurIPS pp 129164–129204

Ouali Y, Bulat A, Martinez B, et al (2024) Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In: ECCV, pp 395–413

Peng W, Meng L, Chen Y, et al (2026) Inst-it: Boosting instance understanding via explicit visual prompt instruction tuning. NeurIPS pp 50062–50092

Peng Y, Wang P, Wang X, et al (2025) Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. arXiv preprint arXiv:250405599

Peng Z, Wang W, Dong L, et al (2024) Grounding multimodal large language models to the world. In: ICLR

Qiu L, Ning S, Zhang C, et al (2026) Da-dpo: Cost-efficient difficulty-aware preference optimization for reducing mllm hallucinations. arXiv preprint arXiv:260100623

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: ICML

Rafailov R, Sharma A, Mitchell E, et al (2023) Direct preference optimization: Your language model is secretly a reward model. NeurIPS pp 53728–53741

Rohrbach A, Hendricks LA, Burns K, et al (2018) Object hallucination in image captioning. In: EMNLP

Saab K, Tu T, Weng WH, et al (2024) Capabilities of gemini models in medicine. arXiv preprint arXiv:240418416

Sarto S, Cornia M, Cucchiara R (2025) Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. arXiv preprint arXiv:250314604

Shao Z, Wang P, Zhu Q, et al (2024) Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:240203300

Shen H, Liu P, Li J, et al (2025) Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:250407615

Shen X, Xiong Y, Zhao C, et al (2024a) Longvu: Spatiotemporal adaptive compression for long

video-language understanding. arXiv preprint arXiv:241017434

Shen Y, Xu Z, Wang Q, et al (2024b) Multimodal instruction tuning with conditional mixture of lora. In: ACL, pp 637–648

Shi D, Glatt R, Klymko C, et al (2025) Oracle-rlaif: An improved fine-tuning framework for multi-modal video models through reinforcement learning from ranking feedback. arXiv preprint arXiv:251002561

Shi E, Shao P, Zhang Y, et al (2026) Linguasafetybench: A benchmark for safety evaluation of multilingual vision-language models. arXiv preprint arXiv:260122737

Shu F, Liao Y, Zhang L, et al (2025) Llava-mod: Making llava tiny via moe-knowledge distillation. In: ICLR, pp 9386–9404

Song D, Wang W, Chen S, et al (2025) Less is more: A simple yet effective token reduction method for efficient multi-modal llms. In: ACL, pp 7614–7623

Stephan M, Khazatsky A, Mitchell E, et al (2024) Rlvf: learning from verbal feedback without overgeneralization. In: ICML, pp 46625–46656

Su Z, Li L, Song M, et al (2025) Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:250508617

Sun Z, Shen S, Cao S, et al (2024a) Aligning large multimodal models with factually augmented rlhf. In: ACL, Finding's

Sun Z, Shen S, Cao S, et al (2024b) Aligning large multimodal models with factually augmented rlhf. In: ACL, pp 13088–13110

Team G, Anil R, Borgeaud S, et al (2023) Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805

Team Q (2025) Qwen3-vl technical report.

Tong S, Brown E, Wu P, et al (2024) Cambrian-1: A fully open, vision-centric exploration of multimodal llms. NeurIPS pp 87310–87356

Touvron H, Lavril T, Izacard G, et al (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971

Tu J, Ni Z, Crispino N, et al (2025) Mlan: Language-based instruction tuning preserves and transfers knowledge in multimodal language models. In: KnowFM, pp 59–74

Viveiros AG, Gonçalves N, Lindemann M, et al (2026) Lantern: Latent visual structured reasoning. arXiv preprint arXiv:260325629

Wan Z, Dou Z, Liu C, et al (2026) Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. NeurIPS

Wang B, Wu F, Han X, et al (2024a) Vigc: Visual instruction generation and correction. In: AAAI, 6, pp 5309–5317

Wang F, Zhou W, Huang JY, et al (2024b) mdpo: Conditional preference optimization for multimodal large language models. In: EMNLP, pp 8078–8088

Wang H, Hu K, Gao L (2025a) Docvideoqa: Towards comprehensive understanding of document-centric videos through question answering. In: ICASSP

Wang J, Wang Y, Xu G, et al (2023) Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:231107397

Wang J, Xu H, Ye J, et al (2024c) Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. arXiv preprint arXiv:240116158

Wang K, Pan J, Shi W, et al (2024d) Measuring multimodal mathematical reasoning with math-vision dataset. NeurIPS pp 95095–95169

Wang P, Yang A, Men R, et al (2022) Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: ICML, pp 23318–23340

Wang P, Bai S, Tan S, et al (2024e) Qwen2-vl: Enhancing vision-language model's perception

of the world at any resolution. arXiv preprint arXiv:240912191

Wang W, Lv Q, Yu W, et al (2024f) Cogvlm: Visual expert for pretrained language models. NeurIPS

Wang W, Gao Z, Chen L, et al (2025b) Visualprm: An effective process reward model for multimodal reasoning. arXiv preprint arXiv:250310291

Wang X, Li C, Yang J, et al (2025c) Llava-critic-r1: Your critic model is secretly a strong policy model. arXiv preprint arXiv:250900676

Wang X, Yang Z, Feng C, et al (2026) Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. NeurIPS

Wang Z, Zhu J, Tang B, et al (2025d) Jigsaw-r1: A study of rule-based visual reinforcement learning with jigsaw puzzles. arXiv preprint arXiv:250523590

Wei L, Li Y, Wang C, et al (2026a) First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. NeurIPS pp 62293–62318

Wei Y, Miao Y, Zhou D, et al (2026b) Moka: Multimodal low-rank adaptation for mllms. NeurIPS pp 137470–137492

Wu M, Yang J, Jiang J, et al (2025) Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. arXiv preprint arXiv:250519255

Wu S, Fei H, Qu L, et al (2023) Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:230905519

Wu Z, Shi K, Zhang C, et al (2026) When models judge themselves: Unsupervised self-evolution for multimodal reasoning. arXiv preprint arXiv:260321289

Xia R, Ye H, Yan X, et al (2025) Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. TIP

Xiaomi LCT (2025) Mimo-vl technical report. URL https://arxiv.org/abs/2506.03569, arXiv:2506.03569

Xie Y, Li G, Xu X, et al (2024) V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In: EMNLP, pp 13258–13273

Xiong T, Wang X, Guo D, et al (2025) Llava-critic: Learning to evaluate multimodal models. In: CVPR

Xu J, Guo Z, Hu H, et al (2025a) Qwen3-omni technical report. arXiv preprint arXiv:250917765

Xu S, Li X, Yuan H, et al (2024a) Llavadi: What matters for multimodal large language models distillation. arXiv preprint arXiv:240719409

Xu Y, Li C, Zhou H, et al (2025b) Visual planning: Let's think only with images. arXiv preprint arXiv:250511409

Xu Z, Shen Y, Huang L (2023) Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In: ACL, pp 11445–11465

Xu Z, Zhang Y, Xie E, et al (2024b) Drivegpt4: Interpretable end-to-end autonomous driving via large language model. RA-L (10):8186–8193

Xu Z, Nguyen KD, Mukherjee P, et al (2025c) Learning to inference adaptively for multimodal large language models. In: ICCV, pp 3552–3563

Xu Z, Chen D, Ling Z, et al (2026) Mindgym: What matters in question synthesis for thinking-centric fine-tuning? NeurIPS

Yang S, Chen Y, Tian Z, et al (2025a) Visionzip: Longer is better but not necessary in vision language models. In: CVPR

Yang S, Niu Y, Liu Y, et al (2026) Look-back: Implicit visual re-focusing in mllm reasoning. In: AAAI, pp 11694–11702

Yang Y, He X, Pan H, et al (2025b) R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In: ICCV

Yang Z, Luo X, Han D, et al (2025c) Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: CVPR

Yao Y, Yu T, Zhang A, et al (2024) Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:240801800

Ye J, Hu A, Xu H, et al (2023a) Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In: EMNLP, pp 2841–2858

Ye Q, Xu H, Xu G, et al (2023b) mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:230414178

You H, Zhang H, Gan Z, et al (2024) Ferret: Refer and ground anything anywhere at any granularity. In: ICLR, pp 57153–57180

You Z, Nie S, Zhang X, et al (2026) Llada-v: Large language diffusion models with visual instruction tuning. In: CVPR

Yu Q, Zhang Z, Zhu R, et al (2026) Dapo: An open-source llm reinforcement learning system at scale. NeurIPS pp 113222–113244

Yu T, Yao Y, Zhang H, et al (2024) Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: CVPR

Yu T, Zhang H, Li Q, et al (2025a) Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In: CVPR, pp 19985–19995

Yu T, Zhang YF, Fu C, et al (2025b) Aligning multimodal llm with human preference: A survey. arXiv preprint arXiv:250314504

Yu W, Yang Z, Li L, et al (2023) Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:230802490

Yuan J, Peng T, Jiang Y, et al (2025) Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. arXiv preprint arXiv:250521327

Yuan Q, Lou J, Yu X, et al (2026) Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. arXiv preprint arXiv:260518740

Yue X, Ni Y, Zhang K, et al (2024) Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR, pp 9556–9567

Yue X, Zheng T, Ni Y, et al (2025a) Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: ACL, pp 15134–15186

Yue Z, Lin Z, Song Y, et al (2025b) Mimo-vl technical report. arXiv preprint arXiv:250603569

Zadeh FP, Oh Y, Kim G (2025) Lpoi: Listwise preference optimization for vision language models. In: ACL, pp 26830–26844

Zeng Y, Huang W, Huang S, et al (2025) Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. arXiv preprint arXiv:251001304

Zhang H, Li X, Bing L (2023) Video-llama: An instruction-tuned audio-visual language model for video understanding. In: EMNLP, pp 543–553

Zhang H, You H, Dufter P, et al (2024a) Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv preprint arXiv:240407973

Zhang H, Luo R, Liu X, et al (2025a) Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In: ACL, pp 26318–26331

Zhang H, Zeng P, Gao L, et al (2025b) Text-video retrieval with global-local semantic consistent learning. TIP

Zhang H, Mao Z, Zhang L, et al (2026a) Uncertainty-aware exploratory direct preference optimization for multimodal large language models. In: CVPR, pp 37831–37841

Zhang H, Zeng P, Zhang J, et al (2026b) Omnicharacter++: Towards comprehensive

benchmark for realistic role-playing agents. TPAMI

Zhang J, Fan Q, Zhang Y (2025c) Docassistant: Integrating key-region reading and step-wise reasoning for robust document visual question answering. In: EMNLP, pp 3496–3511

Zhang P, Zhang K, Li B, et al (2024b) Long context transfer from language to vision. arXiv preprint arXiv:240616852

Zhang R, Gui L, Sun Z, et al (2025d) Direct preference optimization of video large multimodal models from language model reward. In: ACL

Zhang S, Sun P, Chen S, et al (2025e) Gpt4roi: Instruction tuning large language model on region-of-interest. In: ECCV, pp 52–70

Zhang S, Dong L, Li X, et al (2026c) Instruction tuning for large language models: A survey. ACM Computing Surveys pp 1–36

Zhang Y, Fan CK, Ma J, et al (2024c) Sparsevlm: Visual token sparsification for efficient vision-language model inference. arXiv preprint arXiv:241004417

Zhang Y, Chen L, Zheng G, et al (2025f) Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In: CVPR, pp 19867–19878

Zhang Y, Yu T, Tian H, et al (2025g) Mm-rlhf: The next step forward in multimodal llm alignment. In: ICML

Zhang Y, Zhang H, Tian H, et al (2025h) Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In: ICLR, pp 89655–89701

Zhang Z, Chen Z, Zhang Z, et al (2025i) Puzzlebench: A fully dynamic evaluation framework for large multimodal models on puzzle solving. arXiv preprint arXiv:250410885

Zhao J, Wei X, Bo L (2025) R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. arXiv preprint arXiv:250305379

Zhao Z, Wang B, Ouyang L, et al (2023) Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:231116839

Zheng Z, Yang M, Hong J, et al (2025) Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:250514362

Zhong S, Gao S, Huang Z, et al (2024) Moextend: Tuning new experts for modality and task extension. In: ACL, pp 494–505

Zhou G, Qiu P, Chen C, et al (2025a) Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. arXiv preprint arXiv:250421277

Zhou H, Li X, Wang R, et al (2025b) R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:250305132

Zhou J, Chen Y, Li H, et al (2026) V-reflection: Transforming mllms from passive observers to active interrogators. arXiv preprint arXiv:260403307

Zhu D, Shen X, Li X, et al (2024) Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR

Zhu J, Wang W, Chen Z, et al (2025) Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:250410479

Zhu L, Ji D, Chen T, et al (2026) Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. NeurIPS

Zong Y, Bohdal O, Yu T, et al (2024) Safety fine-tuning at (almost) no cost: A baseline for vision large language models. arXiv preprint arXiv:240202207