Perception, Reason, Thought, and Plan: A Survey on Large Multimodal Reasoning Models

Yunxin Li*, **Zhenyu Liu***, **Zitao Li***, **Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang Zhuotao Tian, Baoxing Huai, Wenhan Luo, Zheng Zhang, Baotian Hu‡, Min Zhang**

Harbin Institute of Technology, Shenzhen

Project: https://github.com/HITsz-TMG/Awesome-Large-Multimodal-Reasoning-Models

Abstract

Reasoning lies at the heart of intelligence, shaping the ability to make decisions, draw conclusions, and generalize across domains. In artificial intelligence, as systems increasingly operate in open, uncertain, and multimodal environments, reasoning becomes essential for enabling robust and adaptive behavior. Large Multimodal Reasoning Models (LMRMs) have emerged as a promising paradigm, integrating modalities such as text, images, audio, and video to support complex reasoning capabilities. It aims to achieve comprehensive perception, precise understanding, and deep reasoning. As research advances, multimodal reasoning has rapidly evolved from modular, perception-driven pipelines to unified, language-centric frameworks that offer more coherent cross-modal understanding. While instruction tuning and reinforcement learning have improved model reasoning, significant challenges remain in omni-modal generalization, reasoning depth, and agentic behavior. To address these issues, we present a comprehensive and structured survey of multimodal reasoning research, organized around a four-stage developmental roadmap that reflects the field's shifting design philosophies and emerging capabilities. First, we review early efforts based on task-specific modules, where reasoning was implicitly embedded across stages of representation, alignment, and fusion. Next, we examine recent approaches that unify reasoning into multimodal LLMs, with advances such as Multimodal Chain-of-Thought (MCoT) and multimodal reinforcement learning enabling richer and more structured reasoning chains. Finally, drawing on empirical insights from challenging benchmarks and experimental cases of OpenAI O3 and O4-mini, we discuss the conceptual direction of native large multimodal reasoning models (N-LMRMs), which aim to support scalable, agentic, and adaptive reasoning and planning in complex, real-world environments. By synthesizing historical trends and emerging research, this survey aims to clarify the current landscape and inform the design of next-generation multimodal reasoning systems.

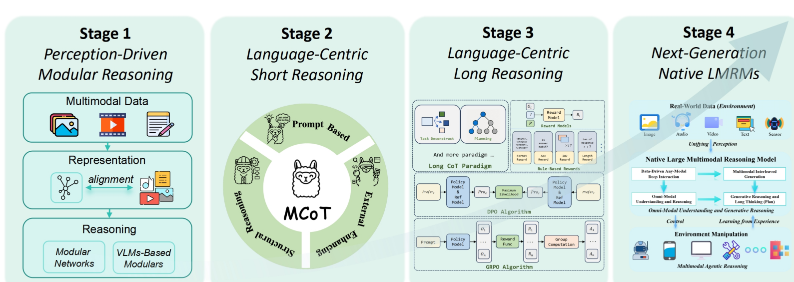

Figure 1: The core path of large multimodal reasoning models

*Equal contribution ‡ Corresponding author, email: hubaotian@hit.edu.cn

Contents

1 Introduction

In both philosophy and artificial intelligence, reasoning is widely regarded as a cornerstone of intelligent behavior (Kahneman, 2011; Su et al., 2024; de Winter et al., 2024; Bi et al., 2025). It enables agents not only to adaptively respond to their environments but also to draw logical inferences, generalize knowledge across diverse contexts, and navigate complex challenges. As AI systems increasingly interact with dynamic, uncertain, and multimodal settings, the ability to perform right reasoning under various environments becomes essential for achieving robust and adaptive intelligence (Yang et al., 2025a; Christakopoulou et al., 2024). In this context, Large Multimodal Reasoning Models (LMRMs) have emerged as a promising direction (Wang et al., 2024k; Zhang et al., 2024c; Yin et al., 2023), which integrate multiple data modalities, such as text, images, audio, and video, and exhibit complex reasoning abilities, including logical deduction, causal inference, analogical mapping, and long-horizon thinking. The core objective of LMRMs is to enable *comprehensive perception, precise understanding, and deep reasoning*, supporting the decision-making process in diverse environments.

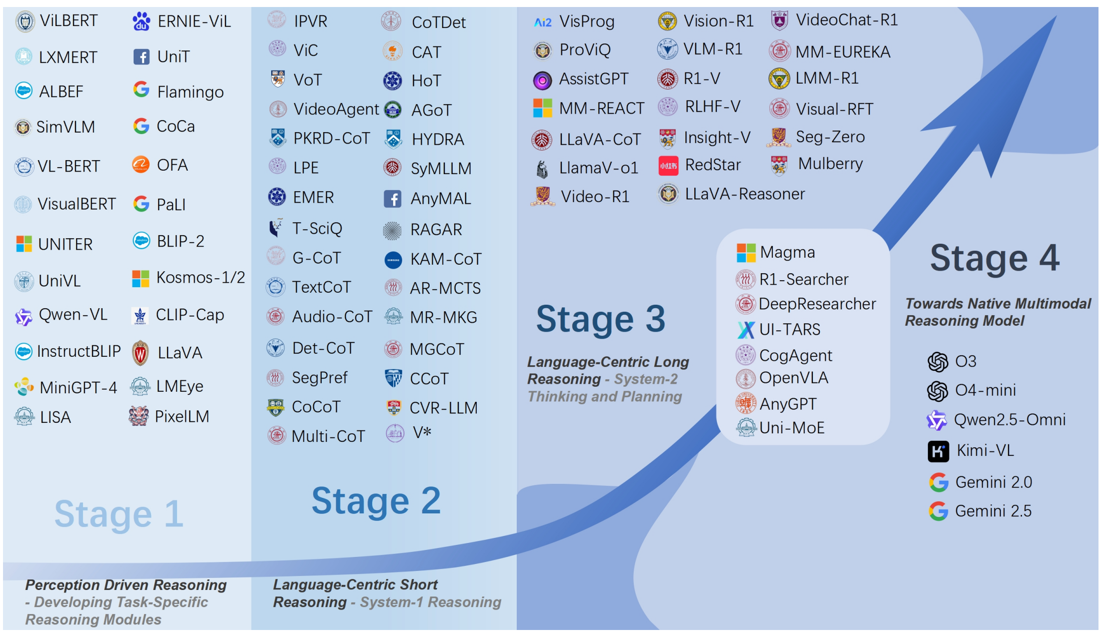

Figure 2: The roadmap of large multimodal reasoning models. The models highlighted in the box are representative models transitioning from Stage 3 towards Stage 4, as indicated by the directional arrow.

Research in multimodal reasoning has progressed rapidly. Early efforts relied on perception-driven, modular pipelines, while recent advances leverage large language models to unify multimodal understanding and reasoning (Huang et al., 2023b; Driess et al., 2023). Instruction tuning (Liu et al., 2023a) and reinforcement learning (DeepSeek-AI et al., 2025) further enhance models' reasoning performance, bringing them closer to human-like deliberative behaviour. Despite this rapid progress, multimodal reasoning is still a core bottleneck of large multimodal models, where they show limiting generalization, depth of reasoning, and agent-like behavior (Yue et al., 2024; Zhang et al., 2024f; Liu et al., 2024f).

Previous surveys in the field have largely focused on either multimodal large language models or the analysis of reasoning methods primarily centred on language, lacking a detailed analysis of recent reinforcement-enhanced multimodal reasoning and technical prospects of LMRMs. Hence, the multimodal reasoning area needs a coherent framework to understand how multimodal reasoning has evolved and where it is heading. Our work addresses a critical gap by providing a comprehensive review and analysis of the entire roadmap of multimodal reasoning models, encompassing early modular designs and state-of-the-art LMRMs. Furthermore, we project future developments of LMRMs, grounded in experimental findings and technical scrutiny.

Specifically, we propose a structured roadmap of multimodal reasoning, organized into three stages (Figure 2): *Perception-Driven Modular Reasoning*, where reasoning is implicit within task-specific modules; *Language-Centric Short Reasoning (System-1)*, where multimodal reasoning emerges via prompt-based and structured short CoT with LLMs; and *Language-Centric Long Reasoning (System-2)*, where long thinking, planning, and agentic behaviors are enabled through extended reasoning chains and reinforcement learning.

Building upon this developmental trajectory, we introduce the notion of *Native Large Multimodal Reasoning Models (N-LMRMs)*, a forward-looking paradigm where reasoning is no longer retrofitted onto language models, but instead natively emerges from omnimodal perception and interaction, and goal-driven cognition. By grounding this vision in recent progress on unified representations, training data synthesis, learning from world experience, and benchmark construction, we outline possible directions for advancing multimodal intelligence beyond current architectural constraints.

Our contributions are mainly threefold:

- This paper presents a comprehensive survey of the Large Multimodal Reasoning Model (LMRM) landscape, encompassing over 540 publications. Our analysis contextualizes and addresses key reasoning limitations in current models (Sec. 2).
- We propose a three-stage roadmap for the development of LMRMs from modular reasoning to multimodal chain-of-thought (MCoT), and finally to long-horizon, system-2 reasoning. Each stage is further analyzed with detailed taxonomies and representative methods (Sec. 3).
- We introduce and analyze Native Large Multimodal Reasoning Models (N-LMRMs), providing a thorough overview of initial progress, including architectures, learning methods, datasets, and benchmarks, thus setting the stage for future multimodal agentic reasoning (Sec. 4).
- We reorganize existing datasets and benchmarks (update to 2025.04) of multimodal understanding and reasoning (Sec. 5) to clarify their categories and evaluation dimensions.

2 Evolving Paradigms of Multimodal Reasoning and Discussion

The evolution of multimodal reasoning has undergone a series of significant paradigm shifts, reflecting a deeper integration of perceptual inputs with structured cognitive processes. In this section, we outline **four** key stages in the development of multimodal reasoning systems, each embodying distinct model design, capabilities, and technical challenges. This historical perspective not only situates the current state of the field but also clarifies the motivations for the directions explored in later sections of this survey.

Stage 1: Perception-Driven Modular Reasoning - Designing Task-Specific Reasoning Systems

In the initial stage, multimodal reasoning capabilities were developed through modular, reasoning modules (Andreas et al., 2016; Yang et al., 2016; Xiong et al., 2016). These systems typically employed convolutional neural networks (CNNs) and recurrent architectures such as long short-term memory (LSTM) networks within supervised learning frameworks. Due to challenges such as limited multimodal data, immature neural architectures, and underdeveloped learning methodologies, early research adopted modular designs that decomposed the reasoning process into separate components: representation, alignment, fusion, and reasoning (§3.1.1). As the field gradually shifted toward a pretraining-finetuning paradigm (Devlin et al., 2019; Radford et al., 2018, 2021), the emergence of large-scale multimodal datasets and deeper neural networks facilitated the rise of pretrained vision–language models (VLMs) (Chen et al., 2020; Li et al., 2020; Yu et al., 2022, 2021), which aimed to unify the processes of representation, alignment, and fusion (§3.1.2).

However, this unification primarily emphasized visual representation and cross-modal fusion, often at the expense of deeper semantic modelling of language. As a result, the reasoning process frequently defaulted to a classification-based paradigm, limiting context-aware and generalized reasoning. The multimodal reasoning systems still rely on additional modules or task-specific enhancements. Overall, the reasoning in this stage remained largely implicit by foundational perceptual processing and neural computation. The emerging multimodal language models will enhance implicit reasoning by introducing powerful language models and large-scale visual data.

Stage 2: Language-Centric Short Reasoning - System-1 Reasoning

The advent of multimodal large language models (MLLMs) (Liu et al., 2023a; Bai et al., 2023; Chen et al., 2024j; Zhang et al., 2023c) marked a pivotal shift in multimodal reasoning: moving from modular systems to end-to-end language-centric frameworks. These models achieved strong performance in tasks such as visual commonsense reasoning (VCR) (Zellers et al., 2019; Yu et al., 2024c), visual question answering (VQA) (Goyal et al., 2017; Singh et al., 2019), and visual grounding (Peng et al., 2023; Rasheed et al., 2024; Liu et al., 2024f; Lai et al., 2024; Rasheed et al., 2024; Ren et al., 2024).

However, early MLLM architectures largely relied on surface-level pattern matching and static knowledge retrieval, falling short in dynamic hypothesis generation, multi-step logical progression, and context-sensitive adaptation. This limitation catalyzed the development of Chain-of-Thought (CoT) reasoning (Kojima et al., 2022), which transforms implicit reasoning into explicit intermediate steps, internalizing the thought processes within end-to-end generation. By aligning the representational capacity of Stage 1's multimodal fusion with the linguistic expressiveness of LLMs, CoT enables more contextualized and interpretable reasoning.

Building on CoT's success in pure language models, researchers extended it to the multimodal domain through the development of Multimodal Chain-of-Thought (MCoT) (Zhang et al., 2023g; Fei et al., 2024; Zhang et al., 2023b; Shao et al., 2024). Early approaches primarily focused on prompt-based adaptations (§3.2.1), enabling models to produce step-by-step multimodal reasoning traces by carefully crafted instructions. Subsequent efforts enhanced the reasoning process itself, either by introducing structured decomposition of reasoning paths (§3.2.2) or by leveraging external tools and retrieval augmentation to expand inference capabilities beyond the model's static knowledge (§3.2.3).

Nevertheless, reasoning at this stage predominantly remained short and reactive—characteristic of fast, intuitive System-1 reasoning. Models are effective for familiar or bounded tasks but struggle with abstraction, compositionality, and planning. These challenges spurred the development of more deliberate, structured reasoning paradigms, setting the stage for the next major transition.

Stage 3: Language-Centric Long Reasoning - System-2 Thinking and Planning

While MCoT has significantly advanced the reasoning capabilities of MLLMs, it remains insufficient for addressing the complexity of real-world multimodal tasks (Zhang et al., 2024f; Yu et al., 2024c; Yue et al., 2024). Most MCoT methods operate through short, reactive chains—resembling fast, intuitive System-1 reasoning. These approaches are effective for familiar or bounded problems but struggle with abstraction, compositionality, long-horizon reasoning, and adaptive planning (DeepSeek-AI et al., 2025). To bridge this gap, recent research has turned toward System-2-inspired reasoning (Yao et al., 2023b; Kahneman, 2011), emphasizing slower, deliberate, and methodologically structured cognitive processes. In this view, the reasoning is no longer treated as a mere function but as a core component of intelligent behaviour itself. Extending MCoT along three critical dimensions–*reasoning modalities*, *reasoning paradigms*, and *learning methods*–has become a key trajectory toward a new class of models: **Large Multimodal Reasoning Models (LMRMs)**, capable of deeper, transferable, and cognitively grounded reasoning.

First, from the perspective of reasoning modality, relying solely on textual representations constrains the model's ability to capture modality-specific knowledge. Recent studies (Lin et al., 2025a; Gao et al., 2024a; Li et al., 2025b; Zhou et al., 2024b; Rose et al., 2023) introduce *cross-modal reasoning chains* that leverage visual, auditory, and linguistic signals as joint substrates for inference, enabling richer semantic grounding and more faithful information integration (§3.3.1).

Second, regarding reasoning paradigms, researchers construct longer, higher-quality chains and introduce generalized, methodologically guided reasoning strategies (Jaech et al., 2024; Yao et al., 2024a). These approaches allow models to autonomously decompose complex tasks and apply transferable procedures across diverse contexts. Notably, the O1 family (e.g., GPT-4o (Hurst et al., 2024)) exemplifies near-human-level performance on a broad range of cognitively demanding multimodal tasks (§3.3.2).

Finally, from a learning method perspective, reinforcement learning-enhanced multimodal reasoning has gained increasing momentum. By incorporating agentic data, iterative feedback, and long-horizon optimization objectives, models like DeepSeek-R1 (DeepSeek-AI et al., 2025) improve their planning, robustness, and adaptive generalization. This line of work has catalyzed the emergence of a new generation of R1-like models emphasizing scalable, methodologically grounded multimodal reasoning (§3.3.3).

Together, these developments reflect a broader transition from reactive to deliberative reasoning paradigms, bringing LMRMs closer to achieving adaptive, system-level intelligence in open and dynamic environments.

Stage 4: Towards Native Large Multimodal Reasoning Model (Prospect)

While LMRMs show promise in addressing complex tasks through extended chains of thought, their language-centric architectures impose critical constraints (Kumar et al., 2025; Pfister & Jud, 2025). First, their predominant focus on vision and language modalities (e.g., text, images, videos) limits their applicability in real-world settings, where diverse data types, such as audio, tactile signals, sensor streams, and temporal sequences, are deeply intertwined. Language-generated reasoning alone often struggles to support multimodal generative thinking, reflection, and control. Second, current models exhibit deficiencies in interactive, long-horizon reasoning and adaptive planning. Although they can produce extended reasoning chains in static settings, their ability to engage in real-time, iterative interaction with dynamic environments remains underdeveloped.

To address these gaps, we prospect the development of **native large multimodal reasoning models (N-LMRMs) as a potential paradigm shift in machine intelligence** (§4). In contrast to conventional LMRMs, which retrofit language models with auxiliary modality processors, N-LMRMs will be natively designed to unify multimodal understanding, generation, and agentic reasoning within a fully end-to-end architecture. Real-world data types are encoded within a unified representation space, like VideoPoet (Kondratyuk et al., 2024), while large-scale synthetic data facilitates holistic learning of reasoning and planning in the environment of any modality interaction. This evolution hinges on two transformative capabilities: *1) Multimodal Agentic Reasoning*: N-LMRMs will embody agentic intelligence, enabling proactive, goal-driven interactions with complex environments, such as long-horizon planning—hierarchical task breakdown and memory-enhanced reasoning for coherence in extended interactions; dynamic adaptation—real-time strategy adjustment based on environmental feedback; embodied learning—closed-loop training frameworks enabling models to learn through simulated or physical interactions for better generalization. *2) Omni-Modal Understanding and Generative Reasoning*: N-LMRMs will move beyond modality-specific encoders and decoders by utilizing a unified representational space for smooth cross-modal synthesis and analysis. This approach includes heterogeneous data fusion for the joint embedding of diverse data types, contextual multimodal generation for the coherent creation of composite outputs, and modality-agnostic inference that enables adaptable processing pipelines for the task-agnostic handling of new or any cross-modal data.

Taken together, the evolution from modular perception-driven systems to emerging native multimodal reasoners outlines a clear trajectory toward more unifying, adaptive, comprehensive high-level AI systems. In the following sections, we provide a detailed analysis of each stage, its representative models, and the emerging research directions that shape the future of multimodal reasoning.

3 Roadmap of Multimodal Reasoning Models

3.1 Stage 1 Perception Driven Modular Reasoning - Developing Task-Specific Reasoning Modules

In the early stages of multimodal reasoning, constraints such as limited multimodal data, nascent neural network architectures, and less sophisticated learning methods led to the development of models tailored to specific tasks. These models typically employed distinct modules to achieve multimodal representation, alignment, fusion, and reasoning. According to the model architectures and learning approaches, these models can be summarized as modular reasoning networks and pretrained Vision-Language Models (VLMs) based modular reasoning.

3.1.1 Modular Reasoning Networks

Initial approaches relied on generic CNN and LSTM backbones to derive answers from multimodal data. However, these were quickly improved by architectures that modularized reasoning based on perceptual cues. Neural Module Networks (NMN) (Andreas et al., 2016) dynamically assembled task-specific modules to compose visual and textual features, replacing static fusion. Hierarchical Co-Attention (HieCoAtt) (Lu et al., 2016) introduced modular cross-modal attention to align question semantics with image regions hierarchically. Multimodal Compact Bilinear Pooling (MCB) (Fukui et al., 2016) optimized feature interactions through efficient learnable bilinear modules. Stacked Attention Networks (SANs) (Yang et al., 2016) modularized reasoning via iterative attention hops over visual features. Dynamic Memory Networks (DMN) (Xiong et al., 2016) integrated memory modules for multi-episode reasoning over sequential inputs. ReasonNet (Ilievski & Feng, 2017) decomposed reasoning into entity-relation modules for structured inference. UpDn (Anderson et al., 2018) introduced bottom-up and top-down attention to prioritize object-level features for reasoning (e.g., VQA-v2). MAC (Hudson & Manning, 2018) employed a memory-augmented control unit for iterative compositional reasoning. BAN (Kim et al., 2018) captured high-order interactions using bilinear attention networks across modalities. Heterogeneous Memory Enhanced Multimodal Attention, HeteroMemory (Fan et al., 2019) extended modularity to video by synchronizing appearance and motion modules with temporal fusion. MuRel (Cadene et al., 2019) modeled reasoning as a relational network over object pairs for fine-grained inference. MCAN (Yu et al., 2019b) used modular co-attention with self- and guided-attention for deep cross-modal reasoning.

These advancements illustrate how perception-driven designs - incorporating attention mechanisms, memory components, and compositional modules - facilitate fine-grained reasoning that is aligned with specific task

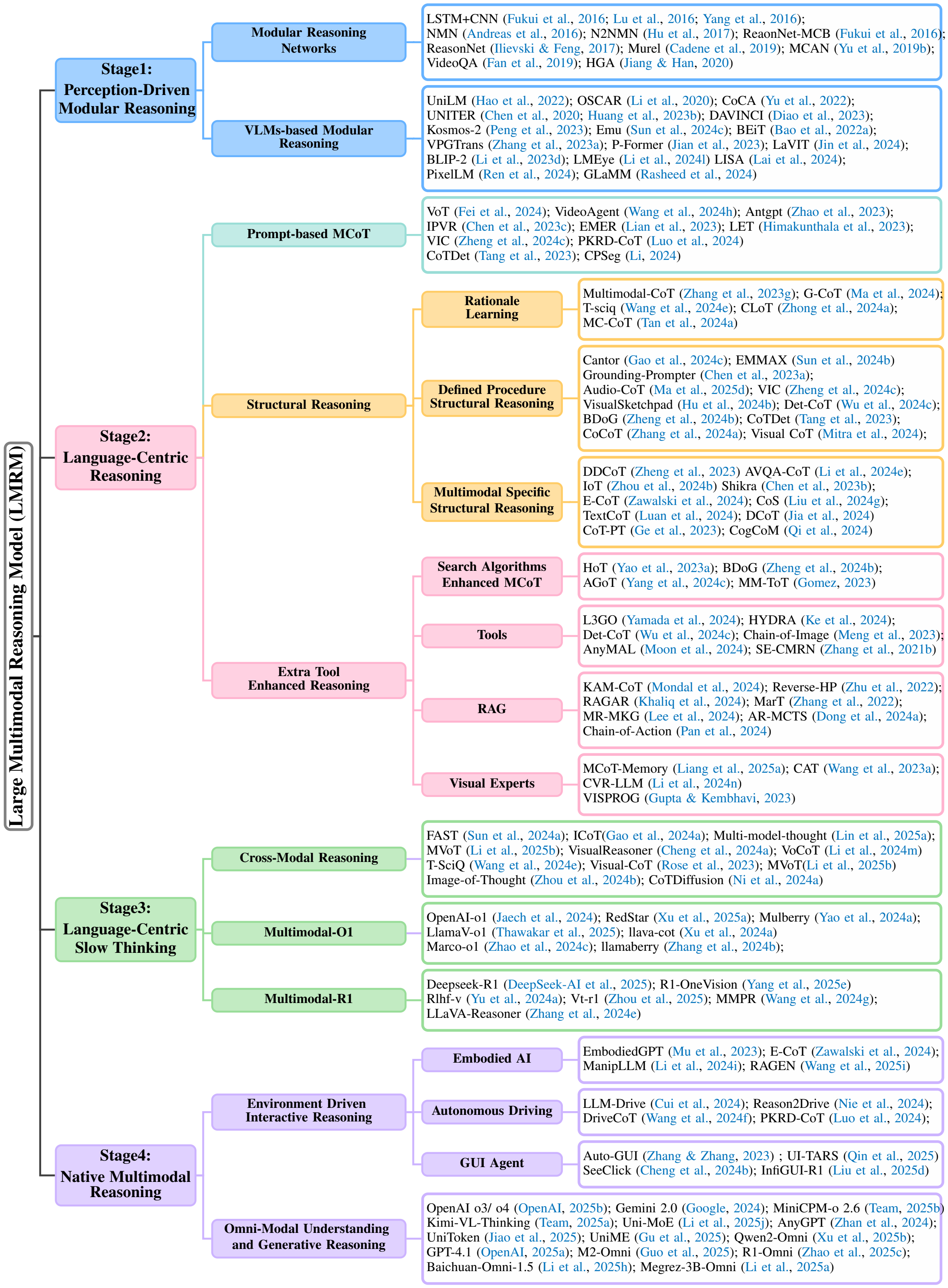

Figure 3: Taxonomy of Large Multimodal Reasoning Models.

requirements. However, the advent of Transformer (Vaswani et al., 2017) architecture, coupled with pretraining-finetuning learning schemes, has propelled multimodal representation, alignment, and fusion. Specifically, Trasnformer-based pretrained VLMs enhance the integration of visual and textual information at the data and model interior, thus enabling perception-driven reasoning capabilities.

3.1.2 Vision-Language Models-based Modular Reasoning

These VLMs are trained with large-scale image-text pairs, advancing perception-driven reasoning tasks, like NLVR^2 (Suhr et al., 2018), TVQA (Lei et al., 2018), GQA (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), VCR (Zellers et al., 2019), and ScienceQA (Saikh et al., 2022). Specifically, VLMs introduced Transformer and employed large-scale image-text data to unify the process of multimodal representation, perception, fusion, and inference. Below are three kinds of pretrained VLMs-based modular reasoning:

Dual-Encoder Contrastive Reasoning. These models leverage dual-stream architectures with contrastive learning to dynamically align and reason over visual and textual features through cross-modal interactions. For example, ViLBERT (Lu et al., 2019) uses dual-stream Transformers with cross-modal attention for dynamic feature alignment. LXMERT (Tan & Bansal, 2019) adds interaction layers between dual encoders to reason over relational embeddings. CLIP (Radford et al., 2021) leverages contrastive pretraining for zero-shot reasoning via aligned embeddings. ALBEF (Li et al., 2021b) integrates contrastive learning with momentum distillation to reason over distilled embeddings. METER (Dou et al., 2022) enhances dual-stream reasoning with a modular encoder-decoder framework for robust alignment (e.g., VCR). SimVLM (Wang et al., 2021) uses prefix-based pretraining to align vision and language for efficient reasoning. VLMo (Bao et al., 2022b) introduces a mixture-of-modality-experts framework for flexible cross-modal reasoning. CoCa (Yu et al., 2022) integrates contrastive and generative heads for versatile reasoning (e.g., NLVR^2). BLIP (Li et al., 2022) introduce the image-text transformer module Q-former and employs vision-language pretraining with contrastive objectives to reason via bootstrapped alignment.

Single-Transformer-Backbone Interactive Reasoning. This paradigm embeds visual and textual inputs in a single Transformer, enabling direct cross-modal reasoning through unified encoding method. VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020), VL-BERT (Su et al., 2019) fuse visual-text inputs in a single Transformer to reason via joint contextual encoding or enhanced cross-modal pretraining. PixelBERT (Huang et al., 2020) employs a CNN and Transformer architecture to process pixels for fine-grained reasoning (e.g., NLVR^2). UniVL (Luo et al., 2020) unifies video-language reasoning with a single Transformer for temporal cross-modal tasks (e.g., TVQA). Oscar (Li et al., 2020), VinVL (Zhang et al., 2021a) anchor reasoning with object tags or enhanced visual features in a unified Transformer, boosting semantic inference (e.g., VCR, GQA). ERNIE-ViL (Yu et al., 2021) integrates scene graph knowledge into a single Transformer, enhancing compositional reasoning through structured visual-language interactions. UniT (Hu & Singh, 2021) streamlines multimodal tasks with a shared self-attention backbone for unified reasoning. PaLI (Chen et al., 2022b) scales single-Transformer reasoning with a multilingual framework for cross-lingual inference (e.g., OK-VQA). Flamingo (Alayrac et al., 2022) employs cross-attention to prioritize dynamic vision-text interactions. BEiT-3 (Wang et al., 2022b) adopts masked data modeling to unify vision-language learning. OFA (Wang et al., 2022a), BLIP-2 (Li et al., 2023d) introduce a unified multimodal framework or a querying Transformer to excel in cross-modal reasoning with improved efficiency (e.g., VQA-v2). Kosmos-1 (Huang et al., 2023b), Kosmos-2 (Peng et al., 2023) enable interleaved input processing or grounding capability for flexible multimodal understanding and precise object localization.

Multimodal LLMs-based Implicit Reasoning. This approach projects visual inputs into a large language model's text space, leveraging the contextual reasoning capabilities of large language models (Li et al., 2023e) to improve the performance of multimodal reasoning. Their architecture contains pretrained visual encoders and large language models, arr. *Vision-Encoder-LLM*. CLIP-Cap (Mokady et al., 2021) projects CLIP visual features into an LLM for reasoning and captioning tasks. LLaVA (Liu et al., 2023a) enables conversational reasoning by tuning ViT-LLM integration for interactive tasks or scaling for complex VQA tasks. MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023) align a ViT to a frozen LLM via a projection layer or instruction tuning, streamlining visual-text reasoning. Qwen-VL (Bai et al., 2023) incorporates a spatial-aware ViT, enhancing grounded reasoning for spatially complex tasks. mPLUG-Owl (Ye et al., 2023), LMEye (Li et al., 2024l), and Otter (Li et al., 2023a) integrate a modular visual encoder with an LLM for instruction-following and in-context learning for multimodal reasoning.

While the architectural innovations of these three kinds of models have significantly advanced multimodal reasoning for tasks, their reliance on predefined feature alignments or contextual encodings often limits their ability to handle complex, multi-step reasoning scenarios requiring iterative or compositional inference. These

Table 1: The classific works of the initial stage of perception-driven multimodal modular reasoning, where VLMs and MLLMs play a significant role in advancing the performance of multimodal reasoning tasks.

Model	Year	Architecture	Highlight	Training Method
			Neural Modular Reasoning Networks	
NMN (Andreas et al., 2016)	2016	Modular	Dynamically assembles task-specific modules for visual-textual reasoning.	Supervised learning
HieCoAtt (Lu et al., 2016)	2016	Attention-based	Aligns question semantics with image regions via hierarchical cross-modal attention.	Supervised learning
MCB (Fukui et al., 2016)	2016	Bilinear	Optimizes cross-modal feature interactions with efficient bilinear modules.	Supervised learning
SANs (Yang et al., 2016)	2016	Attention-based	Iteratively refines reasoning through multiple attention hops over visual features.	Supervised learning
DMN (Xiong et al., 2016)	2016	Memory-based	Integrates memory modules for multi-episode reasoning over sequential inputs.	Supervised learning
ReasonNet (Ilievski & Feng, 2017)	2017	Modular	Decomposes reasoning into entity-relation modules for structured inference.	Supervised learning
UpDn (Anderson et al., 2018)	2018	Attention-based	Combines bottom-up and top-down attention for object-level reasoning.	Supervised learning
MAC (Hudson & Manning, 2018)	2018	Memory-based	Uses a memory-augmented control unit for iterative compositional reasoning.	Supervised learning
BAN (Kim et al., 2018)	2018	Bilinear	Captures high-order interactions via bilinear attention across modalities.	Supervised learning
HeteroMemory (Fan et al., 2019)	2019	Memory-based	Synchronizes appearance and motion modules for video-based temporal reasoning.	Supervised learning
MuRel (Cadene et al., 2019)	2019	Relational	Models reasoning as a relational network over object pairs for fine-grained inference.	Supervised learning
MCAN (Yu et al., 2019b)	2019	Attention-based	Employs modular co-attention with self- and guided-attention for deep reasoning.	Supervised learning
			VLMs-based Modular Reasoning	
ViLBERT (Lu et al., 2019)	2019	Dual-Encoder	Aligns visual-text features via dual-stream Transformers with cross-modal attention.	Pretraining + fine-tuning
LXMERT (Tan & Bansal, 2019)	2019	Dual-Encoder	Enhances cross-modal reasoning with dual-stream pretraining on diverse tasks.	Pretraining + fine-tuning
X-LXMERT (Tan & Bansal, 2019)	2020	Dual-Encoder	Extends dual-stream reasoning with generative cross-modal pretraining.	Pretraining + fine-tuning
ALBEF (Li et al., 2021b)	2021	Dual-Encoder	Integrates contrastive learning with momentum distillation for robust reasoning.	Contrastive + generative pretraining
SimVLM (Wang et al., 2021)	2021	Dual-Encoder	Uses prefix-based pretraining for flexible cross-modal reasoning.	Pretraining + fine-tuning
VLMo (Bao et al., 2022b)	2022	Dual-Encoder	Employs a mixture-of-modality-experts for dynamic cross-modal reasoning.	Pretraining + fine-tuning
METER (Dou et al., 2022)	2022	Dual-Encoder	Enhances reasoning with a modular encoder-decoder for robust alignment.	Pretraining + fine-tuning
BLIP (Li et al., 2022)	2022	Dual-Encoder	Bootstraps alignment with contrastive learning for efficient reasoning.	Contrastive + generative pretraining
VisualBERT (Li et al., 2019)	2019	Single-Transformer-Backbone	Fuses visual-text inputs in a single Transformer for joint contextual reasoning.	Pretraining + fine-tuning
VL-BERT (Su et al., 2019)	2019	Single-Transformer-Backbone	Enhances cross-modal reasoning with unified visual-language pretraining.	Pretraining + fine-tuning
UNITER (Chen et al., 2020)	2020	Single-Transformer-Backbone	Reasons via joint contextual encoding in a single Transformer backbone.	Pretraining + fine-tuning
PixelBERT (Huang et al., 2020)	2020	Single-Transformer-Backbone	Processes pixels with CNN+Transformer for fine-grained cross-modal reasoning.	Pretraining + fine-tuning
UniVL (Luo et al., 2020)	2020	Single-Transformer-Backbone	Unifies video-language reasoning with a single Transformer for temporal tasks.	Pretraining + fine-tuning
Oscar (Li et al., 2020)	2020	Single-Transformer-Backbone	Anchors reasoning with object tags in a unified Transformer for semantic inference.	Pretraining + fine-tuning
VinVL (Zhang et al., 2021a)	2021	Single-Transformer-Backbone	Boosts reasoning with enhanced visual features in a single Transformer.	Pretraining + fine-tuning
ERNIE-ViL (Yu et al., 2021)	2021	Single-Transformer-Backbone	Integrates scene graph knowledge for structured visual-language reasoning.	Pretraining + fine-tuning
UniT (Hu & Singh, 2021)	2021	Single-Transformer-Backbone	Streamlines multimodal tasks with a shared self-attention Transformer backbone.	Pretraining + fine-tuning
Flamingo (Alayrac et al., 2022)	2022	Single-Transformer-Backbone	Prioritizes dynamic vision-text interactions via cross-attention.	Pretraining + fine-tuning
CoCa (Yu et al., 2022)	2022	Single-Transformer-Backbone	Combines contrastive and generative heads for versatile cross-modal reasoning.	Contrastive + generative pretraining
BEiT-3 (Wang et al., 2022b)	2022	Single-Transformer-Backbone	Unifies vision-language learning with masked data modeling.	Pretraining + fine-tuning
OFA (Wang et al., 2022a)	2022	Single-Transformer-Backbone	Provides a unified multimodal framework for efficient cross-modal reasoning.	Pretraining + fine-tuning
PaLI (Chen et al., 2022b)	2022	Single-Transformer-Backbone	Scales reasoning with a multilingual single-Transformer framework.	Pretraining + fine-tuning
BLIP-2 (Li et al., 2023d)	2023	Single-Transformer-Backbone	Uses a querying Transformer for improved cross-modal reasoning efficiency.	Pretraining + fine-tuning
Kosmos-1 (Huang et al., 2023b)	2023	Single-Transformer-Backbone	Enables interleaved input processing for flexible multimodal understanding.	Pretraining + fine-tuning
Kosmos-2 (Peng et al., 2023)	2023	Single-Transformer-Backbone	Enhances grounding capability for precise object localization and reasoning.	Pretraining + fine-tuning
CLIP-Cap (Mokady et al., 2021)	2021	Vision-Encoder-LLM	Projects CLIP visual features into an LLM for reasoning and captioning.	Fine-tuning
LLaVA (Liu et al., 2023a)	2023	Vision-Encoder-LLM	Tunes ViT-LLM integration for conversational multimodal reasoning.	Instruction tuning
MiniGPT-4 (Zhu et al., 2023)	2023	Vision-Encoder-LLM	Aligns ViT to a frozen LLM via projection for streamlined reasoning.	Fine-tuning
InstructBLIP (Dai et al., 2023)	2023	Vision-Encoder-LLM	Uses instruction tuning to align ViT with LLM for multimodal reasoning.	Instruction tuning
Qwen-VL (Bai et al., 2023)	2023	Vision-Encoder-LLM	Incorporates spatial-aware ViT for enhanced grounded reasoning.	Pretraining + fine-tuning
mPLUG-Owl (Ye et al., 2023)	2023	Vision-Encoder-LLM	Integrates modular visual encoder with LLM for instruction-following reasoning.	Instruction tuning
Otter (Li et al., 2023a)	2023	Vision-Encoder-LLM	Combines modular visual encoder with LLM for in-context multimodal reasoning.	Instruction tuning

constraints highlight the need for Multimodal Chain-of-Thought (MCoT) reasoning (Sec. 3.2) in large-scale models like the development of LLMs, which can dynamically decompose tasks, integrate intermediate reasoning steps, and adaptively align perception and inference for more robust and generalizable performance across diverse multimodal challenges.

Takeaways: Perception-Driven Modular Reasoning

Early multimodal models primarily focused on the representation, alignment, and fusion of information. Reasoning in these models was often implicit, typically requiring separate, task-specific reasoning modules. More recently, multimodal large language models, particularly those adopting a vision encoder-language model structure, have achieved a unified multimodal reasoning architecture and demonstrated improved multi-task reasoning performance.

3.2 Stage 2 Language-Centric Short Reasoning - System-1 Reasoning

With the advent of large-scale multimodal pretraining, MLLMs have started to demonstrate emergent reasoning capabilities. However, such inferences are often shallow, relying primarily on implicit correlations rather than explicit logical processes. To mitigate this limitation, MCoT has emerged as a simple yet effective approach. By incorporating intermediate reasoning steps, MCoT improves cross-modal alignment, knowledge integration, and contextual grounding, all without the need for extensive supervision or significant architectural modifications. In this stage, we categorize existing approaches into three paradigms: prompt-based MCoT, structural reasoning with predefined patterns, and tool-augmented reasoning leveraging lightweight external modules.

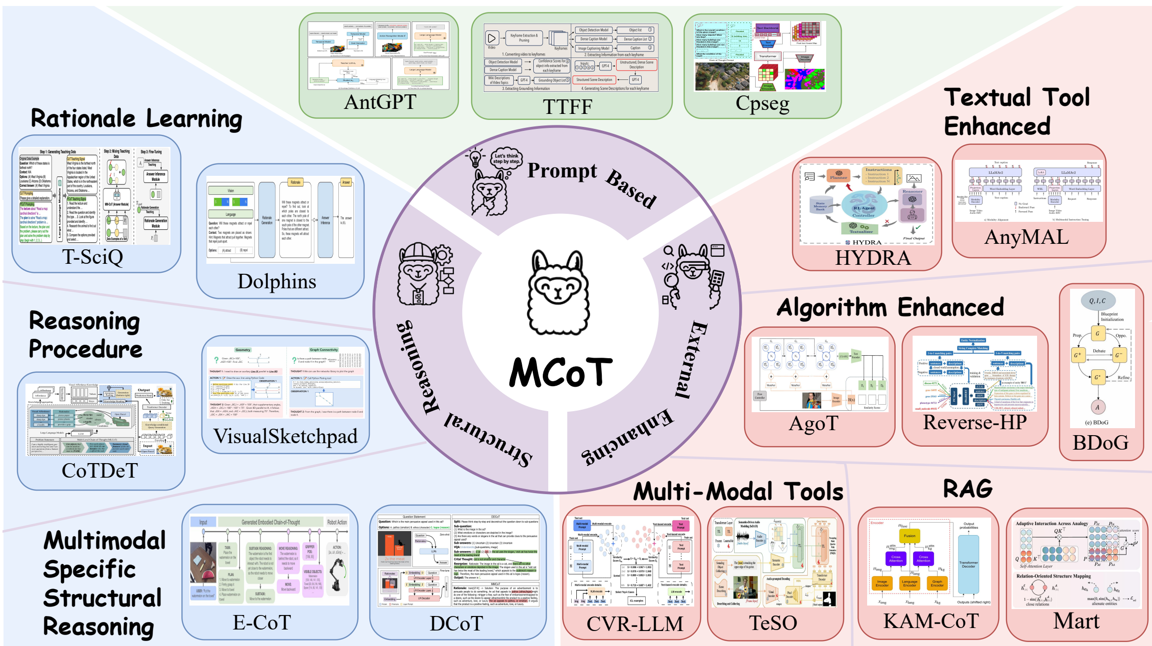

Figure 4: Taxonomy and representative methods of structural reasoning in multimodal chain-of-thought.

3.2.1 Prompt-based MCoT

Prompt-based Multimodal Chain-of-Thought (MCoT) methods extend the textual CoT paradigm to multimodal contexts, enabling step-by-step reasoning across modalities with strong interpretability and minimal additional training. In visual reasoning, IPVR (Chen et al., 2023c) proposed a structured "see-think-confirm" prompting framework that guides LLMs through visual grounding and rationale verification. VIC (Zheng et al., 2024c) prompts textual reasoning chains before visual input to mitigate hallucinations and improve accuracy.

For video understanding, VoT (Fei et al., 2024) leverages spatial-temporal scene graphs to prompt progressive reasoning from low-level perception to high-level interpretation. VideoAgent (Wang et al., 2024h) is an LLM-coordinated system that iteratively prompts key information from long videos with minimal frame usage. LET (Himakunthala et al., 2023) employs a frame-by-frame prompting strategy on the VIP dataset to guide temporal reasoning for video infilling and prediction.

In domain-specific applications, PKRD-CoT (Luo et al., 2024) introduces a zero-shot prompting framework that structures autonomous driving reasoning across perception, knowledge, reasoning, and decision-making. LPE (Xie et al., 2025a) uses prompt-based reasoning on spoken content and emotional cues to generate empathetic responses. EMER (Lian et al., 2023) applies prompting in multimodal emotion recognition to integrate unimodal clues and produce interpretable predictions.

Task-oriented reasoning has also benefited from prompt-based MCoT. CoTDet (Tang et al., 2023) uses multi-level prompting to extract affordance knowledge for object detection. AntGPT (Zhao et al., 2023) prompts LLMs to infer human goals and temporal dynamics from video-based action sequences. CPSeg (Li, 2024) formulates chain-of-thought prompts to align textual and pixel-level semantics for enhanced segmentation.

3.2.2 Structural Reasoning

Unlike prompt-based MCoT methods, which induce reasoning behaviour through handcrafted exemplars or a zero-shot prompting approach, structural reasoning focuses on learning reasoning patterns through supervised training. By integrating explicit procedural structures into the model, these approaches transform loosely guided reasoning into standardized, stage-wise processes, improving scalability, reliability, and efficiency in complex multimodal tasks. We categorize structural reasoning into three representative types: (i) *rationale construction*, which learns to produce atomic reasoning steps as interpretable scaffolds; (ii) *defined reasoning procedures*, which adapt structured texture reasoning schemes to multimodal settings; and (iii) *modality-specific structural reasoning*, which further incorporates modality-aware constraints and designs to better align with the characteristics of visual, auditory, or embodied inputs.

Rationale Construction The foundation of structural reasoning in multimodal contexts begins with effective rationale learning approaches. Multimodal-CoT (Zhang et al., 2023g) proposes a two-stage Multimodal-CoT framework that decouples rationale generation from answer prediction to reduce hallucinations. T-sciq (Wang et al., 2024e) leverages teacher LLMs to generate rationales with varying complexity, showing rationale quality is key to reasoning accuracy. In autonomous driving, G-CoT (Ma et al., 2024) designs Dolphins explicitly linking rationales to visual and historical driving signals for more grounded reasoning. MC-CoT (Tan et al., 2024a) uses a self-consistency strategy to select the most accurate rationale among multiple candidates, boosting smaller models' performance. CLoT (Zhong et al., 2024a) promotes non-linear, explorative rationale construction via Leap-of-Thought to support creative reasoning.

Defined Reasoning Procedure In the realm of enhancing the interpretability of text reasoning processes, numerous studies have proposed structured reasoning stages. Cantor (Gao et al., 2024c), for instance, differentiates between perception and decision-making stages. In the perception stage, low-level attributes such as objects, colours, and shapes are extracted from images or textual descriptions, followed by the decision-making stage that integrates these features for problem-solving. TextCoT (Luan et al., 2024) adopts a three-phase process. The Image Overview stage generates a global description, the Coarse Localization stage pinpoints the answer region using the grounding ability of LMMs, and the Fine-grained Observation stage combines global and local details for accurate answers. Similarly, Grounding-Prompter (Chen et al., 2023a) conducts global understanding, noise evaluation, partition understanding, and prediction. It gradually merges global and local semantics, resists noise, and improves the perception of temporal boundaries. Audio-CoT (Ma et al., 2025d) utilizes three chain-of-thought reasoning paradigms. Manual-CoT depends on handcrafted examples for reasoning guidance, Zero-Shot-CoT achieves zero-shot reasoning with simple prompts, and Desp-CoT facilitates reasoning by generating audio descriptions. VIC (Zheng et al., 2024c) breaks tasks into text-based sub-steps before integrating visual inputs to form final rationales. Visual Sketchpad (Hu et al., 2024b) organizes rationales into thought, action and observation phases during the sketching process. Det-CoT (Wu et al., 2024c) formalizes VQA reasoning as a combination of subtasks and reviews. BDoG (Zheng et al., 2024b) utilizes a dedicated debate and summarization pipeline with unique agents. CoTDet (Tang et al., 2023) achieves object detection via human-like procedure of listing, analyzing and summarizing. CoCoT (Zhang et al., 2024a) systematically compares input similarities and differences. SegPref (Wang et al., 2024j) localizes sounding objects accurately in the visual space through global understanding, sounding object filtering, and noise removal. EMMAX (Sun et al., 2024b) combines grounded planning approaches with predictive movement techniques.

Multimodal Specific Structural Reasoning Recent research has introduced modality-specific reasoning structures tailored to the unique challenges of multimodal inputs, particularly in vision-language tasks. A prominent line of work focuses on region-based grounding, where spatial localization is used to guide structured reasoning. For instance, CoS (Liu et al., 2024g) and TextCoT (Luan et al., 2024) adopt a two-stage pipeline that first identifies regions of interest conditioned on the input question, followed by localized inspection to enable multi-granular reasoning without resolution loss. DCoT (Jia et al., 2024) extends this paradigm by introducing a dual-guidance mechanism that combines bounding-box grounding with the retrieval of semantically similar examples, jointly enhancing fine-grained and context-aware reasoning. Beyond spatial grounding, CoT-PT (Ge et al., 2023) integrates visual and textual embeddings through prompt tuning and gradually refines visual concept representations via coarse-to-fine abstraction.

Another class of approaches focuses on text-guided semantic enrichment. Shikra (Chen et al., 2023b) and TextCoT (Luan et al., 2024) leverage image captions as high-level semantic cues to guide spatial attention and object grounding. This strategy reduces dependence on external detection modules and facilitates more interpretable referential reasoning. Inspired by classical CoT frameworks, DDCoT (Zheng et al., 2023) and

AVQA-CoT (Li et al., 2024e) decompose complex visual or audio-visual queries into sequential sub-questions, enabling compositional reasoning and improved multi-hop inference across modalities.

Finally, E-CoT (Zawalski et al., 2024) extends structured reasoning to embodied scenarios by interleaving task rephrasing, planning, and low-level action execution. This highlights the necessity of reasoning chains that span both semantic and sensorimotor levels in vision-language-action models.

Table 2: The Structural Reasoning, which transforms loosely guided reasoning into standardized, step-by-step processes by integrating explicit procedural structures into the model, enhancing scalability, reliability, and efficiency in complex multimodal tasks.

Name	Modality	Task	Reasoning Structure	Datasets	Highlight
Cantor (2024c)	T,I	VQA	Perception, Decision	-	Decouples perception and reasoning via feature extraction and CoT-style integration.
TextCoT(2024)	T,I	VQA	Caption, Localization, Precise observation	-	first summarizes visual context, then generates CoT-based responses.
Grounding-Prompter(2023a)	T,V,A	Temporal Sentence Grounding	Denoising	VidChapters-7M	Grounding-Prompter performs global parsing, denoising, partitioning before reasoning.
Audio-CoT(2025d)	T,A	AQA	Manual-CoT, Zero-Shot-CoT, Desp-CoT	-	Enhances visual reasoning by utilizing three chain-of-thought paradigms.
VIC(2024c)	I,T	VQA	Thinking before looking	-	Breaks tasks into text-based sub-steps before integrating visual inputs to form final rationales
Visual Sketch-pad(2024b)	I,T	VQA, math QA	Sketch-based reasoning paradigm	-	Organizes rationales into Thought-Action-Observation phases.
Det-CoT(2024c)	I,T	VQA	Subtask decomposition, Execution, and Verification	-	Formalizes VQA reasoning as a combination of subtasks and reviews.
BDoG(2024b)	I,T	VQA	Entity update, Relation update, Graph pruning	-	Utilizes a dedicated debate and summarization pipeline with unique agents.
CoTDet(2023)	I,T	object detection	Object listing, Affordance analysis, Visual feature summarization	COCO-Tasks	Achieves object detection via human-like procedure of listing, analyzing and summarizing.
CoCoT(2024a)	I,T	VQA	Contrastive prompting strategy	-	Systematically contrasts input similarities and differences.
SegPref(2024j)	T,A,V	Temporal Sentence Grounding	Visual summary, Sound filtering, Denoising	Youtube-8M, Semantic-ADE20K	Localizes sounding objects accurately in the visual space through global understanding, sounding object filtering, and noise removal.
EMMAX(2024b)	I,T	Robotic task	Grounded CoT reasoning, Look-ahead spatial reasoning	Dataset based on BridgeV2	Combines grounded planning approaches with predictive movement techniques.
DDCoT (2023)	T,I	VQA	Question Deconstruct,Rationale	ScienceQA	Maintains a critical attitude by identifying reasoning and recognition responsibilities through the combined effect of negative-space design and visual deconstruction.
AVQA-CoT (2024e)	T,A,V	AVQA	Question Deconstruct , Question Selection , Rationale	MUSIC-AVQA	Decomposes complex questions into multiple simpler sub-questions and leverages LLMs to select relevant sub-questions for audio-visual question answering.
CoT-PT (2023)	T,I	Image Classification, Image-Text Retrieval , VQA	Coarse-to-Fine Image Concept Representation	ImageNet	First to successfully adapt CoT for prompt tuning by combining visual and textual embeddings in the vision domain.
IoT (2024b)	T,I	VQA	Visual Action Selection , Execution , Rationale , Summary , Self-Refine	-	Enhances visual reasoning by integrating visual and textual rationales through a model-driven multimodal reasoning chain.
Shikra (2023b)	T,I	VQA , PointQA	Caption , Object Grounding	ScienceQA	Maintains a critical attitude by identifying reasoning and recognition responsibilities through the combined effect of negative-space design and visual deconstruction.
E-CoT (2024)	T,I,A	Policies' Generalization	Task Rephrase , Planning , Task Deconstruct , Object Grounding	Bidgedata v2	Integrates semantic planning with low-level perceptual and motor reasoning, advancing task formulations in embodied intelligence.
CoS (2024g)	T,I	VQA	Object Grounding , Rationale	Llava665K	Guides the model to identify and focus on key image regions relevant to a question, enabling multi-granularity understanding without compromising resolution.
TextCoT (2024)	T,I	VQA	Caption , Object Grounding , Image Zoom	Llava665K , SharedGPT4V	Enables accurate and interpretable multimodal question answering through staged processing: overview, coarse localization, and fine-grained observation.
DCoT (2024)	T,I	VQA	Object Grounding , Fine-Grained Image Generation , Similar Example Retrieve , Rationale	-	Uses a dual-guidance mechanism by combining bounding box cues to focus attention on relevant image regions and retrieving the most suitable examples from a curated demonstration cluster as contextual support.

Takeaways: Structural Reasoning

Structural reasoning methods define standardized reasoning workflows by integrating modular subtasks such as question deconstruct, visual grounding, caption generation, summary, phases, and image procession. These approaches enhance interpretability and consistency by organizing generation task into explicit stages. Recent trends also incorporate modality-aware designs to better align reasoning with visual, auditory, or embodied inputs.

3.2.3 Externally Augmented Reasoning

Externally augmented reasoning introduces advantage algorithm, auxiliary tools or expert modules to compensate for limitations in the model's inherent reasoning capacity. These components are integrated at inference time or coupled during training, enabling more flexible, scalable, and task-specialized reasoning workflows. By decoupling core reasoning steps from the base model, such methods support long-horizon planning, precise grounding, and access to dynamic or domain-specific information. We group externally augmented methods into four categories: (i) *search algorithm-enhanced MCoT*, which navigates reasoning spaces via various search algorithm; (ii) *tool-based augmentation*, which leverages external language tools or systems to guide reasoning execution; (iii) *retrieval-augmented reasoning*, which incorporates relevant multimodal knowledge from external sources into the reasoning path; and (iv) *multimodal enhancing*, which incorporate specialized multimodal modules to support perception-driven reasoning.

Search Algorithm Enhanced MCoT Search strategy-driven MCoT approaches empower models to dynamically navigate and optimize reasoning trajectories throughout the reasoning process. MM-ToT (Gomez, 2023), for instance, leverages GPT-4 and Stable Diffusion, employing depth-first search (DFS) and breadth-first search (BFS) algorithms to identify the most optimal multimodal outputs according to a 0.0–1.0 metric scale. HoT (Yao et al., 2023a) creates interconnected thoughts from multimodal inputs and packages them into a single hyperedge. Unlike this, Aggregation Graph-of-Thought (AGoT) (Yang et al., 2024c) builds a reasoning aggregation graph, which integrates diverse reasoning elements at every step and subsequently incorporates visual data. Blueprint Debate on Graph (BDoG) (Zheng et al., 2024b) takes a distinctive route, discarding search algorithms and instead utilizing three agents—an affirmative debater, a negative debater, and a moderator. These agents engage in iterative debates to address multimodal questions, with the moderator ultimately synthesizing a final answer, thus implicitly constructing a graph-of-thought that explores and aggregates a wide range of thoughts. Overall, compared to prompt-based methods that rely on linear, example-driven inference, search strategy-oriented MCoT variants enable models to explore multiple reasoning pathways, thereby significantly enhancing adaptability and the depth of problem-solving.

Textural Tools To enhance the reasoning capabilities of multimodal Chain-of-Thought (MCoT) frameworks, some works incorporate external textual-enhancing tools that guide, structure, or refine the overall reasoning process through language. L3GO (Yamada et al., 2024) employs GPT-4 with Chain-of-Thought prompting to produce explicit textual reasoning steps, which guide 3D mesh construction in a Blender environment, aided by ControlNet for visual grounding. HYDRA (Ke et al., 2024) and Det-CoT (Wu et al., 2024c) leverage large language models not only as planners, but also as dynamic instruction generators, error diagnosers, and reasoning controllers. These models interact with visual foundation models (e.g., BLIP2, LLaVA) and reinforcement learning agents, while using textual prompts and feedback to iteratively improve visual understanding and decision-making. Both systems integrate a State Memory Bank to maintain dialogue history or prior instructions, enabling incremental CoT reasoning via textual modulation. Chain-of-Image (Meng et al., 2023) introduces SyMLLM, which generates intermediate images from language descriptions, turning complex problems into visual reasoning tasks—yet still grounded in language-based control. Similarly, AnyMAL (Moon et al., 2024) unifies diverse modalities into a textual space for cross-modal reasoning, while SE-CMRN (Zhang et al., 2021b) utilizes syntactic cues via GCNs to improve performance in visual commonsense reasoning.

RAG Several approaches enhance multimodal reasoning through retrieval mechanisms, e.g., solving online questions (Chen et al., 2024k). RAGAR (Khaliq et al., 2024) proposed CoRAG and ToRAG to support political fact-checking through retrieval of multimodal evidence. Chain-of-Action (Pan et al., 2024) retrieves information from heterogeneous sources through configurable reasoning chains. KAM-CoT (Mondal et al., 2024) incorporates Knowledge Graphs as external knowledge sources to augment multimodal reasoning. AR-MCTS (Dong et al., 2024a) integrates dynamic step-wise retrieval with Monte Carlo Tree Search, enabling MLLMs to access relevant knowledge at each reasoning step and automatically generate high-quality reasoning

Table 3: Externally Augmented Reasoning, which enhances a model's reasoning by incorporating external resources like algorithms, tools, or expert modules to overcome its inherent limitations.

Name	Modality	Task	Enhancement Type	External Source	Highlight
MM-ToT (2023)	T,I	Image Generation	Search Algorithm	DFS,BFS	Applies DFS and BFS to select optimal outputs.
HoT (2023a)	T,I	VQA	Search Algorithm	multi-hop random walks on graph	Generates linked thoughts from multimodal data in a hyperedge.
AGoT (2024c)	T,I	Text-Image Retrieval, VQA	Search Algorithm	prompt aggregation and prompt flow operations	Builds a graph to aggregate multi-faceted reasoning with visuals.
BDoG (2024b)	T,I	VQA	Search Algorithm	Graph Condensation: Entity update, Relation update, Graph pruning	Effective three-agent debate forms thought graph for multimodal queries.
L3GO (2024)	T,I	3D Object Generation & Composition	Tools	Blender, ControlNet	Iterative part-based 3D construction through LLM reasoning in a simulation environment.
HYDRA (2024)	T,I	Knowledge-QA, Visual Grounding	Tools	RL agent controller, Visual Foundation Models	RL agent controls multi-stage visual reasoning through dynamic instruction selection.
Det-CoT (2024c)	T,I	object detection	Tools	Visual Processing Prompts	Visual prompts guide MLLM attention for structured detection reasoning.
Chain-of-Image (2023)	T,I	Geometric, chess & commonsense reasoning	Tools	Chain of Images prompting	Generates intermediate images during reasoning for visual pattern recognition.
AnyMAL (2024)	T, I, A, V	Cross-modal reasoning, multimodal QA	Tools	Pre-trained alignment module	Efficient integration of diverse modalities; strong reasoning via LLaMA-2 backend.
SE-CMRN (2021b)	T,I	Visual Commonsense Reasoning	Tools	Syntactic Graph Convolutional Network	Enhances language-guided visual reasoning via syntactic GCN in a dual-branch network.
RAGAR (2024)	T,I	Political Fact-Checking	RAG	DuckDuckGo & SerpAPI	Integrates MLLMs with retrieval-augmented reasoning to verify facts using text and image evidence.
Chain-of-action (2024)	T,I	Info retrieval	RAG	Google Search, ChromaDB	Decomposes questions into reasoning chains with configurable retrieval actions to resolve conflicts between knowledge sources.
KAM-CoT (2024)	T,I, KG	Educational science reasoning	RAG	ConceptNet knowledge graph	Enhances reasoning by retrieving structured knowledge from graphs and integrating it through two-stage training.
AR-MCTS (2024a)	T,I	Multi-step reasoning	RAG	Contriever, CLIP dual-stream	Step-wise retrieval with Monte Carlo Tree Search for verified reasoning.
MR-MKG (2024)	T, I	General multimodal reasoning	RAG	RGAT	Enhances multimodal reasoning by integrating information from multimodal knowledge graphs.
Reverse-HP (2022)	T, I	Disease-related reasoning	RAG	reverse hyperplane projection	Utilizes KG embeddings to enhance reasoning for specific diseases with multimodal data.
MarT (2022)	T, I	Analogical reasoning	RAG	Structure-guided relation transfer	Uses structure mapping theory and relation-oriented transfer for analogical reasoning with KG.
MCoT-Memory (2025a)	T,I	VQA	Multimodal Information Enhancing	LLAVA	Memory framework and scene graph construction for effective long-horizon task planning
MGCoT (2023c)	T,I	VQA	Multimodal Embedding Enhancing	ViT-large encoder	Precise visual feature extraction aiding multimodal reasoning
CCoT (2024)	T,I	VQA	Multimodal Perception Enhancing	Scene Graphs	Utilization of the generated scene graph as an intermediate reasoning step.
CVR-LLM (2024n)	T,I	VQA	Multimodal Embedding Enhancing	BLIP2flant5 & BLIP2 multi-embedding	Precise context-aware image descriptions through iterative self-refinement and effective text-multimodal factors integrations
CAT (2023a)	T,I	Image Captioning	Multimodal Perception Enhancing	SAM	Promising pre-trained image caption generators, SAM, and instruction-tuned large language models integration

annotations. Knowledge graph integration has further expanded multimodal reasoning capabilities through diverse approaches: MR-MKG (Lee et al., 2024) enhances general multimodal reasoning by retrieving relevant triples from MMKGs via RGAT, Reverse-HP (Zhu et al., 2022) enables disease-related reasoning using reverse hyperplane projection on SDKG-11, and MarT (Zhang et al., 2022) employs structure mapping theory for multimodal analogical reasoning through relation-oriented transfer between entities in MarKG.

Multimodal Tools Using visual experts is another effective way to enhance the capabilities of models for multimodal reasoning. MCoT-Memory (Liang et al., 2025a) improves long-horizon planning by incorporating memory retrieval and scene graph updates, retaining high-confidence experiences for robust decision-making. MGCoT (Yao et al., 2023c) uses the ViT-large encoder (for multimodal tasks) to extract visual features, the Stanford CoreNLP system for coreference resolution, and the OpenIE system to extract thought unit nodes, thus enabling efficient GoT reasoning. CCoT (Mitra et al., 2024) enhances the compositional visual understanding and multimodal reasoning capabilities of LMMs through two key steps: scene graph generation and response generation. It utilizes the generated scene graph as an intermediate reasoning step. CVR-LLM (Li et al., 2024n) includes two key components: CaID generates context-aware image descriptions through iterative self-refinement, and CVR-ICL innovatively integrates text and multimodal factors to select context examples, enhancing the performance of LLMs in complex visual reasoning tasks. CAT (Wang et al., 2023a) integrates pre-trained image caption generators, SAM, and instruction-tuned large language models. Through visual controls and language controls, it realizes user-centered image description. VISPROG (Gupta & Kembhavi, 2023) iterates alternately through three steps: initial generation, feedback, and refinement. It utilizes a suitable language model and three prompts and based on few-shot prompting, guides the model to generate feedback and refine the output until the stopping condition is met.

Takeaways: Externally Augmented Reasoning

Externally augmented reasoning introduces auxiliary modules (such as search algorithms, tool agents, retrieval systems, and specialized multimodal processors) to assist or offload parts of the reasoning process. These methods enable more controllable, scalable, and task-adaptive reasoning by decoupling planning, grounding, or perception tasks from the backbone model, often enhancing long-horizon reasoning and domain specialization.

3.3 Stage 3 Language-Centric Long Reasoning - System-2 Thinking and Planning

While structural reasoning introduces predefined patterns to guide MLLMs toward more systematic reasoning, it remains constrained by shallow reasoning depth and limited adaptability. To handle more complex multimodal tasks, recent work aims to develop System-2-style reasoning (Kahneman, 2011). Unlike fast and reactive strategies, this form of reasoning is deliberate, compositional, and guided by explicit planning. By extending reasoning chains, grounding them in multimodal inputs, and training with supervised or reinforcement signals, these models begin to exhibit long-horizon reasoning and adaptive problem decomposition.

3.3.1 Cross-Modal Reasoning

Cross-Modal Reasoning refers to the ability to integrate and reason across multiple modalities, such as text, images, videos. Recent advancements in cross-modal reasoning have emphasized the importance of augmenting multimodal information beyond textual inputs through model-intrinsic capabilities or external tools and algorithms. These methods aim to enhance reasoning accuracy and robustness by dynamically incorporating complementary information from diverse modalities.

External Tools Beyond the use of external tools for multimodal understanding described in §3.2.3, recent approaches increasingly explore tool integration as a vehicle for multimodal reasoning itself. VisProg (Gupta & Kembhavi, 2023) and ProViQ (Choudhury et al., 2024) leverage program generation and procedural execution to enable cross-modal reasoning, dynamically generating executable code or logic paths to solve complex tasks such as video question answering, multi-step visual reasoning, and geometric problem solving. In parallel, methods such as AssistGPT (Gao et al., 2023), MM-ReAct (Yang et al., 2023), and Multi-Modal-Thought (Lin et al., 2025a) adopt modular integration frameworks—such as PEIL and vision expert prompting—to coordinate tool use based on reasoning progression. These systems enable interpretable and adaptive reasoning by calling different tools dynamically during task execution. VisualReasoner (Cheng et al., 2024a) further introduces a data synthesis strategy to generate multi-step reasoning traces, which are then used to train a plug-and-play visual reasoning module applicable to a variety of vision-language backbones. Collectively, these efforts

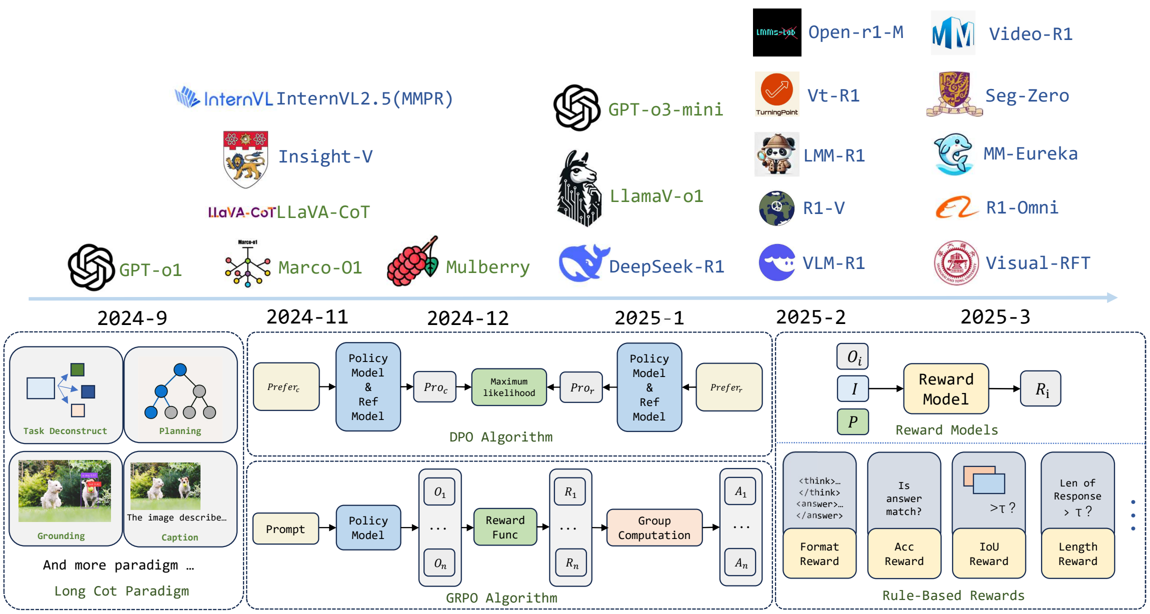

Figure 5: Timeline (top) and core components (bottom) of recent multimodal O1-like and R1-like models. The top part illustrates the chronological emergence of representative models. The bottom part summarizes key components including structured reasoning paradigms, reinforcement learning algorithms (e.g., DPO and GRPO), and the design of rule-based reward models.

extend the landscape of multimodal reasoning by combining program induction, dynamic tool orchestration, and data-driven reasoning supervision.

External Algorithms FAST (Sun et al., 2024a) and ICoT (Gao et al., 2024a) both leverage cognitive processes akin to human thinking, with FAST employing a system switch adapter to dynamically alternate between fast and slow thinking modes, while ICoT utilizes Attention-driven Selection (ADS) to interleave visual and textual reasoning steps. Meanwhile, Image-of-Thought (Zhou et al., 2024b) and CoTDiffusion (Ni et al., 2024a) focus on generating visual rationales, with Image-of-Thought extracting visual info step-by-step and CoTDiffusion creating visual subgoal plans, extending algorithmic augmentation to robotics.

Model-Intrinsic Capabilities These approaches rely on the LMM's inherent ability to generate or infer multimodal information without external tools. T-SciQ (Wang et al., 2024e), Visual-CoT (Rose et al., 2023) and VoCoT (Li et al., 2024m) demonstrated how fine-tuning LMMs on carefully designed CoT datasets (e.g., VoCoT-Instruct80K) could enable single-step multimodal reasoning in charts, documents, and geometry problems. MVoT (Li et al., 2025b) represents an early effort, where a self-contained architecture iteratively refines visual-textual representations for embodied reasoning tasks.

Takeaways: Cross-Modal Reasoning

Cross-modal reasoning methods enhance multimodal inference by integrating visual, auditory, and programmatic cues across modalities. Representative strategies include leveraging external tools, algorithmic control for interleaving modality-specific steps, and model-intrinsic fusion of multimodal representations, enabling more grounded, interpretable, and robust reasoning in open-ended tasks.

3.3.2 Multimodal-O1

With the rise of OpenAI o1, which sparked widespread interest in large reasoning models, open-source reproductions such as Marco-o1 (Zhao et al., 2024c) and llamaberry (Zhang et al., 2024b) utilizing CoT

Table 4: Approaches enhancing Cross-Modal Reasoning, which refers to the ability to integrate and reason across multiple modalities, such as text, images, videos.

Name	Modality	Cross-Modal Reasoning	Task	Highlight
IdealGPT (2023)	T, I	Answer sub-questions about image via gpt	VQA, Text Entailment	Using gpt to iteratively decompose and solve visual reasoning tasks
AssistGPT (2023)	T, I, V	Plan, Execute, Inspect via External Tools (gpt4, OCR, Grounding, et al.)	VQA, Causal Reasoning	Using an interleaved code and language reasoning approach to handle complex multimodal tasks
ProViQ (2024)	T, V	Generate and execute Python programs for the video	Video VQA	Using procedural programs to solve visual subtasks in videos
MM-REACT (2023)	T, I, V	Use CV tools for sub-taskss about image	VQA, Video VQA	Vision experts combined with GPT for multimodal reasoning and action
VisualReasoner (2024a)	T, I	Synthesize multi-step reasoning (Using exteral CV tools) data	GQA, VQA	Proposing a least-to-most visual reasoning paradigm and a data synthesis approach for training
Multi-model-thought (2025a)	T, I	External Tools (Visual Sketchpad)	Geometry, Math, VQA	Investigating inference-time scaling for multi-modal thought across diverse tasks
FaST (2024a)	T, I	System switch adapter for visual reasoning	VQA	Integrating fast and slow thinking mechanisms into visual agents
ICoT (2024a)	T, I	Generate interleaved visual-textual reasoning via ADS	VQA	Using visual patches as reasoning carriers to improve LMMs' fine-grained reasoning
Image-of-Thought (2024b)	T, I	Extract visual rationales step-by-step via IoT prompting	VQA	Using visual rationales to enhance LLMs' reasoning accuracy and interpretability
CoTDiffusion (2024a)	T, I	External Algorithms	Robotics	Generating subgoal images before action to enhance reasoning in long-horizon robot manipulation tasks
T-SciQ (2024e)	T, I	Model-Intrinsic Capabilities	ScienceQA	Using LLM-generated reasoning signals to teach multimodal reasoning for complex science QA
Visual-CoT (2023)	T, I	Model-Intrinsic Capabilities	VQA, DocQA, ChartQA	Using visual-text pairs as reasoning carriers to bridge logical gaps in sequential data
VoCoT (2024m)	T, I	Model-Intrinsic Capabilities	VQA	Using visually-grounded object-centric reasoning paths for multi-step reasoning
MVoT (2025b)	T, I	Model-Intrinsic Capabilities	Spatial Reasoning	Using multimodal reasoning with image visualizations to enhance complex spatial reasoning in LMMs

Table 5: Apporach of Multimodal-o1. It mainly relies on a multi-stage, structured reasoning path to solve problems.

Name	Backbone	Dataset	Modality	Reasoning Paradigm	Task Type	Highlight
Macro-O1 (2024c)	Qwen2-7B-Instruct	Open-O1 CoT + Marco-o1 CoT + Marco-o1 Instruction	T	MCTS-guided Thinking	Math, Translate	MCTS for solution expansion and reasoning action strategy
llamaberry (2024b)	LLaMA-3.1-8B	PRM800K + OpenMathInstruct-1	T	MCTS-guided Thinking	Math	SR-MCTS for search and PPRM for evaluation
LLaVA-CoT (2024a)	Llama-3.2V-11B-cot	LLaVA-CoT-100k	T, I	Summary, Caption, Thinking	Science, General	Introduce LLaVA-CoT-100k and scalable beam search
LlamaV-o1 (2025)	Llama-3.2V-11B-cot	LLaVA-CoT-100k + PixMo	T, I	Summary, Caption, Thinking	Science, General	Introduce VCR-Bench and outperforms
Mulberry (2024a)	Llama-3.2V-11B-cot, LLaVA-Next-8B, Qwen2-VL-7B	Mulberry-260K	T, I	Caption, Rationales, Thinking	Math, General	Introduce Mulberry-260k and CoMCTS for collective learning
RedStar-Geo (2025a)	InternVL2-8B	GeoQA	T, I	Long-Thinking	Math	Competitive with minimal Long-CoT data

fine-tuning began to emerge. CoT fine-tuning activates the model's inherent slow thinking ability through training methods. Compared to traditional CoT approaches, it enhances the model's reasoning capabilities on open-ended questions, introducing mechanisms for self-reflection and error correction. LLaVA-CoT (Xu et al., 2024b), LlamaV-o1 (Thawakar et al., 2025), RedStar (Xu et al., 2025a) and Mulberry (Yao et al., 2024a) extend the reasoning paradigm to the multimodal domain. In contrast to the two-stage reasoning paradigm of 'Thinking -> Answer' in text domains, these works expand the reasoning process to a four-stage approach includes Summary (Rationale), Caption, Thinking and Answer.

Building on CoT fine-tuning, testing-time scaling with various reasoning strategies is also an important method to enhance reasoning capabilities. Best-of-N sampling generates multiple responses for a given prompt, expanding the search space to identify better solutions. Beam Search, on the other hand, does not generate a complete response in one pass but instead selects the most promising intermediate outputs at each step using scoring. LLaVA-CoT (Xu et al., 2024b) and LlamaV-o1 (Thawakar et al., 2025) apply this method to strengthen reasoning abilities. Monte Carlo Tree Search (MCTS) allows for parallel exploration of multiple solution paths, ensuring a more refined search process compared to Beam Search. Marco-o1 (Zhao et al., 2024c), llamaberry (Zhang et al., 2024b) and Mulberry (Yao et al., 2024a) have successfully integrated this approach into the generation process of reasoning models.

Takeaways: Multimodal-O1

Multimodal-O1 models extend System-1 reasoning by deepening CoT workflows through multi-stage generation structures, long-horizon reasoning, and structured supervision. Enhanced by fine-tuning on rationale-rich data and supported by planning algorithms such as Beam Search or MCTS, these models achieve more coherent, interpretable, and scalable multimodal reasoning.

Table 6: Apporach of Multimodal-R1. It mainly employs reinforcement learning approaches to improve the reasoning capability of large multimodal models.

Apporach	Backbone	Dataset	RL Algorithm	Modality	Task Type	RL Framework	Cold Start	Rule-base/RM
RLHF-V (2024a)	LLaVA-13B	RLHF-V-Dataset (1.4k)	DPO	T, I	VQA	Muffin	-	(unknown)
InternVL2.5 (2024g)	InternVL	MMPR (3m)	MPO (DPO)	T, I	VQA	-	-	(unknown)
Insight-V (2024b)	LLaMA3-LLaVA-Next	-	DPO	T, I	VQA	trl	-	(unknown)
LLaVA-Reasoner-DPO (2024e)	LLaMA3-LLaVA-Next	ShareGPT4o-reasoning-dpo (6.6k)	DPO	T, I	VQA	trl	-	(unknown)
VLM-R1 (2025)	Qwen2.5-VL	coco , LISA , Refcoco	GRPO	T, I	Grounding ,Math , Open-Vocabulary Detection	trl	No	Rule-base
R1-V (2025b)	Qwen2-VL	CLEVR , GEOQA	GRPO	T, I	Counting , Math	trl	No	Rule-base
MM-EUREKA (2025)	InternVL2.5	K12 , MMPR	RLOO	T, I	Math	OpenRLHF	Yes	Rule-base
MM-EUREKA-Qwen (2025)	Qwen2.5-VL	K12 , MMPR	GRPO	T, I	Math	OpenRLHF	No	Rule-base
Video-R1 (2025b)	Qwen2.5-VL	Video-R1 (260K)	GRPO	T, I, V	Video VQA	trl	Yes	Rule-base
LMM-R1 (2025)	Qwen2.5-VL	VerMulti	PPO	T, I	Math	OpenRLHF	No	RM
Vision-R1 (2025b)	Qwen2.5-VL	LLaVA-CoT , Mulberry	GRPO	T, I	Math	-	Yes	Rule-base
Visual-RFT (2025f)	Qwen2-VL	coco , LISA , ...	GRPO	T, I	Detection , Classification	trl	No	Rule-base
R1-OneVision (2025e)	Qwen2.5-VL	R1-Onevision-Dataset	GRPO	T, I	Math , Science , General , Doc	-	Yes	Rule-base
Seg-Zero (2025)	Qwen2.5-VL , SAM2	RefCOCOg , ReasonSeg	GRPO	T, I	Grounding	verl	No	Rule-base
VisualThinker-R1-Zero (2025)	Qwen2-VL	SAT dataset	GRPO	T, I	Spatial Reasoning	trl	No	Rule-base
R1-Omni (2025c)	HumanOmni	MAFW , DFEW	GRPO	T, I, A, V	emotion recognition	trl	Yes	Rule-base
OThink-MR1 (2025e)	Qwen2.5-VL	CLEVR , GEOQA	GRPO	T, I	Counting , Math	-	No	Rule-base
Multimodal-Open-R1 (2025)	Qwen2-VL	multimodal-open-r1-8k-verified (based on Math360K and Geo170K)	GRPO	T,I	Math	trl	No	Rule-base
Curr-ReFT (2025)	Qwen2.5-VL	RefCOCOg , Math360K , Geo170K	GRPO	T,I	Detection , Classification , Math	Curr-RL	No	RM
Open-R1-Video (2025)	Qwen2-VL	open-r1-video-4k	GRPO	T, I, V	Video VQA	trl	No	Rule-base
VisRL (2025f)	Qwen2.5-VL	VisCoT	DPO	T,I	VQA	trl	Yes	RM
R1-VL (2025c)	Qwen2-VL	Mulberry-260k	StepGRPO	T,I	Math , ChartQA	not release	No	Rule-base

3.3.3 Multimodal-R1

The DPO in reinforcement learning has been widely used to enhance the reasoning capabilities of large multimodal models in recent years. RLHF-V (Yu et al., 2024a), LLaVA-Reasoner (Zhang et al., 2024e) and Insight-V (Dong et al., 2024b), by leveraging a large amount of self-constructed preference data and directly applying the DPO algorithm for training, have somewhat improved the reasoning ability of the models. MMPR (Wang et al., 2024g) made modifications to the DPO algorithm, adding quality loss obtained from a Binary Classifier and generation loss from traditional SFT on top of the DPO Preference loss, which effectively enhanced the model's CoT capabilities.

With the success of Deepseek-R1, the GRPO algorithm began to be widely applied in multimodal large models. Works including MM-EUREKA (Meng et al., 2025),Vt-R1 (Zhou et al., 2025), LMM-R1 (Yingzhe et al., 2025), R1-V (Chen et al., 2025b) , by adopting a similar approach to the text domain, have applied the GRPO algorithm to mathematical geometry problems, successfully demonstrating the phenomenon of reflection. VLM-R1 (Shen et al., 2025), Visual-RFT (Liu et al., 2025f), and Seg-Zero (Yuqi et al., 2025) utilize the GRPO algorithm to enhance the visual capabilities of multimodal large language models, such as grounding, detection, and classification. This reinforcement learning approach has successfully led to improvements in the model's visual capabilities. Besides, works includes Video-R1 (Feng et al., 2025b) and VideoChat-R1 (Li et al., 2025g) have introduced the GRPO algorithm into the video modality, while R1-Omni (Zhao et al., 2025c) has further extended it to the audio modality. Despite this, existing work is often limited to specific tasks, and current multimodal large models have not yet been able to generalize the long-chain-of-thought abilities learned from tasks such as mathematics to the model's general capabilities, as seen with Deepseek-R1.

Takeaways: Multimodal-R1

Multimodal-R1 methods leverage reinforcement learning—particularly DPO and GRPO, enhancing the model's ability to explore and optimize complex reasoning paths. These approaches improve reasoning depth, coherence, and domain adaptability by aligning model outputs with preference data or multi-modal feedback, laying the groundwork for more generalized long-horizon syatem-2 reasoning.

4 Towards Native Multimodal Reasoning Model

LMRMs have demonstrated potential in handling complex tasks with long chain of thoughts. However, their language-centric architectures constrain their effectiveness in real-world scenarios. Specifically, their reliance on vision and language modalities limits their capacity to process and reason over interleaved diverse data types, while their performance in real-time, iterative interactions with dynamic environments remains underdeveloped. These limitations underscore the need for a new class of models capable of broader multimodal integration and more advanced interactive reasoning.

In this section, we first analyze the performance of state-of-the-art LMRMs on benchmarks designed to assess omni-modal understanding and agentic capabilities, highlighting their limitations in real-world applicability (Sec. 4.1). Subsequently, we introduce the concept of **Native Large Multimodal Reasoning Models (N-LMRMs)**, which represent a paradigm shift in machine intelligence through two foundational capabilities: Multimodal Agentic Reasoning and Omni-Modal Understanding and Generative Reasoning (Sec.4.2). Finally, we will discuss the open challenges in building N-LMRMs and outline promising research directions to overcome these barriers (Sec. 4.3).

4.1 Experimental Findings

Although LMRMs have made significant progress in generating comprehensive thought processes and addressing complex questions such as MMMU (Yue et al., 2024) and MathVista (Lu et al., 2024), autonomously solving these questions is far from real-world utility in the following aspects: 1) Evaluation scopes should cover multiple modalities, including vision, audio, and text. 2) Evaluation capabilities should involve interaction with external environments, requiring long-horizon reasoning and adaptive planning. Here we present a summary of our collected omni-modal and agentic benchmarks in Table 7, followed by an analysis of LMRMs' performance on these benchmarks.

Omni-modal Benchmarks Recent studies have introduced a series of omni-modal benchmarks designed to evaluate the ability of LMRMs to perform unified understanding and reasoning across various data types (e.g. images, audio, text, and video). For example, OmniMMI (Wang et al., 2025g) aims to comprehensively assess the interactive capabilities of streaming video contexts in open-world environments. Experimental results reveal that even commercial models, such as Gemini-1.5-Pro and GPT-4o, achieve an average accuracy of less than 20%. When tasks require unified modality understanding (OmniBench (Li et al., 2024j), TaskAnything and JudgeAnything (Pu et al., 2025), MixEvalL-X (Ni et al., 2024b)), the performance of both open-source and closed-source models is significantly lower than under single-modal conditions. Specifically, in the Audio-Video Question Answering (AVQA) task, such as WorldSense (Hong et al., 2025), Claude 3.5 Sonnet only achieves an average accuracy of 35%, while the best-performing open-source model only achieves an accuracy of 25%. In the case of more challenging multimodal reasoning tasks, such as BabelBench (Wang et al., 2024i) and OmnixR (Chen et al., 2024e), the performance of all models declines sharply as the number

Table 7: A summary of agentic and omni-modal benchmarks, which expose the deep reasoning flaws of current LMRMs. T, I, A, V represent text, image, audio and video respectively.

Dataset	Task	Modality	Characteristic
	Agentic Benchmark		
AgentBench (Liu et al., 2023b)	Code, Web Navigation, General Reasoning	T	Eight Different Environments
WorfBench (Qiao et al., 2024)	Workflow Evaluation	T	Multi-Faceted Scenarios and Intricate Graph Workflows
OSWorld (Xie et al., 2024a)	Computer Using, GUI Navigation	T, I, V	Real Computer Environment Infrastructure
EmbodiedBench (Yang et al., 2025b)	Multimodal Understanding, Spatial Reasoning	T, I	High and Low Action Levels
EmbodiedEval (Cheng et al., 2025)	Attribute QA, Spatial Reasoning	T, I	Broad Abilities Assessment
SPA-Bench (Chen et al., 2024c)	Single and Cross APP Using	T, I	Tasks Across English and Chinese APPs
VisualWebBench (Liu et al., 2024b)	VQA, OCR, Grounding, General Reasoning	T, I	1.5K Human-Curated Instances
VisualWebArena (Koh et al., 2024)	Web Navigation, Visual Understanding	T, I	Realistic Visually Grounded Web Tasks
VisualAgentBench (Liu et al., 2024d)	Household, GUI Navigation, CSS Debugging	T, I	Tasks Across Embodied, GUI and Visual Design
GAIA (Mialon et al., 2023)	Multimodality Handling, Web Browsing, Generally Tool-Use and Reasoning	T, I	Increasing Difficulty Level
BrowseComp (Wei et al., 2025a)	Web Browsing	T	Easy to Verify but Hard to Solve
SWE-Bench Multimodal (Yang et al., 2024b)	Code	T, I	Image Included in Problem Statement
AndroidWorld (Rawles et al., 2024)	APP Using	T, I	Fully Functional Android Environment
GTA (Wang et al., 2024b)	Tool Using	T, I	Tool Using in Real-World Scenarios
WorkArena++ (Boisvert et al., 2024)	Web Search, GUI Navigation, General Reasoning	T, I	Realistic Office Worker Trajectories
WindowsAgentArena (Bonatti et al., 2024)	Windows OS Using	T, I	Realistic Windows OS Environment
	Omni-Modal Benchmark		
OmniMMI (Wang et al., 2025g)	VQA, Proactive Reasoning	T, V, A	In Streaming Video Context
OmniBench (Li et al., 2024j)	Omni-Understanding	T, I, A, V	Simultaneous Multimodal Reasoning
JudgeAnything (Pu et al., 2025)	Multimodal Understanding, Generation and Evaluation	T, I, A, V	MLLM as A Judge Across Any Modality
WorldSense (Hong et al., 2025)	AVQA	T, A, V	Collaboration of Omni-Modality
BabelBench (Wang et al., 2024i)	VQA, Math, Spatial Reasoning, General Reasoning	T, I	Code-Driven Multimodal Data Analysis
OmnixR (Chen et al., 2024e)	Omni-Modal Reasoning	T, I, A, V	Synthetic Dataset and Real-world Dataset
LongVALE (Geng et al., 2024)	AVQA	T, A, V	105K Omni-Modal Events with Temporal Boundaries
MixEvalL-X (Ni et al., 2024b)	Multimodal Understanding and Generation	T, I, A, V	Standardizing Cross-Modal Evaluations

of modalities increases. This suggests that models struggle to generate reasoning paths for image, video, and audio inputs compared to text inputs. These findings collectively highlight that current LMRMs are not yet capable of effectively processing omni-modal inputs.

Agent Benchmarks A diverse range of tasks highlights the complexity and breadth of multimodal agent evaluation settings. These include AgentBench's multi-environment tasks (Liu et al., 2023b, 2024d), WorFBench's intricate workflow planning scenarios (Qiao et al., 2024), OSWorld's and AndroidWorld's full operating system interactions (Xie et al., 2024a; Rawles et al., 2024), EmbodiedBench's vision-based navigation and manipulation challenges (Yang et al., 2025b), VisualWebArena's visually grounded web tasks (Koh et al., 2024), and GAIA's open-ended, tool-augmented queries (Hu et al., 2023). Together, these benchmarks span a wide spectrum of task types and modalities (e.g., text and vision), encompassing both realistic and tool-augmented environments.

Regarding the performance of LMRMs on agent benchmarks, these models generally lead current performance and have made notable progress (Team, 2024, 2025a; Yao et al., 2024b). However, even state-of-the-art models consistently fall short of human-level reliability and struggle with complex, open-ended tasks. Evaluations across benchmarks repeatedly expose common bottlenecks: models often fail at real-world grounding (Gou et al., 2025; Zheng et al., 2024a), coherent long-horizon reasoning and planning (Qian et al., 2025), seamless integration with external tools (Wang et al., 2025d), and maintaining robustness across diverse modalities and domains (Chu et al., 2025). For example, in the BrowseComp benchmark (Wei et al., 2025a), GPT-4o achieves only 0.6% accuracy, rising to 1.9% with browsing tools, highlighting weak tool-interactive planning capability. OpenAI's reasoning model o1 reaches 9.9%, but still leaves significant room for improvement. Notably, OpenAI Deep Research, with targeted tuning for web search, completes 51.5% of tasks via autonomous iterative tool calling and reasoning. The experimental results highlight that current large reasoning models remain deficient in long-horizon reasoning and adaptive planning, which may require specific tuning and architectural enhancements to evolve into truly native agentic systems.

Preliminary Study with o3 and o4-mini Recently, OpenAI released o3 and o4-mini, providing full agentic access to ChatGPT tools and enabling models to "think with images." The integration of visual content

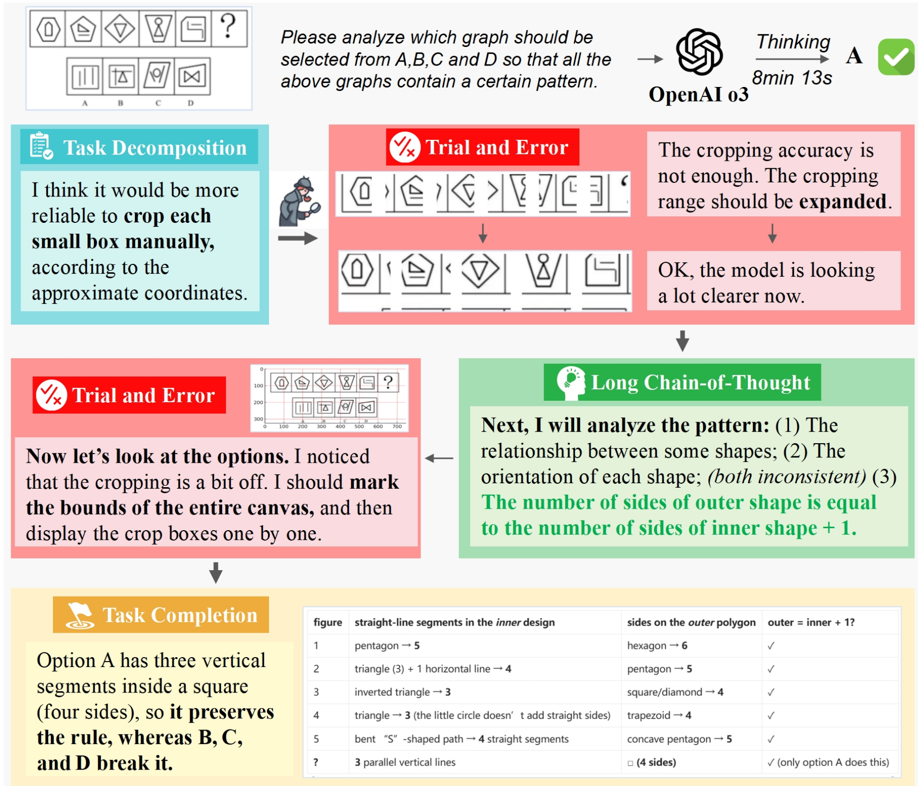

Figure 6: Case study of OpenAI o3's long multimodal chain-of-thought, reaching the correct answer after 8 minutes and 13 seconds of reasoning. The question is from Chinese Civil Service Examination.

enhances multimodal reasoning directly within the thought process. For example, in Figure 6, o3 demonstrates a clear task decomposition during an 8 minute and 13 second thought process. It effectively determines the best way to crop each sub-figure through trial and error, ultimately arriving at the correct solution.

Beyond visual reasoning, we evaluated o3's capabilities in file processing, puzzle solving, location identification, and multimedia content creation. As illustrated in Figure 7 and 8, o3 exhibits strong performance in complex multimodal problem-solving by capturing and leveraging subtle clues in images. However, several challenges are identified: 1) **Language knowledge can interfere with visual input.** As the finger counting case shown n Figure 8, o3 mistakenly identifies the image as the standard raised hand emoji showing four fingers plus a thumb, despite the image clearly displaying six fingers. 2) **OpenAI o3 struggles with input file handling and multimedia content generation.** Due to tool constraints and the lack of Internet access in coding environments, file processing and multimedia creation often result in inaccuracies. In the resume information collection case in Figure 8, phone numbers parsed from resume PDFs can be incorrect, and o3 hallucinates candidates' project experiences by reusing similar content. Additionally, in multimedia creation cases in Figure 7, the generated frames fail to adhere to the "red panda" instructions, and o3 is unable to support interleaved text-image generation. 3) **OpenAI o3 may fabricate reasoning in its thought process.** It occasionally "lies" about its reasoning, constructing incorrect rationales for potentially correct answers (e.g., the puzzle-solving case in Figure 7). This problem needs to be solved urgently, as it could lead to the model attempting to deceive users during the post-training process. In fact, it highlights that the model has not yet mastered the relevant thinking logic to solve the problem.

4.2 Capability of N-LMRMs

Based on the above experimental findings, we introduce the concept of **Native Large Multimodal Reasoning Models (N-LMRMs)**. N-LMRMs are inherently designed to integrate multimodal understanding, generation, and agentic reasoning across any modality, which will be beyond the perception and reasoning scope of o4-mini. This advancement will build upon two transformative capabilities that have been explored largely in parallel: *Multimodal Agentic Reasoning*, which enables proactive, goal-driven interactions through

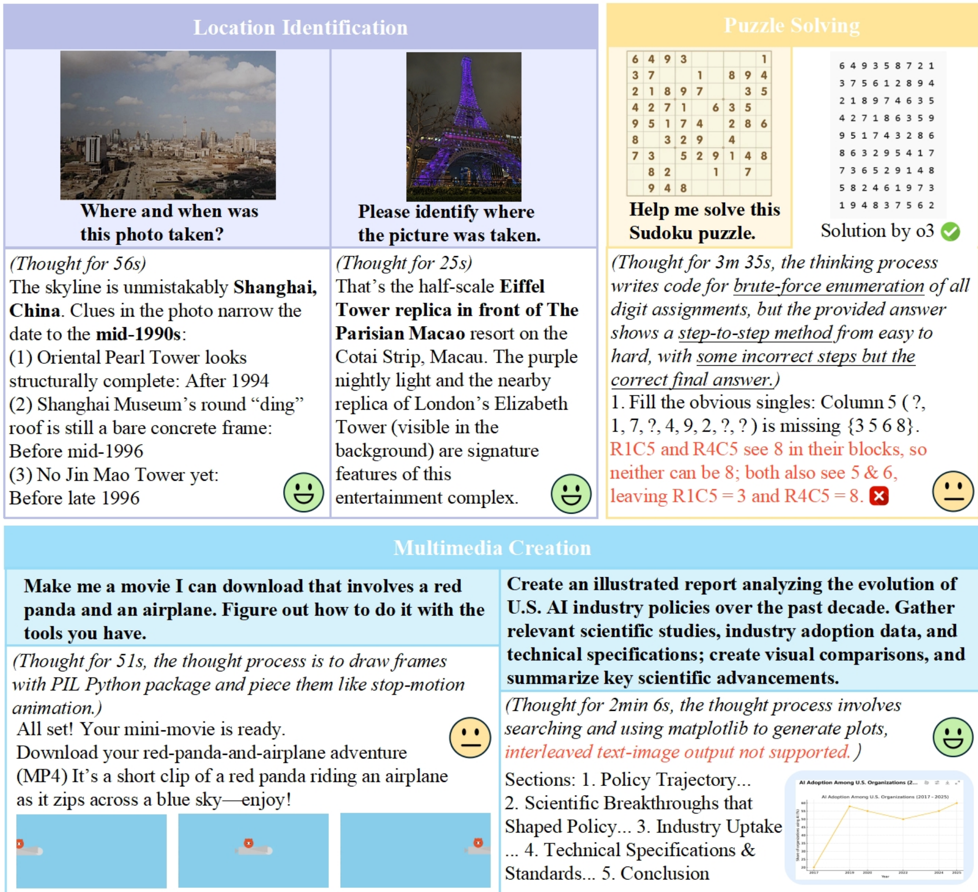

Figure 7: Case study of OpenAI o3: Find locations, solve a puzzle and create multimedia contents.

hierarchical task decomposition, real-time strategic adaptation, and embodied learning; and *Omni-Modal Understanding and Generative Reasoning*, which supports seamless cross-modal synthesis and analysis via unified representations—facilitating heterogeneous data fusion and contextual multimodal interaction. Table 8 summarizes key existing works related to agentic and omni-modal models. These models only explore some of the capabilities of N-LMRMs and do not combine the above two capabilities to build a more powerful large multimodal reasoning model.

Multimodal Agentic Reasoning A core capability of multimodal agentic reasoning is dynamic adaptation, which can adjust strategies in real time based on environmental feedback. Some of the latest products from the industry have initially demonstrated this capability. As Model Context Protocol (MCP) (Anthropic, 2025) and Agent2Agent Protocal (A2A) (Surapaneni et al., 2025) facilitates seamless integration of diverse tools and enables dynamic interaction across various external environments, these protocols underscore the importance of multimodal agentic reasoning, enabling agents to adapt strategies in real-time based on environmental feedback, thereby enhancing their effectiveness in dynamic and multifaceted real-world applications. For instance, **Operater** combines the visual capabilities of GPT-4o with advanced reasoning capabilities achieved through reinforcement learning, which enables it to interact with the operating system and browser in real-time through a graphical user interface (GUI), continuously improving its browsing and data operations during task execution. Similarly, **Claude Computer Use** allows models to manipulate and navigate desktop environments, learning optimal interaction strategies through trial and error.

Moreover, Search-o1 (Li et al., 2025e) utilizes external knowledge retrieval during the reasoning process to fill gaps in their understanding. R1-Searcher (Song et al., 2025b) and DeepResearcher (Zheng et al., 2025e) enhance their ability to autonomously use search engines to collect information through reinforcement learning. By incorporating this autonomous knowledge retrieval into the reasoning process, these systems are able to act with a more refined understanding and adapt their responses to changing tasks. **Gemini 2.0** has the

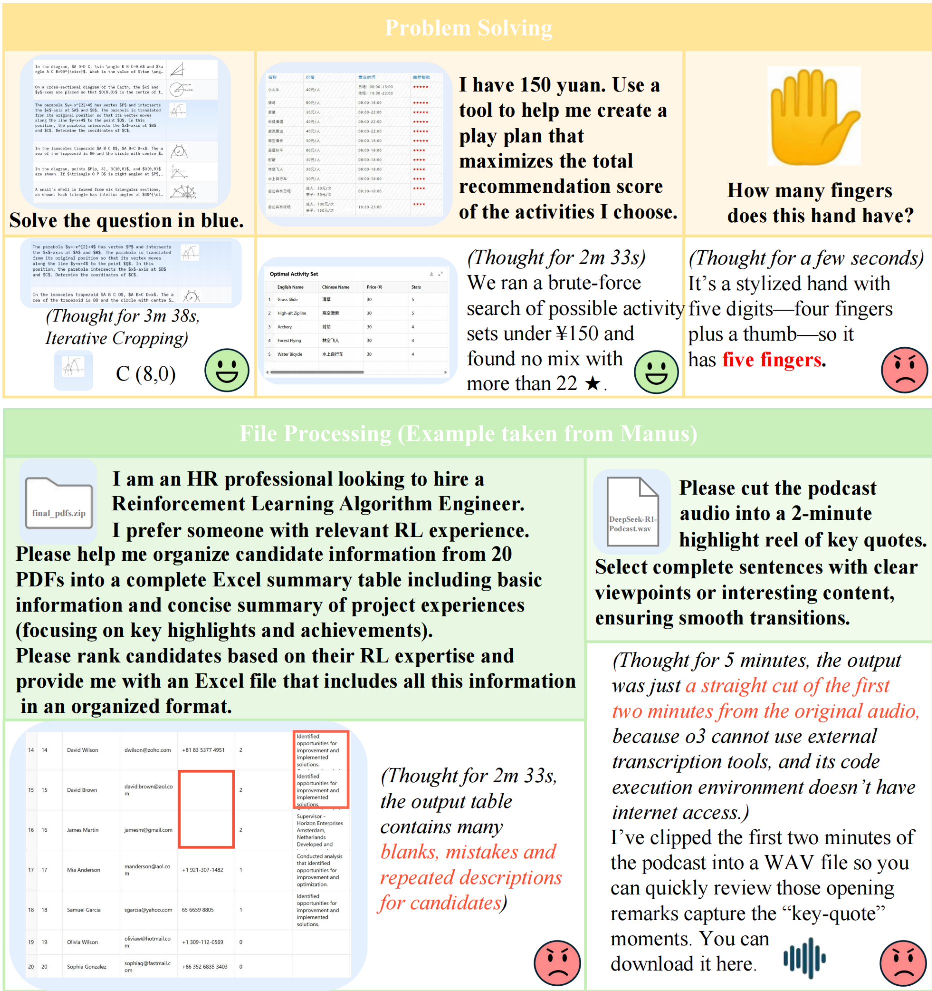

Figure 8: Case study of OpenAI o3: Visual problem solving and file processing.

ability to process and generate multi-modal content. By deeply integrating with Google's various tools and combining its advanced reasoning capabilities, it can effectively decompose tasks and gradually obtain the required information when dealing with multi-step problems. *While current models have demonstrated initial versions of this functionality, they fall short in their ability to engage in sustained, interactive reasoning across diverse modalities*.

Another aspect is the embodied learning of LMRMs to handle the external environment. Embodied learning is exemplified by systems capable of interacting with both digital and physical environments. For example, Magma (Yang et al., 2025a) learns by interacting with real-world data, improving its spatial-temporal reasoning to navigate and manipulate objects effectively in both virtual and physical contexts. Similarly, OpenVLA (Kim et al., 2024) combines a visual encoder with a language model, enabling the system to learn from real-world robot demonstrations. This embodied approach allows the model to acquire both visual and task-specific reasoning skills, enhancing its ability to perform complex, real-world actions that require multimodal understanding and adaptation. In summary, recent RL-scale methods will greatly stimulate the agentic behavior of large-scale models, pushing to the world model.

Omni-Modal Understanding and Generative Reasoning The behaviors of multimodal agents are closely linked to the deep reasoning capabilities of the underlying large multimodal models, particularly in terms of perception range, understanding accuracy, and reasoning depth. Thus, developing a comprehensive omni-modal model for real-world applications and enhancing its deep reasoning ability is foundational.

Table 8: A summary of recent agentic and omni-modal models towards N-LMRMs.

Model	Parameter	Input Modality	Output Modality	Training Strategy	Task	Characteristic
Agentic Models						
R1-Searcher (Song et al., 2025b)	7B, 8B	T	T	RL	Multi-Hop QA	RL-Enhanced LLM Search
Search-o1 (Li et al., 2025e)	32B	T	T	Training-Free	Multi-Hop QA, Math	Agentic Search-Augmented Reasoning
DeepResearcher (Zheng et al., 2025e)	7B	T	T	RL	Multi-Hop QA	RL in Live Search Engines
Magma (Yang et al., 2025a)	8B	T, I, V	T	Pretrain	Multimodal Understanding, Spatial Reasoning	820K Spatial-Verbal Labeled Data
OpenVLA (Kim et al., 2024)	7B	T, I	T	SFT	Spatial Reasoning	970k Real-World Robot Demonstrations
CogAgent (Hong et al., 2024)	18B	T, I	T	Pretrain+SFT	VQA, GUI navigation	Low-High Resolution Encoder Synergy
UI-TARS (Qin et al., 2025)	2B, 7B, 72B	T, I	T	Pretrain+SFT+RL	VQA, GUI navigation	End-to-End GUI Reasoning and Action
Seeclick (Cheng et al., 2024b)	10B	T, I	T	Pretrain+SFT	GUI navigation	Screenshot-Based Task Automation
Omni-Modal Model						
Gemini 2.0 & 2.5	/	T, I, A, V	T, I, A	/	/	/
GPT-4o	/	T, I, A, V	T, I, A	/	/	/
Megrez-3B-Omni (Li et al., 2025a)	3B	T, I, A	T	Pretrain+SFT	VQA, OCR, ASR, Math, Code	Multimodal Encoder-Connector-LLM
Qwen2.5-Omni (Xu et al., 2025b)	7B	T, I, A, V	T, A	Pretrain+SFT	VQA, OCR, ASR, Math, Code	Time-Aligned Multimodal RoPE
Baichuan-Omni-1.5 (Li et al., 2025h)	7B	T, I, A, V	T, A	Pretrain+SFT	VQA, OCR, ASR, Math, GeneralQA	Leading Medical Image Understanding
M2-omni (Guo et al., 2025)	9B, 72B	T, I, A, V	T, I, A	Pretrain+SFT	VQA, OCR, ASR, Math, GeneralQA	Step Balance For Pretraining and Adaptive Balance For SFT
MiniCPM-o 2.6 (Team, 2025b)	8B	T, I, A, V	T, A	Pretrain+SFT+RL	VQA, OCR, ASR, AST	Parallel Multimodal Streaming Processing
Mini-Omni2 (Xie & Wu, 2024)	0.5B	T, I, A	A	Pretrain+SFT	VQA, ASR, AQA, GeneralQA	Real-Time and End-to-End Voice Response
R1-Omni (Zhao et al., 2025c)	0.5B	T, A, V	T	RL	Emotion Recognition	RL with Verifiable Reward
Janus-Pro (Chen et al., 2025d)	1B, 7B	T, I	T, I	Pretrain+SFT	Multimodal Understanding, Text-to-Image	Decoupling Visual Encoding For Understanding and Generation
AnyGPT (Zhan et al., 2024)	7B	T, I, A	T, I, A	Pretrain	Multimodal-to-Text and Text-to-Multimodal	Discrete Representations For Unified Processing
Uni-MoE (Li et al., 2025j)	13B, 20B, 22B, 37B	T, I, A, V	T	Pretrain+SFT	VQA, AQA	Modality-Specific Encoders with Connectors for Unified Representation

Early work, AnyGPT (Zhan et al., 2024), utilizes discrete representations for the unified processing of various modalities, achieving unified understanding and generation across modalities. Recently, Baichuan-Omni-1.5 (Li et al., 2025h) showcases impressive capabilities in collaborative real-time understanding across various modalities. Qwen2.5-Omni (Xu et al., 2025b) uses a new position embedding, named Time-aligned Multimodal RoPE, to synchronize the timestamps of video inputs with audio. More latest open source work, like M2-omni (Guo et al., 2025) and MiniCPM-o (Yu et al., 2024b), is narrowing the performance gap with closed-source models like GPT-4o.

Driven by real-world specific needs, omni-modal models with smaller size are gaining more and more attention. Megrez-3B-Omni (Li et al., 2025a) is an on-device multimodal understanding LLM model that demonstrates excellent performance in tasks such as scene understanding and OCR. Mini-Omni2 (Xie & Wu, 2024), a visual-audio assistant capable of providing real-time, end-to-end voice responses to visoin and audio queries. R1-Omni (Zhao et al., 2025c) focuses on emotion recognition from visual and auditory information.

Despite these advancements, current research in multimodal AI primarily focuses on enhancing the comprehension and generation of unified multimodal representations. The development of reasoning capabilities that effectively integrate and interrogate cross-modal interactions remains critically underexplored. Bridging this gap is essential for realizing native multimodal reasoning models—systems inherently designed to process, analyze, and synthesize interconnected modalities with human-like sophistication.

4.3 Technical Prospects

The technical prospect of Native Large Multimodal Reasoning Models (N-LMRMs) aims to natively unify understanding, generation, and reasoning across diverse data types, from language and vision to audio, tactile, sensor readings, temporal sequences, and structured data, bringing us closer to systems that can see, hear, talk, and act in a unified and cohesive manner. However, building such N-LMRMs poses significant challenges. These models must be architecturally designed to handle heterogeneous modalities within a single system, genetically use and combine diverse tools through long multimodal reasoning chains, and support continuous learning from real-world interactions. This section outlines key challenges in building N-LMRMs and proposes several potential pathways to address them.

Unified Representations and Cross-Modal Fusion. A fundamental challenge is creating a single model architecture that can process and generate different modalities in a coherent way. Traditional approaches often use separate encoders for each modality (Lyu et al., 2023; Li et al., 2024l). In contrast, native omni-modal

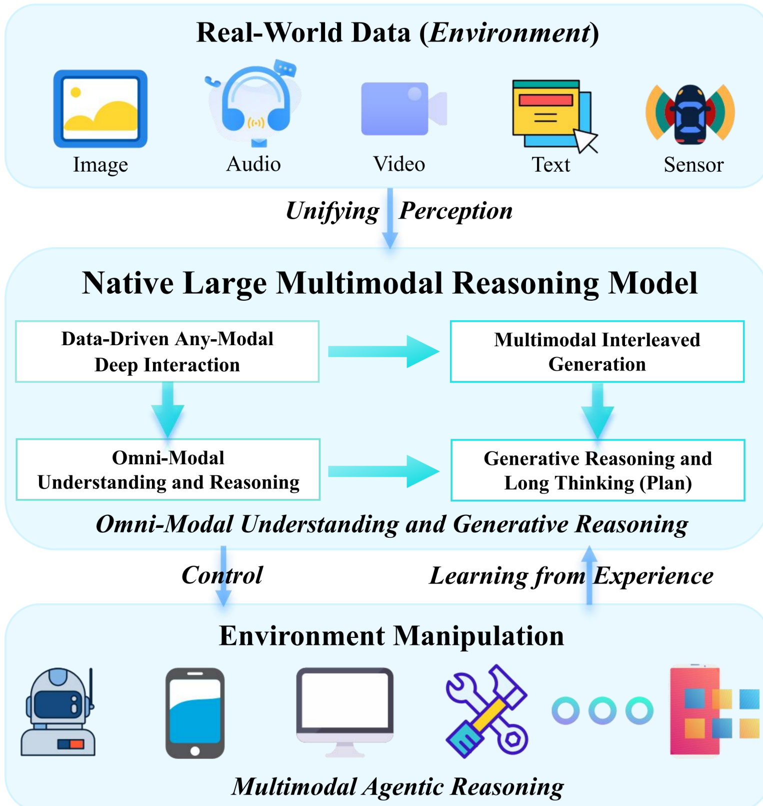

Figure 9: Overview of next-generation native large multimodal reasoning model. The envisioned system aims to achieve comprehensive perception across diverse real-world data modalities, enabling precise omnimodal understanding and in-depth generative reasoning. This foundational model will lead to more advanced forms of intelligent behavior, learning from world experience and realizing lifelong learning and self-improvement.

models seek a more unified design that allows for seamless interaction between modalities. One possible solution is to homogenize all inputs and outputs into a common format and process any modality uniformly. This approach requires careful design to prevent negative interference, where one modality may dominate or impair the representation of others (Leng et al., 2024; Chen et al., 2024g). Thus an emerging solution is Mixture-of-Experts (MoE) architectures, with experts specialized for certain modalities are only activated for relevant inputs, while a core language model serves as the backbone for language intelligence (Chen et al., 2024i; Li et al., 2025j; Team, 2025a; Shukor et al., 2025).

Interleaved Multimodal Long Chain-of-Thought. Building on unified representations, N-LMRMs can extend traditional long internal chains of thought into interleaved reasoning processes across multiple modalities. This enables a new axis for test-time compute scaling that seamlessly blends different modalities (Wang et al., 2025a). OpenAI's recently released o3 and o4-mini represent pioneering steps in this direction, i.e. reasoning with images in their chain of thought (OpenAI, 2025b), by automatically processing with tools that can zoom, crop, flip, or enhance images. Importantly, these capabilities come natively, without relying on separate specialized models (Wu & Xie, 2023; Hu et al., 2024b; Feng et al., 2025a; Qian et al., 2025; Wang et al., 2025d). Driven by the promising generalization capabilities of reinforcement learning across domains such as software engineering (OpenAI, 2025), IMO-level math (DeepSeek-AI et al., 2025), creative writing (Zhao et al., 2024c), and GUI manipulation (Qin et al., 2025), scaling reinforcement learning to more modalities, longer tool-augmented reasoning chains, and a broader set of reasoning tasks could be the recipe for the next generation of N-LMRMs, capable of simulating cross-modal reasoning and elevating machine intelligence.

Learning and Evolving from World Experiences. In dynamically evolving intelligent systems, the core value of LMRMs-based "World Model[2]" lies not only in its real-time modelling and reasoning capabilities in complex environments, like autonomous driving (Wang et al., 2024m) but also in its evolutionary mechanism for life-long learning (Thrun & Mitchell, 1995) through continuous interaction with the environment. When the MCP and A2A create a high-density network of tools and agent clusters, the system can transform each interaction into structured experiences through multidimensional engagement with the environment, tools, and other agents. This includes everything from pattern recognition in real-time data streams to causal reasoning across tool operation chains, from collaborative feedback in communication networks to autonomous adaptation in abnormal scenarios.

This continuous learning paradigm enables LMRMs to overcome the limitations of static knowledge bases. By iteratively accumulating world experiences, it dynamically updates its cognitive architecture and decision-making strategies. Particularly in open environments, the autonomous learning mechanism drives the model to actively explore the potential of tool combinations. In the process of solving new problems, it simultaneously stores transferable knowledge, ultimately forming an intelligent system that possesses specialized reasoning capabilities while maintaining cross-scenario generalization resilience. We think the interactive learning method of online reinforcement learning and offline verification methods may iteratively and continuously stimulate the capabilities of LMRMs, which have been utilized in the GUI agentic model (Qin et al., 2025; Zheng et al., 2025a; Wang et al., 2024n) to continually improve the performance.

Data Synthesis. The current capabilities of LMRMs are largely data-driven. To enhance these models during the pre-training stage, it is crucial to develop a high-quality data synthesis pipeline that tailors their functionalities. Most existing efforts (Chang et al., 2024; Huang et al., 2025c; Xu et al., 2024c) in data synthesis focus on improving single-modal or cross-modal understanding and reasoning, particularly in domains like vision, language, and speech. However, there has been limited exploration of more complex aspects, such as aligning three or more modalities, creating multimodal interactive chains of thought and visual generation, implementing multi-step planning in dynamic environments, and coordinating multi-tool calls and parallel tool usage. These areas present significant opportunities for advancing multimodal reasoning models.

In conclusion, we introduce the concept of N-LMRM as an initial step towards transitioning from capable reasoners to autonomous agents. Additionally, in alignment with OpenAI's five-stage pathway to AGI (OpenAI, 2023), we are laying the groundwork for subsequent stages, including self-evolving innovators (Yamada et al., 2025) and multi-agent organizations (Zhang et al., 2025d). Building on our research proposal, future work can explore more agentic and omni-modal capabilities, advancing the development of increasingly autonomous machine intelligence.

Takeaways: Native Large Multimodal Reasoning Model (LMRMs)

In this section, we examined the latest large multimodal model (e.g., O3 and O4-mini) and their performance on challenging tasks and benchmarks. We then presented the future trajectory for native multimodal large models in terms of capability scope and level, including omnimodal perception and understanding, multimodal interactive generative reasoning, and intelligent agent behavior. To realize this vision, we discussed approaches related to unified perception, learning methods, and data synthesis. We hope that native LMRMs will achieve comprehensive perception, precise understanding, and deep reasoning as a paradigm shift in machine intelligence.

5 Dataset and Benchmark

In exploring the development and optimization of Multimodal Reasoning Models, a surge of tasks and benchmarks have been proposed to conduct empirical ability evaluation and analysis for evaluating model performance across various aspects, e.g., video understanding and visual reasoning. In this section, we summarize and categorize existing datasets that are useful to facilitate the development of Multimodal Reasoning Models into four major types based on capacity: **(1)** Understanding; **(2)** Generation; **(3)** Reasoning; and **(4)** Planing. Then, we summarize commonly used metrics and evaluation aspects for these benchmarks or datasets. Benchmarks are designed with specific ability evaluation, and we classify four primary categories as shown in Figure 10 and eleven subcategories, as shown in Table 9.

[2] https://sites.google.com/view/worldmodel-iclr2025/

Table 9: Overview of Multimodal Benchmarks and Datasets (Training), categorized by task: Understanding (Visual-centric, Audio-centric), Generation (Cross-modal, Joint Multimodal), Reasoning (General Visual, Domain-Specific), and Planning (GUI, Embodied & Simulated Environments). These benchmarks often require short or long reasoning for successful task completion, e.g., challenging visual and audio generation.

Ability	Task	Benchmark	Dataset
Multimodal Understanding	Visual Centric	VQA (Kafle & Kanan, 2016), GQA (Hudson & Manning, 2019) DocVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019) OCR-VQA (Mishra et al., 2019), CMMLU (Li et al., 2024f) C-Eval (Huang et al., 2023c), MTVQA (Tang et al., 2024) Perception-Test (Patraucean et al., 2023), Video-MMMU (Hu et al., 2025b) Video-MME (Fu et al., 2024a), MMBench (Liu et al., 2024f) Seed-Bench (Li et al., 2023c), MME-RealWorld (Zhang et al., 2024f) MMMU (Yue et al., 2024), MM-Vet (Yu et al., 2024c) MMT-Bench (Ying et al., 2024), Hallu-PI (Ding et al., 2024) ColorBench (Liang et al., 2025b), DVQA (Kafle et al., 2018) MMStar (Chen et al., 2024f), TRIG-Bench (Li et al., 2025d) MM-IFEval (Ding et al., 2025), All-Angles Bench (Yeh et al., 2025) M3Exam (Zhang et al., 2023e), Exams-V (Das et al., 2024) TikTalkCoref (Li et al., 2025f), AgMMU (Gauba et al., 2025) Kaleidoscope (Salazar et al., 2025), VideoComp (Kim et al., 2025) CliME (Borah et al., 2025), TDBench (Hou et al., 2025) RefCOCOm (Liu et al., 2025a) SBVQA (Alasmary & Al-Ahmadi, 2023), H2VU-Benchmark (Wu et al., 2025) 4D-Bench (Zhu et al., 2025), V2P-Bench (Zhao et al., 2025e) RSMMVP (Adejumo et al., 2025), HIS-Bench (Zhao et al., 2025b) MMLA (Zhang et al., 2025b), SARLANG-1M (Wei et al., 2025b)	ALIGN (Jia et al., 2021), LTIP (Wu et al., 2024b) YFCC100M (Thomee et al., 2016), DocVQA (Mathew et al., 2021) Visual Genome (Krishna et al., 2016), Wukong (Gu et al., 2022) CC3M (Sharma et al., 2018), ActivityNet-QA (Yu et al., 2019a) SBU-Caption (Ordonez et al., 2011), AI2D (Hiippala et al., 2021) LAION-5B (Schuhmann et al., 2022), LAION-400M (Schuhmann et al., 2021) MS-COCO (Lin et al., 2014b), Virpt (Yang et al., 2024a) OpenVid-1M (Nan et al., 2024), VidGen-1M (Tan et al., 2024b) Flickr30k (Plummer et al., 2017), COYO-700M (Lu et al., 2023) WebVid (Bain et al., 2022), Youku-mPLUG (Xu et al., 2023a) VideoCC3M (Nagrani et al., 2022), FILIP (Yao et al., 2021) CLIP (Radford et al., 2021), YouTube8M (Abu-El-Haija et al., 2016) OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022) TikTalkCoref (Li et al., 2025f), MRES-32M (Liu et al., 2025a) EarthScape (Massey & Imran, 2025)
	Audio Centric	AudioBench (Wang et al., 2024a), VoiceBench (Chen et al., 2024h) Fleurs (Conneau et al., 2022), MusicBench (Melechovský et al., 2024) Air-Bench (Yang et al., 2024d), MMAU (Sakshi et al., 2024) SD-eval (Ao et al., 2024), CoVoST2 (Wang et al., 2020) MusicNet (Thickstun et al., 2017), AVE-PM (Liu et al., 2025b) ACVUBench (Yang et al., 2025f)	Librispeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020) Aishell (Bu et al., 2017), Fleurs (Conneau et al., 2022), MELD (Poria et al., 2019) CoVoST2 (Wang et al., 2020), SIFT-50M (Pandey et al., 2025) Clotho (Drossos et al., 2020), AudioCaps (Kim et al., 2019) ClothoAQA (Lipping et al., 2022), MusicNet (Thickstun et al., 2017) NSynth (Engel et al., 2017), MusicCaps (Agostinelli et al., 2023)
Multimodal Generation	Cross-modal Generation	GenEval (Ghosh et al., 2023), T2I-CompBench++ (Huang et al., 2025a) DPG-Bench (Hu et al., 2024a), GenAI-Bench (Li et al., 2024a) VBench (Huang et al., 2024a), VideoScore (He et al., 2024) WorldSimBench (Qin et al., 2024), WorldModelBench (Li et al., 2025c) MagicBrush (Zhang et al., 2023d), VBench++ (Huang et al., 2024b) MJHQ-30K (Li et al., 2024d), VBench 2.0 (Zheng et al., 2025b) AIGCBench (Fan et al., 2019), EvalCrafter (Liu et al., 2024e) LMM4LMM (Wang et al., 2025e), RISEBench (Zhao et al., 2025d) WikiVideo (Martin et al., 2025), OmniDiff (Liu et al., 2025c) ExtremeAIGC (Chandna et al., 2025), MCiteBench (Hu et al., 2025a)	MS-COCO (Lin et al., 2014b), Flickr30k (Plummer et al., 2017) Conceptual Captions (Sharma et al., 2018), RedCaps (Desai et al., 2021) CommonPool (Gadre et al., 2023), LLaVA-Pretrain (Liu et al., 2023a) Aishell1 (Bu et al., 2017), ThreeDWorld (Gan et al., 2021) X2I (Xiao et al., 2024), GAIA-1 (Hu et al., 2023) UniSim (Yang et al., 2024e), VidProM (Wang & Yang, 2024) LWM (Liu et al., 2024a), Genesis (Authors, 2024) HQ-Edit (Hui et al., 2024), InstructPix2Pix (Brooks et al., 2023) MagicBrush (Zhang et al., 2023d)
	Joint Multimodal Generation	MM-Interleaved (Tian et al., 2024), ANOLE (Chern et al., 2024) InterleavedEval (Liu et al., 2024c), OpenLEAF (An et al., 2024) OpenING (Zhou et al., 2024a), M2RAG (Ma et al., 2025c) SEED-Bench (Li et al., 2023c), SEED-Bench-2 (Li et al., 2023b) MME-Unify (Xie et al., 2025b)	DreamLLM (Dong et al., 2023), SEED-Story (Yang et al., 2024f) NextGPT (Wu et al., 2024a), DreamFactory (Xie et al., 2024b) DreamRunner (Wang et al., 2024o), EVA (Chi et al., 2024)
Multimodal Reasoning	General Visual Reasoning	NaturalBench (Li et al., 2024b), VCR (Zellers et al., 2019) PhysBench (Chow et al., 2025), MMBench (Liu et al., 2024f) MMMU (Yue et al., 2024), AGIEval (Zhong et al., 2024b) MMStar (Chen et al., 2024f), InfographicVQA (Mathew et al., 2022) VCRBench (Qi et al., 2025), VisualPuzzles (Song et al., 2025d) IV-Bench (Ma et al., 2025a), VisuLogic (Xu et al., 2025d) FG-BMK (Yu et al., 2025), Video-MMLU (Song et al., 2025a) DVBench (Zeng et al., 2025), GeoSense (Xu et al., 2025c) FLIP Plesner et al. (2025), ViLBench (Tu et al., 2025b) HAVEN (Gao et al., 2025a), MAGIC-VQA (Yang et al., 2025c) PM4Bench (Gao et al., 2025b), FAVOR-Bench (Tu et al., 2025a) VLRMBenc (Ruan et al., 2025), UrbanVideo-Bench (Zhao et al., 2025a) FortisAVQA (Ma et al., 2025b), VideoVista-CulturalLingo (Chen et al., 2025e)	VCR (Zellers et al., 2019), TDIUC (Kafle & Kanan, 2017) MMPR (Wang et al., 2024g), ChartQA (Masry et al., 2022) SWAG (Zellers et al., 2018), LLaVA-CoT (Xu et al., 2024b) CLEVR (Johnson et al., 2016), Mulberry-260K (Yao et al., 2024a) ShareGPT4oReasoning (Zhang et al., 2024e), R1-Onevision (Yang et al., 2025e) Video-R1-data (Feng et al., 2025b), Visual-CoT (Shao et al., 2024)
	Domain-specific Reasoning	MathVista (Lu et al., 2024), MATH-Vision (Wang et al., 2024c) VLM-Bench (Zheng et al., 2022), GemBench (Garcia et al., 2024) GeoQA (Chen et al., 2022a), VIMA-Bench (Jiang et al., 2022) WorldSimBench (Qin et al., 2024), WorldModelBench (Li et al., 2025c) ScienceQA (Lu et al., 2022), ChartQA ((Masry et al., 2022)) MathQA (Amini et al., 2019), Habitat (Savva et al., 2019) AI2-THOR (Kolve et al., 2017), Gibson (Xia et al., 2018) iGibson (Li et al., 2021a), Isaac Lab (Mittal et al., 2023) VCBench (Wang et al., 2025h), VCT (Götting et al., 2025) 3MDBench (Sviridov et al., 2025), PuzzleBench (Zhang et al., 2025e) ColorBench (Liang et al., 2025b), VisualPuzzles (Song et al., 2025d) Plot2XML (Cui et al., 2025), NoTeS-Bank (Pal et al., 2025) EIBench (Lin et al., 2025b), XLRS-Bench (Wang et al., 2025b) STI-Bench (Li et al., 2025i), EgoToM (Li et al., 2025k) DomainCQA (Zhong et al., 2025), MMCR-Bench (Yan et al., 2025a), Misleading ChartQA (Chen et al., 2025g) FlowVerse (Chen et al., 2025a), VisNumBench (Weng et al., 2025) MicroVQA (Burgess et al., 2025), MPBench (Xu et al., 2025e) Open3DVQA (Zhan et al., 2025), ProBench (Yang et al., 2025d) Chart-HQA (Chen et al., 2025c), MMSciBench (Ye et al., 2025)	Habitat (Savva et al., 2019), AI2-THOR (Kolve et al., 2017) Gibson (Xia et al., 2018), GeoQA (Chen et al., 2022a) Isaac Lab (Mittal et al., 2023), ProcTHOR (Deitke et al., 2022) CALVIN (Mees et al., 2022), SRM&SRMEval (Miao et al., 2025)
Multimodal Planning	GUI Navigation	WebArena (Zhou et al., 2024c), Mind2Web (Deng et al., 2023) VisualWebBench (Liu et al., 2024b), OSWorld (Xie et al., 2024a) OmniACT (Kapoor et al., 2024), VisualAgentBench (Liu et al., 2024d) LlamaTouch (Zhang et al., 2024d), Windows Agent Arena (Bonatti et al., 2024) Ferret-UI (You et al., 2024), WebShop (Yao et al., 2022) SWE-BENCH M (Yang et al., 2024b), MineDojo (Fan et al., 2022) TeamCraft (Long et al., 2024), V-MAGE (Zheng et al., 2025d) TongUI (Zhang et al., 2025a), BEARCUBS (Song et al., 2025c)	AMEX (Chai et al., 2024), RiCo (Deka et al., 2017) WebSRC (Chen et al., 2021), E-ANT (Wang et al., 2024d) AndroidEnv (Toyama et al., 2021), GUI-World (Chen et al., 2024b) MBE-ARI (Noronha et al., 2025)
	Embodied and Simulated Environments	MineDojo (Fan et al., 2022), MuEP (Li et al., 2024g) GVCCI (Kim et al., 2023), BEHAVIOR-1K (Li et al., 2024c) Habitat 3.0 (Puig et al., 2024), SAPIEN (Xiang et al., 2020) HomeRobot (Yenamandra et al., 2023), HoloAssist (Wang et al., 2023b) DrivingDojo (Rietsch et al., 2022), WolfBench (Qiao et al., 2024) MBE-ARI (Noronha et al., 2025), VisEscape (Lim et al., 2025)	MineDojo (Fan et al., 2022), Habitat 3.0 (Puig et al., 2024) SAPIEN (Xiang et al., 2020), HomeRobot (Yenamandra et al., 2023) HoloAssist (Wang et al., 2023b), DrivingDojo (Rietsch et al., 2022) OmmiHD-Scenes (Zheng et al., 2025c)

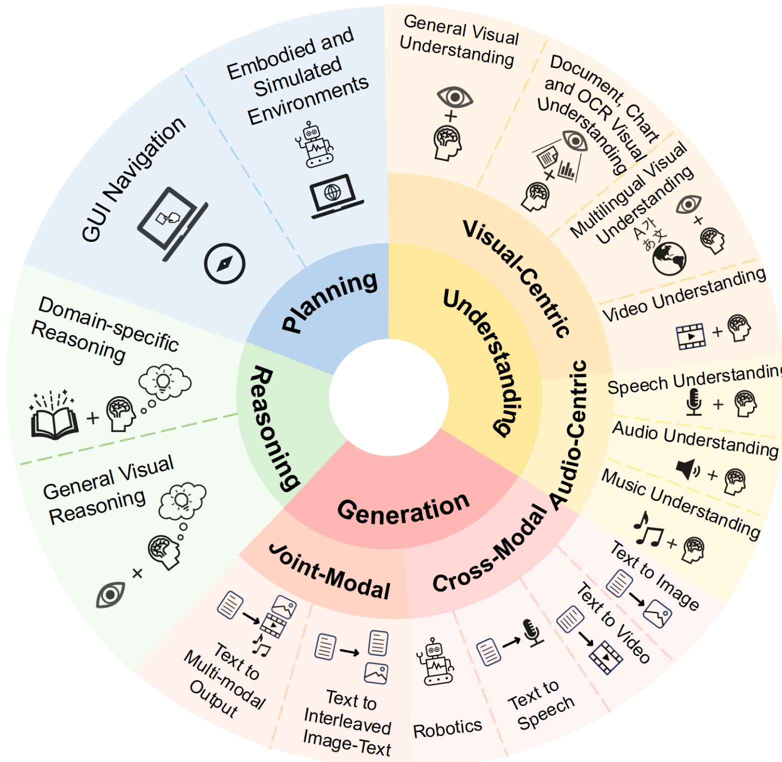

Figure 10: The outlines of datasets and benchmarks. We reorganize the multimodal datasets and benchmarks into four main categories: Understanding, Generation, Reasoning, and Planning.

5.1 Multimodal Understanding

Multimodal Understanding refers to the ability of models to process and interpret information from multiple modalities, such as visual and auditory data, to perform tasks that require comprehension, reasoning, and generation. These tasks are crucial for developing models capable of interacting with and responding to the real world in a more human-like manner. Based on the task definition, existing multimodal understanding tasks can be roughly categorized into two main areas: **1)** Visual-Centric Understanding, which encompasses the model's ability to understand and reason about visual content, and **2)** Audio-Centric Understanding, which focuses on tasks involving audio, such as speech, music, and environmental sounds.

5.1.1 Visual-Centric Understanding

Visual-centric understanding evaluates a model's ability to comprehend and reason about visual data, such as images and videos, across a variety of specialized tasks. These tasks can be broadly categorized into the following domains: general visual understanding, document and chart interpretation, multilingual visual reasoning, video understanding, mathematical and scientific reasoning, and comprehensive benchmarks. Each domain addresses different facets of visual understanding, from object recognition and spatial reasoning in natural images to the interpretation of structured visual data, such as documents and graphs. Below, we explore each of these categories in detail, highlighting their key features and challenges.

General Visual Understanding General visual question-answering (VQA) datasets have evolved significantly in both complexity and scope. Early datasets, such as VQA (Kafle & Kanan, 2016) and GQA (Ainslie et al., 2023), primarily focused on object recognition, attribute identification, and simple spatial reasoning within natural images. These datasets typically contain image-question-answer triplets, with questions formatted simply (e.g., "What color is the car?"). The focus was largely on natural images and basic perception. More recent datasets, such as ALIGN (Jia et al., 2021) aim to address more complex visual-language tasks, including image-text alignment and multimodal representations. Visual Genome (Krishna et al., 2016) extends visual understanding by including relationships and object-level information, thus pushing the boundaries of reasoning. The LAION-400M dataset (Schuhmann et al., 2021), one of the largest collections of image-text pairs, enables large-scale training for visual-language models. The LAION-5B dataset (Schuhmann et al., 2022) provides a strong dataset for large-scale image-text representations, and FILIP (Yao et al., 2021) and YFCC100M (Thomee et al., 2016) integrates both vision and language, enhancing models' performance across diverse benchmarks.

Document, Chart, and OCR Visual Understanding Document, chart, and OCR-based VQA datasets form a specialized domain focusing on understanding structured visual information that includes textual elements. Document VQA, exemplified by DocVQA (Mathew et al., 2021), targets document understanding, requiring models to locate and interpret text within documents to answer questions. Chart VQA, such as DVQA (Kafle et al., 2018), focuses on interpreting visual data representations, including bar charts, line graphs, and pie charts, testing the model's ability to understand these structures. OCR-VQA datasets like TextVQA (Singh et al., 2019) and OCR-VQA (Mishra et al., 2019) emphasize reading and reasoning about text embedded within natural images. These datasets share several distinctive characteristics: 1) the critical integration of OCR with visual understanding, 2) multi-step reasoning that combines both textual and visual elements, and 3) domain-specific knowledge about document structures, chart conventions, or text layouts. Unlike general VQA datasets, these collections heavily emphasize the interplay between visual and textual content, requiring models to bridge modalities in more structured contexts. Additionally, datasets like AI2D (Hiippala et al., 2021) focus on diagrams and structured visual representations, enhancing reasoning over graphical content.

Multilingual Visual Understanding Multilingual visual understanding datasets cater to the increasing demand for language diversity in multimodal systems. Datasets like CMMLU (Li et al., 2024f), C-Eval (Huang et al., 2023c), Exams-v (Das et al., 2024), M3exam (Zhang et al., 2023e), VideoVista-CulturalLingo (Chen et al., 2025e), and MTVQA (Tang et al., 2024) cover beyond English-centric VQA systems. These datasets are characterized by: 1) integration of questions and annotations in multiple languages, covering various language families, 2) testing visual understanding and linguistic capabilities across different cultural contexts, and 3) requiring models to understand visual concepts that may have specific cultural interpretations or references. Unlike single-language VQA datasets, these multilingual datasets evaluate and enhance the cross-lingual transfer abilities of MLLMs.

Video Understanding Video understanding datasets, e.g., ActivityNet-QA (Yu et al., 2019a) and Perception-Test (Patraucean et al., 2023), are increasingly used for training and evaluating models in dynamic visual tasks. These datasets, compared to static image datasets, require models to address time-based understanding, involving dynamic visual features across multiple frames. They include annotations for actions, events, and temporal relationships, and cover diverse video durations, ranging from short clips to several-minute-long videos. Existing video evaluation datasets have expanded to tackle challenges such as the scientific domain (e.g., Video-MMMU (Hu et al., 2025b)), long video domains (e.g., Video-MME (Fu et al., 2024a)), and comprehensive video understanding and reasoning (e.g., VideoVista (Li et al., 2024k)). VideoVista provides a versatile benchmark featuring 14 categories of videos with durations from a few seconds to over 10 minutes and encompasses 19 understanding tasks and 8 reasoning tasks. It utilizes an automatic annotation framework powered by GPT-4o, enhancing its scalability and diversity. Datasets like YouTube8M (Abu-El-Haija et al., 2016) have become foundational for large-scale video classification and multimodal understanding. Additionally, VidGen-1M (Tan et al., 2024b) and WebVid (Bain et al., 2022) serve as training datasets and focus on enhancing video comprehension by integrating multimodal text and visual signals.

Comprehensive Benchmarks Integrated evaluation benchmarks, such as MMBench (Liu et al., 2024f), Seed-Bench (Li et al., 2023c), and MME-RealWorld (Zhang et al., 2024f), have emerged to provide a more holistic evaluation of existing multimodal models. These benchmarks test how well models integrate visual and linguistic understanding in real-world scenarios, including 1) multidimensional evaluation frameworks that assess various aspects of visual understanding, from perception to reasoning and knowledge integration, 2) carefully designed questions aimed at exploring specific abilities and identifying weaknesses, and 3) standardized evaluation pipelines for fair comparison across models. Unlike early task-specific datasets, these benchmarks offer a comprehensive measure of models' overall capabilities.

Visual-centric Understanding emphasizes models' abilities to process and reason about visual data, from basic object recognition in images to complex multimodal reasoning in videos and documents. By addressing various specialized tasks, such as general visual understanding, document interpretation, multilingual reasoning, and video comprehension, these benchmarks provide a comprehensive view of a model's visual capabilities. These evaluations are essential for ensuring that models can integrate visual perception with reasoning, which is critical for real-world applications.

5.1.2 Audio-Centric Understanding

Audio-Centric Understanding refers to the evaluation of models' capabilities in processing, interpreting, and responding to various forms of audio input, such as speech, environmental sounds, and music. As these modalities become increasingly integral to machine learning tasks, evaluating how well models understand and interact with audio data has become a key focus. The evaluation spans different aspects of speech, audio, and music understanding, with various benchmarks and datasets designed to assess accuracy, translation, emotion recognition, and general comprehension in audio-related tasks. These evaluations help gauge the effectiveness of models in understanding the full range of audio data encountered in real-world applications.

Speech Understanding Speech evaluation datasets play a crucial role in assessing models' performance in the audio domain. These datasets primarily measure whether a model can accurately and clearly understand human speech in real-world settings. Existing datasets evaluate speech understanding from several perspectives: 1) Accuracy of speech recognition: Librispeech (Panayotov et al., 2015) is a dataset of audiobooks read by various speakers, serving as a widely used evaluation metric for English speech recognition. Common Voice (Ardila et al., 2020) collects voice recordings from volunteers globally, providing a diverse voice dataset for model training. The Aishell (Bu et al., 2017) series is the standard for Chinese speech recognition. Fleurs (Conneau et al., 2022) evaluates speech recognition and speech-to-text translation models across multiple languages. 2) Speech multilingual translation tasks: CoVoST2 (Wang et al., 2020) is a multilingual speech-to-text translation dataset that evaluates models' real-time speech recognition translation capabilities. 3) Emotion recognition: The MELD (Poria et al., 2019) dataset assesses models' ability to recognize emotions in speech, using emotional voices from multiple speakers in TV dramas. These datasets comprehensively assess models' ability to understand speech, considering factors such as content accuracy, diverse speech tasks, and additional acoustic information.

Audio Understanding Environmental sound understanding is another essential aspect of audio comprehension, involving the extraction and recognition of information from non-human voices. Compared to human speech, environmental sounds provide more complex and varied information. Mainstream evaluation datasets primarily assess audio understanding in two key areas: 1) Audio captioning: Clotho (Drossos et al., 2020) contains sounds from free sound platforms, primarily used for the audio captioning task. Similarly, AudioCaps (Kim et al., 2019), sourced from the AudioSet dataset, also focuses on audio captioning and has a broader application scope. 2) Audio question answering (AQA): ClothoAQA (Lipping et al., 2022) is a crowdsourced dataset designed for the AQA task and AQUALLM (Behera et al., 2023) is constructed by an automatic audio QA generation framework based on LLMs. These benchmarks include various audio types paired with questions and answers, helping models learn to understand audio content and generate accurate responses to audio-related questions.

Music Understanding Music, with its structural characteristics and complex variations, has become a significant area of research in audio understanding. Two primary directions are considered in music evaluation: Mainstream datasets like MusicNet (Thickstun et al., 2017) and NSynth (Engel et al., 2017) evaluate models' ability to recognize music theory elements such as instruments, notes, pitches, and rhythms in the audio. Additionally, MusicCaps (Agostinelli et al., 2023) and MusicBench (Melechovský et al., 2024) are used for captioning entire musical tracks, testing models' ability to understand both the detailed content and overall structure of music compositions.

Comprehensive Benchmarks As Large Audio-Language Models (LALMs) continue to evolve, more models now possess the ability to understand both speech and diverse sounds. Consequently, researchers are proposing new evaluation benchmarks to comprehensively assess models' audio understanding capabilities. VoiceBench (Chen et al., 2024h) focuses on models' ability to understand speech in varied contexts, including evaluations of basic capabilities, colloquial expressions, and performance in noisy environments. AudioBench (Wang et al., 2024a) integrates diverse speech tasks (e.g., Automatic Speech Recognition, Speech Question Answering), sound tasks (e.g., Audio Captioning, Audio Question Answering), and tasks related to human voices (e.g., accent, age, and gender). Air-Bench (Yang et al., 2024d) and MMAU (Sakshi et al., 2024) expand upon this by including music tasks in their evaluations. SD-eval (Ao et al., 2024) combines speech tasks with environmental sound tasks, enabling models to understand complex, mixed audio scenarios.

These benchmarks not only incorporate earlier evaluation methods but also provide a more comprehensive framework for assessing speech understanding across a wide range of real-world applications.

Audio-Centric Understanding offers a comprehensive framework for evaluating models' capabilities in processing and understanding audio data. It spans tasks from speech recognition to environmental sound and music interpretation. These evaluations are crucial for ensuring models' versatility and effectiveness in real-world applications, advancing their ability to handle complex audio data.

5.2 Multimodal Generation

Multimodal Generation is a key capability of Multimodal Reasoning Models, encompassing the creation of novel content across different data types, such as text, images, audio, or video. This generative ability is critical not only for creative applications but also for tasks where models need to communicate their understanding or reasoning results in a multimodal format.

These tasks can be broadly categorized based on how information flows between modalities and the nature of the generated output: **(1)** Cross-modal Generation, which evaluates a model's ability to generate content in one modality based on input from another; and **(2)** Joint Multimodal Generation, which assesses a model's ability to simultaneously generate content across multiple modalities.

5.2.1 Cross-modal Generation

Cross-modal generation involves tasks where models generate content in one modality based on input from another. This includes tasks like text-to-image, text-to-video, and text-to-speech generation, where models must effectively map one type of input (e.g., text) to a different form (e.g., image, video, or speech). These tasks challenge models to transform and align information from one modality to another, often requiring the handling of complex or conditional prompts. In this section, we explore how datasets and benchmarks have been developed to evaluate model performance across various cross-modal tasks, focusing on alignment, coherence, and semantic generation.

Text to Image The field of text-to-image generation (T2I) has seen significant advancements, driven by diverse datasets and benchmarks tailored to tasks such as text-to-image generation, editing, and conditional generation.

For text-to-image generation, datasets like MSCOCO (30K) (Lin et al., 2014a), CC12M (Changpinyo et al., 2021), and Flickr30k (Plummer et al., 2017) offer large-scale, general-purpose image-text pairs, emphasizing everyday scenes and objects. In contrast, datasets like RedCaps (Desai et al., 2021) and COMMONPOOL (Gadre et al., 2023) introduce more complex text descriptions and higher-resolution images. Benchmarks such as GenEval (Ghosh et al., 2023) and ELLA (Hu et al., 2024a) focus on evaluating text-to-image alignment, assessing how accurately the generated images match the textual descriptions. Meanwhile, GenAI-Bench (Li et al., 2024a) and T2I-CompBench++ (Huang et al., 2023a) emphasize the handling of complex prompts and object interactions, highlighting the need for effective compositional generation and improved semantic alignment.

For text-to-image editing, datasets like MagicBrush (Zhang et al., 2023d), InstructPix2Pix (Brooks et al., 2023), and HQ-Edit (Hui et al., 2024) focus on instruction-based editing, with HQ-Edit extending tasks to high-definition images. UltraEdit (Zhao et al., 2024a) and SEED-Data-Edit (Ge et al., 2024) introduce multi-turn editing tasks, improving training for large language models (LLMs) in multi-turn dialogues. These datasets assess the varying demands of image editing, with MagicBrush evaluating creative aspects and Emu Edit (Sheynin et al., 2023) focusing on precision and coherence in high-quality edits based on textual instructions.

For conditional text-to-image generation, datasets like ADE20K (Zhou et al., 2016) and CocoStuff (Caesar et al., 2016) offer detailed segmentation maps and scene parsing annotations, enabling models to generate images with specific scene structures. UniControl (Qin et al., 2023) introduces more comprehensive data, requiring models to handle multiple conditional inputs simultaneously. Benchmarks like UniCombine (Wang et al., 2025c) focus on evaluating instruction execution completeness, visual coherence, and consistency with constraints.

Text to Video In text-to-video generation, high-quality datasets and comprehensive benchmarks are critical for advancing research. Datasets like VidGen-1M (Tan et al., 2024b), OpenVid-1M (Nan et al., 2024), and VidProM (Wang & Yang, 2024) cover a wide range of video content and corresponding descriptive texts.

Benchmarking tools such as AIGCBench (Fan et al., 2019), EvalCrafter (Liu et al., 2024e), and VBench (Huang et al., 2024a) evaluate models across various metrics like relevance, coherence, and visual quality. Specialized benchmarks like VideoScore (He et al., 2024), WorldSimBench (Qin et al., 2024), and WorldScore (Duan et al., 2025) expand evaluation to cover video quality and real-world accuracy, with VideoScore assessing user satisfaction.

Text to Speech Text-to-speech (TTS) generation has benefited from high-caliber datasets and benchmarks that enable the development of Large Audio-Language Models (LALMs). Early models used synthetic datasets to evaluate speech dialogue capabilities, employing datasets like LlaMA-Questions (Nachmani et al., 2024), Web Questions (Berant et al., 2013), and Trivia QA (Joshi et al., 2017). Evaluations were based on comparing word error rates and accuracy between text and audio outputs. Recent benchmarks like ADU-Bench (Gao et al., 2024b) assess speech dialogue capabilities across regular, professional, multilingual, and ambiguous scenarios, while URO-Bench (Yan et al., 2025b) includes evaluations of speech style, such as intonation and emotion.

Robotics In robotics, datasets and benchmarks provide high-fidelity, multi-modal environments for evaluating model performance. Datasets like ThreeDWorld (Gan et al., 2021) and GAIA-1 (Hu et al., 2023) offer interactive simulation platforms for robotics tasks like autonomous driving. On the benchmark side, Genesis (Engelcke et al., 2019) provides a standardized evaluation framework to assess models across a range of robotics tasks, ensuring real-world applicability.

In summary, cross-modal generation is a pivotal area of multimodal AI, focusing on tasks such as text-to-image, text-to-video, and text-to-speech generation. These tasks challenge models to transform and align information across modalities. As advancements continue, the focus is on improving the handling of complex prompts, multi-step reasoning, and semantic alignment, with models poised to perform increasingly sophisticated transformations and interactions across modalities.

5.2.2 Joint Multimodal Generation

Joint multimodal generation refers to the simultaneous creation of content across multiple modalities, such as generating both text and images or combining text, audio, and video into a cohesive output. This presents additional complexity as models must ensure coherence and alignment between the generated modalities. Tasks like text-to-interleaved image-text and text-to-multimodal output exemplify this, requiring models to generate content that complements and fits within the broader context of the narrative. Specialized datasets and benchmarks have been developed to support these tasks, providing a rich environment for training models to create contextually relevant multimodal outputs.

Text to Interleaved Image-Text The development of multimodal large language models (MLLMs) has significantly advanced interleaved image-text generation, with datasets like MM-Interleaved (Tian et al., 2024) and ANOLE (Chern et al., 2024) supporting model training with high-quality annotated image-text pairs. These datasets emphasize the need for models to generate contextually relevant and visually coherent content. Benchmarks like InterleavedEval (Liu et al., 2024c) and OpenLEAF (An et al., 2024) focus on evaluating models' ability to generate coherent and aligned image-text pairs, while OpenING (Zhou et al., 2024a) provides a more diverse set of tasks to assess interleaved image-text generation.

Text to Multimodal Output Recent developments in text-to-multimodal output focus on enhancing multimodal generation by combining cross-modal and joint multimodal data. Models like NextGPT (Wu et al., 2024a) and DreamFactory (Xie et al., 2024b) leverage training-free approaches to transform text into multimodal stories, integrating video evaluation benchmarks like Vbench. Other models, such as EVA (Chi et al., 2024), incorporate embodied world models to simulate and anticipate events in video sequences based on text inputs.

In summary, joint multimodal generation involves the simultaneous creation of content across multiple modalities, requiring models to maintain coherence and alignment between them. As research advances, future developments will likely focus on improving intermodal coherence, adaptability, and seamless generation, opening up new possibilities for dynamic, multi-dimensional content creation and interactive user experiences.

5.3 Multimodal Reasoning

Multimodal reasoning goes beyond simple understanding or generation by requiring models to integrate information from multiple modalities. This allows them to make inferences, solve problems, and answer complex questions that demand a deeper comprehension of the relationships between different types of data.

We can broadly categorize multimodal reasoning models into two primary categories: **(1)** General Visual Reasoning, which evaluates a model's ability to understand visual content and apply general knowledge, logic, and common sense to solve tasks; and **(2)** Domain-specific Reasoning, which evaluates specific, often more technical, reasoning abilities such as mathematical problem-solving based on visual input.

5.3.1 General Visual Reasoning

General visual reasoning is one of the most critical capabilities in Multimodal Reasoning Models. It requires models not only to perceive visual information but also to comprehend, analyze, and reason about it using extensive knowledge, logical deduction, and common sense across a variety of scenarios.

To rigorously assess this ability, a wide range of benchmarks has been developed, each targeting distinct aspects of visual reasoning. Moving beyond simple question answering tasks (e.g., VQA), Visual Commonsense Reasoning benchmarks like VCR (Zellers et al., 2019), and specialized datasets like PhysBench (Chow et al., 2025) for physical reasoning, and VideoPhy (Bansal et al., 2024) for understanding physical common sense in videos, challenge models to apply everyday knowledge to interpret visual situations.

Ambitions for broader AI capabilities are reflected in Multimodal General Intelligence Benchmarks. These include comprehensive evaluations like MMBench (Liu et al., 2024f) (covering multilingual aspects), MMMU (Yue et al., 2024) (spanning diverse disciplines), AGIEval (Zhong et al., 2024b) (focused on human-centric evaluation), VideoVista (Li et al., 2024k) and MMStar (Chen et al., 2024f) (video-centric). These benchmarks incorporate visual reasoning as a key component alongside other modalities and tasks. Additionally, visual reasoning over diagrams and structured visuals is crucial, with benchmarks like AI2D (Kembhavi et al., 2016) and InfographicVQA (Mathew et al., 2022) challenging models to interpret spatial layouts, understand relationships, and extract information from diagrams, charts, and infographics.

A critical element in these benchmarks is the datasets used for training and evaluating models. Several datasets, such as SWAG (Zellers et al., 2018), are designed to train models to predict the likely continuation of actions in visual scenes. The LLava-CoT dataset (Xu et al., 2024b) enables models to reason about visual commonsense tasks by integrating large language models. CLEVR (Johnson et al., 2016) challenges models to perform complex reasoning on synthetic images of everyday objects. Other datasets like Mulberry-260K (Yao et al., 2024a) and ShareGPT4oReasoning (Zhang et al., 2024e) further train models for visual commonsense reasoning and multimodal dialogues, respectively.

Video-R1-data (Feng et al., 2025b) helps train models for reasoning about dynamic visual content in video sequences. Finally, Visual-CoT (Shao et al., 2024) supports training models requiring both visual understanding and reasoning across a variety of tasks. This dynamic and ever-evolving landscape of benchmarks and datasets is essential for advancing multimodal reasoning models.

5.3.2 Domain-specific Reasoning

Domain-specific reasoning benchmarks play a crucial role in evaluating the specialized reasoning capabilities of multimodal models in specific fields. For mathematical reasoning, datasets like MathVista (Lu et al., 2024) and MATH-Vision (Wang et al., 2024c) assess a model's ability to solve mathematical problems in visual contexts, requiring both visual understanding and mathematical inference. Similarly, benchmarks like ChartQA (Masry et al., 2022) and ScienceQA (Lu et al., 2022) focus on reasoning in specific domains.

In robotics, several benchmarks assess different aspects of embodied AI with a strong emphasis on reasoning. Simulation environments such as Habitat (Savva et al., 2019), AI2-THOR (Kolve et al., 2017), and iGibson (Li et al., 2021a) require agents to reason about navigation, interaction, and spatial understanding in complex 3D settings. Benchmarks like Isaac Lab (Mittal et al., 2023) and ProcTHOR (Deitke et al., 2022) focus on reasoning for manipulation tasks in diverse environments. Others, such as WebArena (Zhou et al., 2024c), test reasoning about web content, while language-guided reasoning is evaluated through benchmarks like CALVIN (Mees et al., 2022).

For physical reasoning, datasets like PhysBench (Chow et al., 2025), VideoPhy (Bansal et al., 2024), and CRAVE (Sun et al., 2025) assess models' understanding of physical laws and common sense across visual and video contexts. Finally, benchmarks like GAIA-1 (Hu et al., 2023) and RoboGen (Wang et al., 2024l) support the development of world models by evaluating how well models can simulate and reason about real-world dynamics and interactions.

These domain-specific benchmarks are crucial for pushing the boundaries of multimodal reasoning in specialized areas, enabling the development of more capable and intelligent multimodal reasoning models for specific applications.

In summary, multimodal reasoning is a critical area of AI that requires models to integrate and reason across multiple modalities, such as text, images, and video, to solve complex tasks. It is divided into General Visual Reasoning, which applies logic and common sense to visual content, and Domain-specific Reasoning, which evaluates specialized reasoning abilities in fields like mathematics, robotics, and physical laws. These tasks continually push multimodal reasoning models to evolve and approach human-level reasoning. As the field progresses, the future of multimodal reasoning will focus on creating more integrated systems capable of generalizing across diverse tasks and real-world scenarios, enabling more adaptive, intelligent, and versatile AI solutions.

5.4 Multimodal Planning

Multimodal planning benchmarks are essential for evaluating agents' abilities to integrate and process diverse inputs—such as visual, textual, and interactive data—while performing complex, multi-step tasks. These benchmarks cover a wide range of challenges, including web navigation, graphical user interfaces (GUIs), embodied environments, and open-ended simulations. By testing planning, reasoning, and adaptability, they provide a comprehensive view of an agent's capabilities. We categorize these benchmarks into two key areas to highlight their unique contributions and innovations.

5.4.1 GUI Navigation

Benchmarks in GUI navigation assess agents' abilities to plan and execute tasks across digital interfaces, requiring robust visual-language grounding and multi-step reasoning. WebArena (Zhou et al., 2024c) and Mind2Web (Deng et al., 2023) offer realistic web environments for navigation and information extraction, with Mind2Web further introducing cross-website tasks to test generalizability. VisualWebBench (Liu et al., 2024b) advances visual-intensive planning with 1.5K tasks focused on cross-page integration and element localization. Windows Agent Arena (Bonatti et al., 2024) evaluates cross-application planning in desktop environments, while Ferret-UI (You et al., 2024) focuses on grounded UI understanding for executing multi-step instructions. Benchmarks like WebShop (Yao et al., 2022) test visual-language grounding in simulated e-commerce environments. Similarly, OSWorld (Xie et al., 2024a) and OmniACT (Kapoor et al., 2024) provide real desktop OS environments, supporting cross-application workflows such as file manipulation and data processing. VisualAgentBench (Liu et al., 2024d) extends this paradigm by systematically evaluating large multimodal models across GUI, embodied, and visual design tasks, establishing a unified benchmark for planning and acting in visually rich digital environments. This is complemented by benchmarks like LlamaTouch (Zhang et al., 2024d), which scales mobile UI automation with 495 tasks, testing multi-step operations such as app navigation.

5.4.2 Embodied and Simulated Environments

Embodied and simulated environments emphasize planning in dynamic, interactive settings, where agents must adapt to physical or virtual worlds. MineDojo (Fan et al., 2022) provides an open-ended benchmark in Minecraft, enabling the training and evaluation of generalist agents across diverse tasks in a rich, interactive environment. Its flexibility supports multimodal planning for object interaction, navigation, and resource management. MuEP (Li et al., 2024g) focuses on embodied planning with visual-language inputs for tasks like path planning in simulated environments. GVCCI (Kim et al., 2023) introduces a lifelong learning framework that generates synthetic data to enhance visual grounding for language-guided robotic manipulation, achieving significant performance gains without human supervision. BEHAVIOR-1K (Li et al., 2024c) offers a dataset of 1,000 household activities, enabling robots to plan complex tasks by integrating visual, semantic, and action data. Habitat 3.0 (Puig et al., 2024) advances human-robot collaboration in simulated homes, supporting multimodal planning for navigation and interaction. SAPIEN (Xiang et al., 2020) provides a high-fidelity environment for part-based object manipulation, enhancing robotic planning precision. HomeRobot (Yenamandra et al., 2023) and its OpenVocabManip benchmark (Yenamandra et al., 2024) pioneer open-vocabulary mobile manipulation, combining language, perception, and action for generalizable tasks. HoloAssist (Wang et al., 2023b) captures egocentric human-robot interactions, facilitating planning for real-world collaborative tasks. DrivingDojo (Rietsch et al., 2022) tests dynamic decision-making in real-time driving scenarios using video and multi-agent data. Finally, V-MAGE (Zheng et al., 2025d) presents a game-based evaluation framework to assess Multimodal Large Language Models (MLLMs) on tasks like positioning, trajectory tracking, and visual memory, offering a novel approach to quantify planning abilities.

Multimodal planning benchmarks have made significant progress in evaluating agents across diverse tasks, from web navigation to embodied environments. However, challenges remain, such as long-horizon planning, handling noisy inputs, and real-world adaptability. Future benchmarks should focus on open-world environments, real-time human feedback, and collaborative planning, particularly in multi-agent or human-AI scenarios. Addressing these gaps will help advance the development of agents capable of handling unpredictable, real-world tasks with greater flexibility and generalization.

5.5 Evaluation Method

The mainstream evaluation methods currently include Exact/Fuzzy Match, Option Matching, LLM/MLLM Scoring, and Agentic Evaluation.

Exact/Fuzzy Matching Exact/Fuzzy Matching is primarily used in general open-ended VQA tasks including VQAv2 (Antol et al., 2015), OKVQA (Marino et al., 2019). These evaluation datasets typically provide multiple human-annotated candidate answers, and the predicted answers, processed by rules, are matched against the candidate answers either exactly or fuzzily. The final evaluation score is then calculated based on certain rules. For example, in VQAv2 (Antol et al., 2015) evaluation, a match with a single candidate answer is worth only 1/3 of a point, and a full score of 1 point requires a match with all three candidate answers; DocVQA (Mathew et al., 2021), on the other hand, uses Levenshtein distance to measure the accuracy of the predicted results.

Options Matching Due to the diversity of answers, exact and fuzzy matching methods are often unable to encompass all candidate options. To ensure fairness and accuracy in evaluation, the Options Matching approach was introduced. In this method, the system prompt includes several candidate options, and the model is required to select the most appropriate one. Moreover, to reduce the possibility of the model exhibiting a preference for a specific option during the selection process, works such as MMBench (Liu et al., 2024f) have adopted the CircularEval methodology to minimize stochastic variations in the evaluation.

LLM/MLLM Scoring Although option selection ensures fairness, it deviates considerably from the nature of open-ended questions and real-world scenarios. As a result, LLM-based evaluation methods have been introduced into the assessment of open-ended questions (Fu et al., 2024b; Zhang et al., 2023f). This approach involves inputting specific prompts, questions, standard answers, and model predictions into an LLM or MLLM, such as GPT-4o, to generate scores (Chen et al., 2024a; Xu et al., 2024d; Saad-Falcon et al., 2024). The prompts typically include scoring guidelines, reference examples, and other information designed to guide the model toward providing fair and balanced scores.

Agentic Evaluation During the evaluation process, the capabilities of a single model are inherently limited, which may lead to shortcomings when processing diverse multimodal information. To this end, agent-based approaches can leverage tools to mitigate the inherent limitations of the model itself. For instance, CIGEval (Wang et al., 2025f) expands the visual understanding capabilities of MLLMs by integrating a multi-functional toolbox, thereby enabling more fine-grained evaluation. Moreover, multi-agent discussions have shown effectiveness in downstream tasks by fostering consensus and producing more robust solutions (Xu et al., 2023b; Chen et al., 2024d; Xu et al., 2025f), a benefit that also extends to evaluation settings. Methods that leverage cooperative or adversarial interactions among multiple agents to assess outputs have demonstrated more reliable and interpretable evaluations (Chan et al., 2024; Li et al., 2024h; Zhao et al., 2024b; Liang et al., 2024).

6 Conclusion

In this paper, we survey the evolution of multimodal reasoning models, highlighting pivotal advancements and paradigm-shifting milestones in the field. While current models predominantly adopt a language-centric reasoning paradigm—delivering impressive results in tasks like visual question answering, visual math, and video understanding—critical challenges persist. Notably, visual-centric long reasoning (e.g., understanding 3D contexts, addressing complex visual information-seeking questions) and interactive multimodal reasoning (e.g., dynamic cross-modal dialogue or iterative feedback loops) remain underdeveloped frontiers requiring deeper exploration.

Building on empirical evaluations and experimental insights, we propose a forward-looking concept for inherently multimodal large models that transcend language-dominated architectures. Such models should prioritize three core capabilities: Multimodal Agentic Reasoning: Enabling proactive environmental interaction (e.g., embodied AI agents that learn through real-world trial and error). Omini-Modal Understanding and Generative Reasoning: Integrating any-modal semantics (e.g., aligning abstract concepts across vision, audio, and text) while resolving ambiguities in complex, open-world contexts; Producing coherent, context-aware outputs across modalities (e.g., generating diagrams from spoken instructions or synthesizing video narratives from text). By addressing these dimensions, future models could achieve human-like contextual adaptability, bridging the gap between isolated task performance and generalized, real-world problem-solving.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016. URL https://arxiv.org/abs/1609.08675.

Abduljaleel Adejumo, Faegheh Yeganli, Clifford Broni-bediako, Aoran Xiao, Naoto Yokoya, and Mennatullah Siam. A vision centric remote sensing benchmark, 2025. URL https://arxiv.org/abs/2503.15816.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. CoRR, abs/2301.11325, 2023. doi: 10.48550/ARXIV.2301.11325. URL https://doi.org/10.48550/arXiv.2301.11325.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pp. 4895–4901. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.298. URL https://doi.org/10.18653/v1/2023.emnlp-main.298.

Faris Alasmary and Saad Al-Ahmadi. SBVQA 2.0: Robust end-to-end speech-based visual question answering for open-ended questions. IEEE Access, 11:140967–140980, 2023. doi: 10.1109/ACCESS.2023.3339537.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. ArXiv preprint, abs/2204.14198, 2022. URL https://arxiv.org/abs/2204.14198.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 2357–2367. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1245. URL https://doi.org/10.18653/v1/n19-1245.

Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. OpenLEAF: A novel benchmark for open-domain interleaved image-text generation. In Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), pp. 11137–11144, Melbourne, VIC, Australia, 2024. ACM. doi: 10.1145/3664647.3685511.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6077–6086, 2018.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 39–48, 2016.

Anthropic. Introducing the model context protocol, April 2025. URL https://www.anthropic.com/news/model-context-protocol. Anthropic News. Accessed: 2025-04-17.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In International Conference on Computer Vision (ICCV), 2015.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/681fe4ec554beabdc9c84a1780cd5a8a-Abstract-Datasets_and_Benchmarks_Track.html.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pp. 4218–4222. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022. URL https://arxiv.org/abs/2104.00650.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. CoRR, abs/2406.03520, 2024. doi: 10.48550/ARXIV.2406.03520. URL https://doi.org/10.48550/arXiv.2406.03520.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022a. URL https://openreview.net/forum?id=p-BhZSz59o4.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems, 35:32897–32912, 2022b.

Swarup Ranjan Behera, Krishna Mohan Injeti, Jaya Sai Kiran Patibandla, Praveen Kumar Pokala, and Balakrishna Reddy Pailla. Aquallm: audio question answering data generation using large language models. arXiv preprint arXiv:2312.17343, 2023.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1533–1544. ACL, 2013. URL https://aclanthology.org/D13-1160/.

Jing Bi, Susan Liang, Xiaofei Zhou, Pinxin Liu, Junjia Guo, Yunlong Tang, Luchuan Song, Chao Huang, Guangyu Sun, Jinxi He, et al. Why reasoning matters? a survey of advancements in multimodal reasoning (v1). arXiv preprint arXiv:2504.03151, 2025.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. Advances in Neural Information Processing Systems, 37:5996–6051, 2024.

Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Jang, and Zack Hui. Windows agent arena: Evaluating multi-modal OS agents at scale. CoRR, abs/2409.08264, 2024. doi: 10.48550/ARXIV.2409.08264. URL https://doi.org/10.48550/arXiv.2409.08264.

Abhilekh Borah, Hasnat Md Abdullah, Kangda Wei, and Ruihong Huang. Clime: Evaluating multimodal climate discourse on social media and the climate alignment quotient (caq), 2025. URL https://arxiv.org/abs/2504.03906.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 18392–18402. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01764. URL https://doi.org/10.1109/CVPR52729.2023.01764.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017, pp. 1–5. IEEE, 2017. doi: 10.1109/ICSDA.2017.8384449. URL https://doi.org/10.1109/ICSDA.2017.8384449.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research, 2025. URL https://arxiv.org/abs/2503.13399.

Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1989–1998, 2019.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. arXiv preprint arXiv:1612.03716, 2016.

Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Dingyu Zhang, Peng Gao, Shuai Ren, and Hongsheng Li. AMEX: android multi-annotation expo dataset for mobile GUI agents. CoRR, abs/2407.17490, 2024. doi: 10.48550/ARXIV.2407.17490. URL https://doi.org/10.48550/arXiv.2407.17490.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Bhavik Chandna, Mariam Aboujenane, and Usman Naseem. Extremeaigc: Benchmarking lmm vulnerability to ai-generated extremist content, 2025. URL https://arxiv.org/abs/2503.09964.

Hsin-Yu Chang, Pei-Yu Chen, Tun-Hsiang Chou, Chang-Sheng Kao, Hsuan-Yun Yu, Yen-Ting Lin, and Yun-Nung Chen. A survey of data synthesis approaches. arXiv preprint arXiv:2407.03672, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3558–3568, 2021.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024a.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, Tianshuo Zhou, Yue Yu, Chujie Gao, Qihui Zhang, Yi Gui, Zhen Li, Yao Wan, Pan Zhou, Jianfeng Gao, and Lichao Sun. GUI-WORLD: A dataset for gui-oriented multimodal llm-based agents. CoRR, abs/2406.10819, 2024b. doi: 10.48550/ARXIV.2406.10819. URL https://doi.org/10.48550/arXiv.2406.10819.

Felix Chen, Hangjie Yuan, Yunqiu Xu, Tao Feng, Jun Cen, Pengwei Liu, Zeying Huang, and Yi Yang. Mathflow: Enhancing the perceptual flow of mllms for visual mathematical problems, 2025a. URL https://arxiv.org/abs/2503.16549.

Houlun Chen, Xin Wang, Hong Chen, Zihan Song, Jia Jia, and Wenwu Zhu. Grounding-prompter: Prompting llm with multimodal information for temporal sentence grounding in long videos. arXiv preprint arXiv:2312.17117, 2023a.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning, 2022a. URL https://arxiv.org/abs/2105.14517.

Jingxuan Chen, Derek Yuen, Bin Xie, Yuhao Yang, Gongwei Chen, Zhihao Wu, Li Yixing, Xurui Zhou, Weiwen Liu, Shuai Wang, et al. Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In NeurIPS 2024 Workshop on Open-World Agents, 2024c.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7066–7085, Bangkok, Thailand, August 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023b.

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

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. Omnixr: Evaluating omni-modality language models on reasoning across modalities. arXiv preprint arXiv:2410.12219, 2024e.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024f. URL http://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Lu Chen, Xingyu Chen, Zihan Zhao, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. CoRR, abs/2101.09465, 2021. URL https://arxiv.org/abs/2101.09465.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 16449–16469, Miami, Florida, USA, November 2024g. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.960. URL https://aclanthology.org/2024.findings-emnlp.960/.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794, 2022b.

Xiangnan Chen, Yuancheng Fang, Qian Xiao, Juncheng Li, Jun Lin, Siliang Tang, Yi Yang, and Yueting Zhuang. Chart-hqa: A benchmark for hypothetical question answering in charts, 2025c. URL https://arxiv.org/abs/2503.04095.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811, 2025d.

Xinyu Chen, Yunxin Li, Haoyuan Shi, Baotian Hu, Wenhan Luo, Yaowei Wang, and Min Zhang. Videovista-culturallingo: 360° horizons-bridging cultures, languages, and domains in video comprehension, 2025e. URL https://arxiv.org/abs/2504.17821.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, volume 12375 of Lecture Notes in Computer Science, pp. 104–120. Springer, 2020. doi: 10.1007/978-3-030-58577-8_7. URL https://doi.org/10.1007/978-3-030-58577-8_7.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. CoRR, abs/2410.17196, 2024h. doi: 10.48550/ARXIV.2410.17196. URL https://doi.org/10.48550/arXiv.2410.17196.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, and Jing Shao. Octavius: Mitigating task interference in MLLMs via loRA-moe. In The Twelfth International Conference on Learning Representations, 2024i. URL https://openreview.net/forum?id=rTDyN8yajn.

Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. arXiv preprint arXiv:2503.07523, 2025f.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 24185–24198, 2024j.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. arXiv preprint arXiv:2301.05226, 2023c.

Zixin Chen, Sicheng Song, Kashun Shum, Yanna Lin, Rui Sheng, and Huamin Qu. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering, 2025g. URL https://arxiv.org/abs/2503.18172.

Ziyang Chen, Xiaobin Wang, Yong Jiang, Jinzhi Liao, Pengjun Xie, Fei Huang, and Xiang Zhao. An adaptive framework for generating systematic explanatory answer in online q&a platforms. arXiv preprint arXiv:2410.17694, 2024k.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. From the least to the most: Building a plug-and-play visual reasoner via data synthesis. arXiv preprint arXiv:2406.19934, 2024a.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. arXiv preprint arXiv:2401.10935, 2024b.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, et al. Embodiedeval: Evaluate multimodal llms as embodied agents. arXiv preprint arXiv:2501.11858, 2025.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. arXiv preprint arXiv:2407.06135, 2024.

Xiaowei Chi, Hengyuan Zhang, Chun-Kai Fan, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-min Chan, Wei Xue, Wenhan Luo, Shanghang Zhang, and Yike Guo. Eva: An embodied world model for future video anticipation. arXiv preprint arXiv:2410.15461, 2024.

Rohan Choudhury, Koichiro Niinuma, Kris M. Kitani, and Laszlo A. Jeni. Zero-shot video question answering with procedural programs. In European Conference on Computer Vision (ECCV), pp. 1113–1130. Springer, 2024.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. CoRR, abs/2501.16411, 2025. doi: 10.48550/ARXIV.2501.16411. URL https://doi.org/10.48550/arXiv.2501.16411.

Konstantina Christakopoulou, Shibl Mourad, and Maja J. Mataric. Agents thinking fast and slow: A talker-reasoner architecture. CoRR, abs/2410.08328, 2024. doi: 10.48550/ARXIV.2410.08328. URL https://doi.org/10.48550/arXiv.2410.08328.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL https://arxiv.org/abs/2501.17161.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: few-shot learning evaluation of universal representations of speech. In IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023, pp. 798–805. IEEE, 2022. doi: 10.1109/SLT54892.2023.10023141. URL https://doi.org/10.1109/SLT54892.2023.10023141.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. IEEE Intelligent Transportation Systems Magazine, 2024.

Zhiqing Cui, Jiahao Yuan, Hanqing Wang, Yanshu Li, Chenxu Du, and Zhenglong Ding. Draw with thought: Unleashing multimodal reasoning for scientific diagram generation, 2025. URL https://arxiv.org/abs/2504.09479.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. arXiv preprint arXiv:2403.10378, 2024.

Joost C. F. de Winter, Dimitra Dodou, and Yke Bauke Eisma. System 2 thinking in openai's o1-preview model: Near-perfect performance on a mathematics exam. CoRR, abs/2410.07114, 2024. doi: 10.48550/ARXIV.2410.07114. URL https://doi.org/10.48550/arXiv.2410.07114.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied AI using procedural generation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/27c546ab1e4f1d7d638e6a8dfbad9a07-Abstract-Conference.html.

Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In Proceedings of the 30th Annual Symposium on User Interface Software and Technology, UIST '17, 2017.

Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07065.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samual Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural

Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/e00da03b685a0dd18fb6a08af0923de0-Abstract-round1.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, and Jiawei Wang. Write and paint: Generative vision-language models are unified modal learners. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/forum?id=HgQR0mXQ1_a.

Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024, pp. 10707–10715. ACM, 2024. doi: 10.1145/3664647.3681251. URL https://doi.org/10.1145/3664647.3681251.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following, 2025. URL https://arxiv.org/abs/2504.07957.

Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. Progressive multimodal reasoning via active retrieval. arXiv preprint arXiv:2412.14835, 2024a.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499, 2023.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. arXiv preprint arXiv:2411.14432, 2024b.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18166–18176, 2022.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. ArXiv preprint, abs/2303.03378, 2023. URL https://arxiv.org/abs/2303.03378.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, pp. 736–740. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9052990. URL https://doi.org/10.1109/ICASSP40776.2020.9052990.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. arXiv preprint arXiv:2504.00983, 2025.

Jesse H. Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In Doina Precup and Yee Whye Teh (eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pp. 1068–1077. PMLR, 2017. URL http://proceedings.mlr.press/v70/engel17a.html.

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. arXiv preprint arXiv:1907.13052, 2019.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 1999–2007. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00210. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Fan_Heterogeneous_Memory_Enhanced_Multimodal_Attention_Model_for_Video_Question_Answering_CVPR_2019_paper.html.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/74a67268c5cc5910f64938cac4526a90-Abstract-Datasets_and_Benchmarks.html.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In Forty-first International Conference on Machine Learning, 2024.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms, 2025a. URL https://arxiv.org/abs/2504.11536.

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

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. CoRR, abs/2405.21075, 2024a. doi: 10.48550/ARXIV.2405.21075. URL https://doi.org/10.48550/arXiv.2405.21075.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6556–6576, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365. URL https://aclanthology.org/2024.naacl-long.365/.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December

10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/56332d41d55ad7ad8024aac625881be7-Abstract-Datasets_and_Benchmarks.html.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel Bear, Dan Gutfreund, David D. Cox, Antonio Torralba, James J. DiCarlo, Josh Tenenbaum, Josh H. McDermott, and Dan Yamins. Threedworld: A platform for interactive multi-modal physical simulation. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/735b90b4568125ed6c3f678819b6e058-Abstract-round1.html.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv preprint arXiv:2306.08640, 2023.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation, 2025a. URL https://arxiv.org/abs/2503.19622.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. arXiv preprint arXiv:2411.19488, 2024a.

Junyuan Gao, Jiahe Song, Jiang Wu, Runchuan Zhu, Guanlin Shen, Shasha Wang, Xingjian Wei, Haote Yang, Songyang Zhang, Weijia Li, Bin Wang, Dahua Lin, Lijun Wu, and Conghui He. Pm4bench: A parallel multilingual multi-modal multi-task benchmark for large vision language model, 2025b. URL https://arxiv.org/abs/2503.18484.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. CoRR, abs/2412.05167, 2024b. doi: 10.48550/ARXIV.2412.05167. URL https://doi.org/10.48550/arXiv.2412.05167.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 9096–9105, 2024c.

Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. CoRR, abs/2410.01345, 2024. doi: 10.48550/ARXIV.2410.01345. URL https://doi.org/10.48550/arXiv.2410.01345.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S. Adve, and Yu-Xiong Wang. Agmmu: A comprehensive agricultural multimodal understanding and reasoning benchmark, 2025. URL https://arxiv.org/abs/2504.10568.

Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models. arXiv preprint arXiv:2304.07919, 2023.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. arXiv preprint arXiv:2405.04007, 2024.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. arXiv preprint arXiv:2411.19772, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Kye Gomez. Multimodal-tot. https://github.com/kyegomez/MultiModal-ToT, 2023.

Google. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=kxnoqaisCT.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark, 2022. URL https://arxiv.org/abs/2202.06767.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms, 2025. URL https://arxiv.org/abs/2504.17432.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Jingdong Chen, Ming Yang, et al. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance. arXiv preprint arXiv:2502.18778, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14953–14962, 2023.

Jasper Götting, Pedro Medeiros, Jon G Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. Virology capabilities test (vct): A multimodal virology qa benchmark, 2025. URL https://arxiv.org/abs/2504.16137.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. CoRR, abs/2206.06336, 2022. doi: 10.48550/ARXIV.2206.06336. URL https://doi.org/10.48550/arXiv.2206.06336.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pp. 2105–2123. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.127.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. Lang. Resour. Evaluation, 55(3):661–688, 2021. doi: 10.1007/S10579-020-09517-1. URL https://doi.org/10.1007/s10579-020-09517-1.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought. arXiv preprint arXiv:2305.13903, 2023.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. arXiv preprint arXiv:2502.04326, 2025.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14281–14290, 2024.

Kaiyuan Hou, Minghui Zhao, Lilin Xu, Yuang Fan, and Xiaofan Jiang. Tdbench: Benchmarking vision-language models in understanding top-down images, 2025. URL https://arxiv.org/abs/2504.03748.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. CoRR, abs/2309.17080, 2023. doi: 10.48550/ARXIV.2309.17080. URL https://doi.org/10.48550/arXiv.2309.17080.

Caiyu Hu, Yikai Zhang, Tinghui Zhu, Yiwei Ye, and Yanghua Xiao. Mcitebench: A benchmark for multimodal citation text generation in mllms, 2025a. URL https://arxiv.org/abs/2503.02589.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. CoRR, abs/2501.13826, 2025b. doi: 10.48550/ARXIV.2501.13826. URL https://doi.org/10.48550/arXiv.2501.13826.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1439–1449, 2021.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In Proceedings of the IEEE international conference on computer vision, pp. 804–813, 2017.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: equip diffusion models with LLM for enhanced semantic alignment. CoRR, abs/2403.05135, 2024a. doi: 10.48550/ARXIV.2403.05135. URL https://doi.org/10.48550/arXiv.2403.05135.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403, 2024b.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. arXiv preprint arXiv:2307.06350, 2023a.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation, 2025a. URL https://arxiv.org/abs/2307.06350.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/e425b75bac5742a008d643826428787c-Abstract-Conference.html.

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

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 24176–24184, 2025c.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023c. URL http://papers.nips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 21807–21818. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. CoRR, abs/2411.13503, 2024b. doi: 10.48550/ARXIV.2411.13503. URL https://doi.org/10.48550/arXiv.2411.13503.

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067, 2018.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. CoRR, abs/2404.09990, 2024. doi: 10.48550/ARXIV.2404.09990. URL https://doi.org/10.48550/arXiv.2404.09990.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 551–562, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f61d6947467ccd3aa5af24db320235dd-Abstract.html.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card. CoRR, abs/2412.16720, 2024. doi: 10.48550/ARXIV.2412.16720. URL https://doi.org/10.48550/arXiv.2412.16720.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. Dcot: Dual chain-of-thought prompting for large multimodal models. In The 16th Asian Conference on Machine Learning (Conference Track), 2024.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Bootstrapping vision-language learning with decoupled language pre-training. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/002262941c9edfd472a79298b2ac5e17-Abstract-Conference.html.

Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 11109–11116, 2020.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: general robot manipulation with multimodal prompts. CoRR, abs/2210.03094, 2022. doi: 10.48550/ARXIV.2210.03094. URL https://doi.org/10.48550/arXiv.2210.03094.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding, 2025. URL https://arxiv.org/abs/2504.04423.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=FlvtjAB0gl.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. CoRR, abs/1610.01465, 2016. URL http://arxiv.org/abs/1610.01465.

Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms, 2017. URL https://arxiv.org/abs/1703.09684.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. DVQA: understanding data visualizations via question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 5648–5656. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00592. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kafle_DVQA_Understanding_Data_CVPR_2018_paper.html.

Daniel Kahneman. Thinking, fast and slow. macmillan, 2011.

Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem AlShikh, and Ruslan Salakhutdinov. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVIII, volume 15126 of Lecture Notes in Computer Science, pp. 161–178. Springer, 2024. doi: 10.1007/978-3-031-73113-6_10. URL https://doi.org/10.1007/978-3-031-73113-6_10.

Fucai Ke, Zhixi Cai, Simindokht Jahangard, Weiqing Wang, Pari Delir Haghighi, and Hamid Rezatofighi. Hydra: A hyper agent for dynamic compositional visual reasoning. In European Conference on Computer Vision, pp. 132–149. Springer, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV, volume 9908 of Lecture Notes in Computer Science, pp. 235–251. Springer, 2016. doi: 10.1007/978-3-319-46493-0_15. URL https://doi.org/10.1007/978-3-319-46493-0_15.

M Abdul Khaliq, P Chang, M Ma, Bernhard Pflugfelder, and F Miletić. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. arXiv preprint arXiv:2404.12065, 2024.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 119–132. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1011. URL https://doi.org/10.18653/v1/n19-1011.

Dahun Kim, AJ Piergiovanni, Ganesh Mallya, and Anelia Angelova. Videocomp: Advancing fine-grained compositional and temporal alignment in video-text models, 2025. URL https://arxiv.org/abs/2504.03970.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. Advances in neural information processing systems, 31, 2018.

Junghyun Kim, Gi-Cheon Kang, Jaein Kim, Suyeon Shin, and Byoung-Tak Zhang. GVCCI: lifelong learning of visual grounding for language-guided robotic manipulation. In IROS, pp. 952–959, 2023. doi: 10.1109/IROS55552.2023.10342021. URL https://doi.org/10.1109/IROS55552.2023.10342021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. arXiv preprint arXiv:2401.13649, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: an interactive 3d environment for visual AI. CoRR, abs/1712.05474, 2017. URL http://arxiv.org/abs/1712.05474.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation. In Proceedings of the International Conference on Machine Learning (ICML), 2024. doi: 10.48550/arXiv.2312.14125. URL https://arxiv.org/pdf/2312.14125.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthinking: Slowdown attacks on reasoning llms. arXiv preprint arXiv:2502.02542, 2025.

EvolvingLMMs Lab. Multimodal open r1, 2025. URL https://github.com/EvolvingLMMs-Lab/open-r1-multimodal. Accessed: 2025-02-28.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9579–9589, 2024.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. arXiv preprint arXiv:2406.02030, 2024.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696, 2018.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio, 2024. URL https://arxiv.org/abs/2410.12787.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. CoRR, abs/2406.13743, 2024a. doi: 10.48550/ARXIV.2406.13743. URL https://doi.org/10.48550/arXiv.2406.13743.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/1e69ff56d0ebff0752ff29caaddc25dd-Abstract-Datasets_and_Benchmarks_Track.html.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023a. URL https://arxiv.org/abs/2305.03726.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models, 2023b. URL https://arxiv.org/abs/2311.17092.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. CoRR, abs/2307.16125, 2023c. doi: 10.48550/ARXIV.2307.16125. URL https://doi.org/10.48550/arXiv.2307.16125.

Boxun Li, Yadong Li, Zhiyuan Li, Congyi Liu, Weilin Liu, Guowei Niu, Zheyue Tan, Haiyang Xu, Zhuyu Yao, Tao Yuan, et al. Megrez-omni technical report. arXiv preprint arXiv:2502.15803, 2025a.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, C. Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. CoRR, abs/2108.03272, 2021a. URL https://arxiv.org/abs/2108.03272.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Jose Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1, 000 everyday activities and realistic simulation. CoRR, abs/2403.09227, 2024c. doi: 10.48550/ARXIV.2403.09227. URL https://doi.org/10.48550/arXiv.2403.09227.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542, 2025b.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. CoRR, abs/2502.20694, 2025c. doi: 10.48550/ARXIV.2502.20694. URL https://doi.org/10.48550/arXiv.2502.20694.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024d.

Guangyao Li, Henghui Du, and Di Hu. Avqa-cot: When cot meets question answering in audio-visual scenarios. In CVPR Workshops, 2024e.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: measuring massive multitask language understanding in chinese. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, pp. 11260–11285. Association for Computational Linguistics, 2024f. doi: 10.18653/V1/2024.FINDINGS-ACL.671. URL https://doi.org/10.18653/v1/2024.findings-acl.671.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34:9694–9705, 2021b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of Proceedings of Machine Learning Research, pp. 19730–19742. PMLR, 2023d. URL https://proceedings.mlr.press/v202/li23q.html.

Kanxue Li, Baosheng Yu, Qi Zheng, Yibing Zhan, Yuhui Zhang, Tianle Zhang, Yijun Yang, Yue Chen, Lei Sun, Qiong Cao, Li Shen, Lusong Li, Dapeng Tao, and Xiaodong He. Muep: A multimodal benchmark for embodied planning with foundation models. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024, pp. 129–138. ijcai.org, 2024g. URL https://www.ijcai.org/proceedings/2024/15.

Lei Li. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 513–522, 2024.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. Towards visual text grounding of multimodal large language model, 2025d. URL https://arxiv.org/abs/2504.04974.

Ruosen Li, Teerth Patel, and Xinya Du. PRD: Peer rank and discussion improve large language model based evaluations. Transactions on Machine Learning Research, 2024h. ISSN 2835-8856. URL https://openreview.net/forum?id=YVD1QqWRaj.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18061–18070, 2024i.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. arXiv preprint arXiv:2501.05366, 2025e.

Xingyu Li, Chen Gong, and Guohong Fu. Multimodal coreference resolution for chinese social media dialogues: Dataset and benchmark approach, 2025f. URL https://arxiv.org/abs/2504.14321.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958, 2025g.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, volume 12375 of Lecture Notes in Computer Science, pp. 121–137. Springer, 2020. doi: 10.1007/978-3-030-58577-8_8. URL https://doi.org/10.1007/978-3-030-58577-8_8.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. arXiv preprint arXiv:2501.15368, 2025h.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. arXiv preprint arXiv:2409.15272, 2024j.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding?, 2025i. URL https://arxiv.org/abs/2503.23765.

Yunxin Li, Baotian Hu, Xinyu Chen, Yuxin Ding, Lin Ma, and Min Zhang. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. arXiv preprint arXiv:2305.04530, 2023e.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. CoRR, abs/2406.11303, 2024k. doi: 10.48550/ARXIV.2406.11303. URL https://doi.org/10.48550/arXiv.2406.11303.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. Lmeye: An interactive perception network for large language models. IEEE Trans. Multim., 26:10952–10964, 2024l. doi: 10.1109/TMM.2024.3428317. URL https://doi.org/10.1109/TMM.2024.3428317.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025j.

Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. Egotom: Benchmarking theory of mind reasoning from egocentric videos, 2025k. URL https://arxiv.org/abs/2503.22152.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. arXiv preprint arXiv:2405.16919, 2024m.

Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. arXiv preprint arXiv:2409.13980, 2024n.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. CoRR, 2023.

Sirui Liang, Baoli Zhang, Jun Zhao, and Kang Liu. ABSEval: An agent-based framework for script evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 12418–12434, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.691. URL https://aclanthology.org/2024.emnlp-main.691/.

Xiwen Liang, Min Lin, Weiqi Ruan, Yuecheng Liu, Yuzheng Zhuang, and Xiaodan Liang. Memory-driven multimodal chain of thought for embodied long-horizon task planning. Openreview, 2025a.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness, 2025b. URL https://arxiv.org/abs/2504.10514.

Seungwon Lim, Sungwoong Kim, Jihwan Yu, Sungjae Lee, Jiwan Chung, and Youngjae Yu. Visescape: A benchmark for evaluating exploration-driven decision-making in virtual escape rooms, 2025. URL https://arxiv.org/abs/2503.14427.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. arXiv preprint arXiv:1405.0312, 2014a.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of Lecture Notes in Computer Science, pp. 740–755. Springer, 2014b. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Yujie Lin, Ante Wang, Moye Chen, Jingyao Liu, Hao Liu, Jinsong Su, and Xinyan Xiao. Investigating inference-time scaling for chain of multi-modal thought: A preliminary study. arXiv preprint arXiv:2502.11514, 2025a.

Yuxiang Lin, Jingdong Sun, Zhi-Qi Cheng, Jue Wang, Haomin Liang, Zebang Cheng, Yifei Dong, Jun-Yan He, Xiaojiang Peng, and Xian-Sheng Hua. Why we feel: Breaking boundaries in emotional reasoning with multimodal large language models, 2025b. URL https://arxiv.org/abs/2504.07521.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In 30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022, pp. 1140–1144. IEEE, 2022. URL https://ieeexplore.ieee.org/document/9909680.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. CoRR, abs/2402.08268, 2024a. doi: 10.48550/ARXIV.2402.08268. URL https://doi.org/10.48550/arXiv.2402.08268.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Jing Liu, Wenxuan Wang, Yisi Zhang, Yepeng Tang, Xingjian He, Longteng Guo, Tongtian Yue, and Xinlong Wang. Towards unified referring expression segmentation across omni-level visual target granularities, 2025a. URL https://arxiv.org/abs/2504.01954.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? CoRR, abs/2404.05955, 2024b. doi: 10.48550/ARXIV.2404.05955. URL https://doi.org/10.48550/arXiv.2404.05955.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 15489–15507, 2024c.

Wuyang Liu, Yi Chai, Yongpeng Yan, and Yanzhen Ren. Audio-visual event localization on portrait mode short videos, 2025b. URL https://arxiv.org/abs/2504.06884.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688, 2023b.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, Jiadai Sun, Xinyue Yang, Yu Yang, Zehan Qi, Shuntian Yao, Xueqiao Sun, Siyi Cheng, Qinkai Zheng, Hao Yu, Hanchen Zhang, Wenyi Hong, Ming Ding, Lihang Pan, Xiaotao Gu, Aohan Zeng, Zhengxiao Du, Chan Hee Song, Yu Su, Yuxiao Dong, and Jie Tang. Visualagentbench: Towards large multimodal models as visual foundation agents. CoRR, abs/2408.06327, 2024d. doi: 10.48550/ARXIV.2408.06327. URL https://doi.org/10.48550/arXiv.2408.06327.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 22139–22149. IEEE, 2024e. doi: 10.1109/CVPR52733.2024.02090. URL https://doi.org/10.1109/CVPR52733.2024.02090.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI, volume 15064 of Lecture Notes in Computer Science, pp. 216–233. Springer, 2024f. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Yuan Liu, Saihui Hou, Saijie Hou, Jiabao Du, Shibei Meng, and Yongzhen Huang. Omnidiff: A comprehensive benchmark for fine-grained image difference captioning, 2025c. URL https://arxiv.org/abs/2503.11093.

Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners, 2025d. URL https://arxiv.org/abs/2504.14239.

Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. Othink-mr1: Stimulating multimodal generalized reasoning capabilities through dynamic reinforcement learning. arXiv preprint arXiv:2503.16081, 2025e.

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

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. arXiv preprint arXiv:2403.12966, 2024g.

Qian Long, Zhi Li, Ran Gong, Ying Nian Wu, Demetri Terzopoulos, and Xiaofeng Gao. Teamcraft: A benchmark for multi-modal multi-agent systems in minecraft, 2024. URL https://arxiv.org/abs/2412.05255.

Cheng-Ze Lu, Xiaojie Jin, Qibin Hou, Jun Hao Liew, Ming-Ming Cheng, and Jiashi Feng. Delving deeper into data scaling in masked image modeling, 2023. URL https://arxiv.org/abs/2305.15248.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems, 29, 2016.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32, 2019.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom in for enhanced multimodal text-rich image understanding. arXiv preprint arXiv:2404.09797, 2024.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353, 2020.

Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. arXiv preprint arXiv:2412.02025, 2024.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023. URL https://arxiv.org/abs/2306.09093.

David Ma, Yuanxing Zhang, Jincheng Ren, Jarvis Guo, Yifan Yao, Zhenlin Wei, Zhenzhu Yang, Zhongyuan Peng, Boyu Feng, Jun Ma, Xiao Gu, Zhoufutu Wen, King Zhu, Yancheng He, Meng Cao, Shiwen Ni, Jiaheng Liu, Wenhao Huang, Ge Zhang, and Xiaojie Jin. Iv-bench: A benchmark for image-grounded video perception and reasoning in multimodal llms, 2025a. URL https://arxiv.org/abs/2504.15415.

Jie Ma, Zhitao Gao, Qi Chai, Jun Liu, Pinghui Wang, Jing Tao, and Zhou Su. Fortisavqa and maven: a benchmark dataset and debiasing framework for robust multimodal reasoning, 2025b. URL https://arxiv.org/abs/2504.00487.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In European Conference on Computer Vision, pp. 403–420. Springer, 2024.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Yu-Shi Zhu, Tong Zhang, Heyan Huang, and Xian-Ling Mao. Multi-modal retrieval augmented multi-modal generation: Datasets, evaluation metrics and strong baselines, 2025c. URL https://arxiv.org/abs/2411.16365.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. arXiv preprint arXiv:2501.07246, 2025d.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 3195–3204. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00331. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. Wikivideo: Article generation from multiple videos, 2025. URL https://arxiv.org/abs/2504.00939.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL https://doi.org/10.18653/v1/2022.findings-acl.177.

Matthew Massey and Abdullah-Al-Zubaer Imran. Earthscape: A multimodal dataset for surficial geologic mapping and earth surface analysis, 2025. URL https://arxiv.org/abs/2503.15625.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021, pp. 2199–2208. IEEE, 2021. doi: 10.1109/WACV48630.2021.00225. URL https://doi.org/10.1109/WACV48630.2021.00225.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022, pp. 2582–2591. IEEE, 2022. doi: 10.1109/WACV51458.2022.00264. URL https://doi.org/10.1109/WACV51458.2022.00264.

Oier Mees, Lukás Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics Autom. Lett., 7 (3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108. URL https://doi.org/10.1109/LRA.2022.3180108.

Jan Melechovský, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pp. 8293–8316. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.459. URL https://doi.org/10.18653/v1/2024.naacl-long.459.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning, 2025. URL https://github.com/ModalMinds/MM-EUREKA.

Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. arXiv preprint arXiv:2311.09241, 2023.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In The Twelfth International Conference on Learning Representations, 2023.

Bingchen Miao, Yang Wu, Minghe Gao, Qifan Yu, Wendong Bu, Wenqiao Zhang, Yunfei Li, Siliang Tang, Tat-Seng Chua, and Juncheng Li. Boosting virtual agent learning and reasoning: A step-wise, multi-dimensional, and generalist reward model with benchmark, 2025. URL https://arxiv.org/abs/2503.18665.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: visual question answering by reading text in images. In 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019, pp. 947–952. IEEE, 2019. doi: 10.1109/ICDAR.2019.00156. URL https://doi.org/10.1109/ICDAR.2019.00156.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14420–14431, 2024.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. IEEE Robotics Autom. Lett., 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034. URL https://doi.org/10.1109/LRA.2023.3270034.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734, 2021.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 18798–18806, 2024.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Anymal: An efficient and scalable any-modality augmented language model. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 1314–1332, 2024.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems, 36:25081–25094, 2023.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, R. J. Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered LLM. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=izrOLJov5y.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions, 2022. URL https://arxiv.org/abs/2204.00679.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. CoRR, abs/2407.02371, 2024. doi: 10.48550/ARXIV.2407.02371. URL https://doi.org/10.48550/arXiv.2407.02371.

Fei Ni, Jianye Hao, Shiguang Wu, Longxin Kou, Jiashun Liu, Yan Zheng, Bin Wang, and Yuzheng Zhuang. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13991–14000, 2024a.

Jinjie Ni, Yifan Song, Deepanway Ghosal, Bo Li, David Junhao Zhang, Xiang Yue, Fuzhao Xue, Zian Zheng, Kaichen Zhang, Mahir Shah, et al. Mixeval-x: Any-to-any evaluations from real-world data mixtures. arXiv preprint arXiv:2410.13754, 2024b.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In European Conference on Computer Vision, pp. 292–308. Springer, 2024.

Ian Noronha, Advait Prasad Jawaji, Juan Camilo Soto, Jiajun An, Yan Gu, and Upinder Kaur. Mbe-ari: A multimodal dataset mapping bi-directional engagement in animal-robot interaction, 2025. URL https://arxiv.org/abs/2504.08646.

OpenAI. Planning for agi and beyond, February 2023. URL https://openai.com/index/planning-for-agi-and-beyond/. Accessed: 2025-04-18.

OpenAI. Competitive programming with large reasoning models, 2025. URL https://arxiv.org/abs/2502.06807.

OpenAI. Introducing gpt-4.1 in the api, 2025a. https://openai.com/index/gpt-4-1/, Last accessed on 2025-04-14.

OpenAI. Introducing openai o3 and o4-mini, April 2025b. URL https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-04-18.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, pp. 1143–1151, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

Aniket Pal, Sanket Biswas, Alloy Das, Ayush Lodh, Priyanka Banerjee, Soumitri Chattopadhyay, Dimosthenis Karatzas, Josep Llados, and C. V. Jawahar. Notes-bank: Benchmarking neural transcription and search for scientific notes understanding, 2025. URL https://arxiv.org/abs/2504.09249.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. arXiv preprint arXiv:2403.17359, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Prabhat Pandey, Rupak Vignesh Swaminathan, K V Vijay Girish, Arunasish Sen, Jian Xie, Grant P. Strimel, and Andreas Schwarz. Sift-50m: A large-scale multilingual dataset for speech instruction fine-tuning, 2025. URL https://arxiv.org/abs/2504.09081.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract-Datasets_and_Benchmarks.html.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. CoRR, abs/2306.14824, 2023. doi: 10.48550/ARXIV.2306.14824. URL https://doi.org/10.48550/arXiv.2306.14824.

Rolf Pfister and Hansueli Jud. Understanding and benchmarking artificial intelligence: Openai's o3 is not agi. arXiv preprint arXiv:2501.07458, 2025.

Andreas Plesner, Turlan Kuzhagaliyev, and Roger Wattenhofer. Flip reasoning challenge, 2025. URL https://arxiv.org/abs/2504.12256.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Int. J. Comput. Vis., 123(1):74–93, 2017. doi: 10.1007/S11263-016-0965-7. URL https://doi.org/10.1007/s11263-016-0965-7.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1050. URL https://doi.org/10.18653/v1/p19-1050.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, et al. Judge anything: Mllm as a judge across any modality. arXiv preprint arXiv:2503.17489, 2025.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimir Vondrus, Théophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=4znwzG92CE.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. arXiv preprint arXiv:2402.04236, 2024.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2504.07956.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs, 2025. URL https://arxiv.org/abs/2504.13958.

Shuofei Qiao, Runnan Fang, Zhisong Qiu, Xiaobin Wang, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Benchmarking agentic workflow generation. arXiv preprint arXiv:2410.07869, 2024.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. arXiv preprint arXiv:2305.11147, 2023.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators. CoRR, abs/2410.18072, 2024. doi: 10.48550/ARXIV.2410.18072. URL https://doi.org/10.48550/arXiv.2410.18072.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In Proceedings of NAACL, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Glamm: Pixel grounding large multimodal model. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 13009–13018. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01236. URL https://doi.org/10.1109/CVPR52733.2024.01236.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. arXiv preprint arXiv:2405.14573, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26374–26383, 2024.

Sebastian Rietsch, Shih-Yuan Huang, Georgios D. Kontes, Axel Plinge, and Christopher Mutschler. Driver dojo: A benchmark for generalizable reinforcement learning for autonomous driving. CoRR, abs/2207.11432, 2022. doi: 10.48550/ARXIV.2207.11432. URL https://doi.org/10.48550/arXiv.2207.11432.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. arXiv preprint arXiv:2305.02317, 2023.

Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. Vlrmbench: A comprehensive and challenging benchmark for vision-language reward models, 2025. URL https://arxiv.org/abs/2503.07478.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 338–354, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. International Journal on Digital Libraries, 23(3): 289–301, 2022.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. CoRR, abs/2410.19168, 2024. doi: 10.48550/ARXIV.2410.19168. URL https://doi.org/10.48550/arXiv.2410.19168.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, Ahmed Ruby, Nicholas Popovič, Marek Suppa, Azmine Toushik Wasi, Ram Mohan Rao Kadiyala, Olga Tsymboi, Maksim Kostritsya, Bardia Soltani Moakhar, Gabriel da Costa Merlin, Otávio Ferracioli Coletti, Maral Jabbari Shiviari, MohammadAmin farahani fard, Silvia Fernandez, María Grandury, Dmitry Abulkhanov, Drishti Sharma, Andre Guarnier De Mitri, Leticia Bossatto Marchezi, Johan Obando-Ceron, Nazar Kohut, Beyza Ermis, Desmond Elliott, Enzo Ferrante, Sara Hooker, and Marzieh Fadaee. Kaleidoscope: In-language exams for massively multilingual vision evaluation, 2025. URL https://arxiv.org/abs/2504.07072.

Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 9338–9346. IEEE, 2019. doi: 10.1109/ICCV.2019.00943. URL https://doi.org/10.1109/ICCV.2019.00943.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. URL https://arxiv.org/abs/2111.02114.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII, volume 13668 of Lecture Notes in Computer Science, pp. 146–162. Springer, 2022. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024. URL https://arxiv.org/abs/2403.16999.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao (eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pp. 2556–2565. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1238. URL https://aclanthology.org/P18-1238/.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025. URL https://arxiv.org/abs/2504.07615.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. arXiv preprint arXiv:2311.10089, 2023.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models, 2025. URL https://arxiv.org/abs/2504.07951.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00851. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark, 2025a. URL https://arxiv.org/abs/2504.14693.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. arXiv preprint arXiv:2503.05592, 2025b.

Yixiao Song, Katherine Thai, Chau Minh Pham, Yapei Chang, Mazin Nadaf, and Mohit Iyyer. Bearcubs: A benchmark for computer-using web agents, 2025c. URL https://arxiv.org/abs/2503.07919.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge, 2025d. URL https://arxiv.org/abs/2504.10342.

DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuandong Tian, and Qinqing Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. CoRR, abs/2410.09918, 2024. doi: 10.48550/ARXIV.2410.09918. URL https://doi.org/10.48550/arXiv.2410.09918.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530, 2019.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491, 2018.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. arXiv preprint arXiv:2408.08862, 2024a.

Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U Tan, Deepanway Ghosal, Soujanya Poria, et al. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. arXiv preprint arXiv:2412.11974, 2024b.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024c. URL https://openreview.net/forum?id=mL8Q9OOamV.

Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency, 2025. URL https://arxiv.org/abs/2502.04076.

Rao Surapaneni, Miku Jha, Michael Vakoc, and Todd Segal. Announcing the agent2agent protocol (a2a), April 2025. URL https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/. Google Developers Blog. Long-Term Contributors. Accessed: 2025-04-17.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3mdbench: Medical multimodal multi-agent dialogue benchmark, 2025. URL https://arxiv.org/abs/2504.13861.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In European Conference on Computer Vision, pp. 305–322. Springer, 2024a.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation, 2024b. URL https://arxiv.org/abs/2408.02629.

Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3068–3078, 2023.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. MTVQA: benchmarking multilingual text-centric visual question answering. CoRR, abs/2405.11985, 2024. doi: 10.48550/ARXIV.2405.11985. URL https://doi.org/10.48550/arXiv.2405.11985.

Kimi Team. Kimi-vl technical report, 2025a. URL https://arxiv.org/abs/2504.07491.

OpenBMB MiniCPM-o Team. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2025b.

Qwen Team. Qvq: To see the world with wisdom, December 2024. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. arXiv preprint arXiv:2501.06186, 2025.

John Thickstun, Zaïd Harchaoui, and Sham M. Kakade. Learning features of music from scratch. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkFBJv9gg.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. Communications of the ACM, 59 (2):64–73, January 2016. ISSN 1557-7317. doi: 10.1145/2812802. URL http://dx.doi.org/10.1145/2812802.

Sebastian Thrun and Tom M Mitchell. Lifelong robot learning. Robotics and autonomous systems, 15(1-2): 25–46, 1995.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. arXiv preprint arXiv:2401.10208, 2024.

Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. Androidenv: A reinforcement learning platform for android. CoRR, abs/2105.13231, 2021. URL https://arxiv.org/abs/2105.13231.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding, 2025a. URL https://arxiv.org/abs/2503.14935.

Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. Vilbench: A suite for vision-language process reward modeling, 2025b. URL https://arxiv.org/abs/2503.20271.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models. CoRR, abs/2406.16020, 2024a. doi: 10.48550/ARXIV.2406.16020. URL https://doi.org/10.48550/arXiv.2406.16020.

Changhan Wang, Anne Wu, and Juan Miguel Pino. Covost 2: A massively multilingual speech-to-text translation corpus. CoRR, abs/2007.10310, 2020. URL https://arxiv.org/abs/2007.10310.

Chen Wang, Fei Xia, Wenhao Yu, Tingnan Zhang, Ruohan Zhang, C. Karen Liu, Li Fei-Fei, Jie Tan, and Jacky Liang. Chain-of-modality: Learning manipulation programs from multimodal human videos with vision-language-models, 2025a. URL https://arxiv.org/abs/2504.13351.

Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery?, 2025b. URL https://arxiv.org/abs/2503.23771.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, Bo Peng, and Yabiao Wang. Unicombine: Unified multi-conditional combination with diffusion transformer. arXiv preprint arXiv:2503.09277, 2025c.

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. Otc: Optimal tool calls via reinforcement learning, 2025d. URL https://arxiv.org/abs/2504.14870.

Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms, 2025e. URL https://arxiv.org/abs/2504.08358.

Jifang Wang, Xue Yang, Longyue Wang, Zhenran Xu, Yiyu Wang, Yaowei Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, and Min Zhang. A unified agentic framework for evaluating conditional image generation, 2025f. URL https://arxiv.org/abs/2504.07046.

Jize Wang, Ma Zerun, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. Gta: a benchmark for general tool agents. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024b.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024c. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Ke Wang, Tianyu Xia, Zhangxuan Gu, Yi Zhao, Shuheng Shen, Changhua Meng, Weiqiang Wang, and Ke Xu. E-ANT: A large-scale dataset for efficient automatic GUI navigation. CoRR, abs/2406.14250, 2024d. doi: 10.48550/ARXIV.2406.14250. URL https://doi.org/10.48550/arXiv.2406.14250.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pp. 19162–19170. AAAI Press, 2024e. doi: 10.1609/AAAI.V38I17.29884. URL https://doi.org/10.1609/aaai.v38i17.29884.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International conference on machine learning, pp. 23318–23340. PMLR, 2022a.

Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:2305.02677, 2023a.

Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. arXiv preprint arXiv:2403.16996, 2024f.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. CoRR, abs/2411.10442, 2024g. doi: 10.48550/ARXIV.2411.10442. URL https://doi.org/10.48550/arXiv.2411.10442.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/78b3e7836e3b7dea79d809b0c99cb097-Abstract-Datasets_and_Benchmarks_Track.html.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. ArXiv preprint, abs/2208.10442, 2022b. URL https://arxiv.org/abs/2208.10442.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In European Conference on Computer Vision, pp. 58–76. Springer, 2024h.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive AI assistants in the real world. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 20213–20224. IEEE, 2023b. doi: 10.1109/ICCV51070.2023.01854. URL https://doi.org/10.1109/ICCV51070.2023.01854.

Xuwu Wang, Qiwen Cui, Yunzhe Tao, Yiran Wang, Ziwei Chai, Xiaotian Han, Boyi Liu, Jianbo Yuan, Jing Su, Guoyin Wang, et al. Babelbench: An omni benchmark for code-driven analysis of multimodal and multistructured data. arXiv preprint arXiv:2410.00773, 2024i.

Yaoting Wang, Peiwen Sun, Yuanchao Li, Honggang Zhang, and Di Hu. Can textual semantics mitigate sounding object segmentation preference? 2024j.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. arXiv preprint arXiv:2401.06805, 2024k.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. In Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024l. URL https://openreview.net/forum?id=SQIDlJd3hN.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14749–14759, 2024m.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. arXiv preprint arXiv:2503.22952, 2025g.

Zhikai Wang, Jiashuo Sun, Wenqi Zhang, Zhiqiang Hu, Xin Li, Fan Wang, and Deli Zhao. Benchmarking multimodal mathematical reasoning with explicit visual dependency, 2025h. URL https://arxiv.org/abs/2504.18589.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning, 2025i. URL https://arxiv.org/abs/2504.20073.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904, 2021.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory, 2024n. URL https://arxiv.org/abs/2409.07429.

Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. Dreamrunner: Fine-grained storytelling video generation with retrieval-augmented motion adaptation. arXiv preprint arXiv:2411.16657, 2024o.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. 2025a.

Yimin Wei, Aoran Xiao, Yexian Ren, Yuting Zhu, Hongruixuan Chen, Junshi Xia, and Naoto Yokoya. Sarlang-1m: A benchmark for vision-language modeling in sar image understanding, 2025b. URL https://arxiv.org/abs/2504.03254.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. Visnumbench: Evaluating number sense of multimodal large language models, 2025. URL https://arxiv.org/abs/2503.14939.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Qi Wu, Quanlong Zheng, Yanhao Zhang, Junlin Xie, Jinguo Luo, Kuo Wang, Peng Liu, Qingsong Xie, Ru Zhen, Haonan Lu, and Zhenyu Yang. H2vu-benchmark: A comprehensive benchmark for hierarchical holistic video understanding, 2025. URL https://arxiv.org/abs/2503.24008.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In Proceedings of the International Conference on Machine Learning, pp. 53366–53397, 2024a.

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding, 2024b. URL https://arxiv.org/abs/2410.05249.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In European Conference on Computer Vision, pp. 164–182. Springer, 2024c.

Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 9068–9079. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00945. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Xia_Gibson_Env_Real-World_CVPR_2018_paper.html.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 11094–11104. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01111. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Xiang_SAPIEN_A_SimulAted_Part-Based_Interactive_ENvironment_CVPR_2020_paper.html.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. URL https://arxiv.org/abs/2409.11340.

Wang Xiaodong and Peng Peixi. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

Jingran Xie, Shun Lei, Yue Yu, Yang Xiang, Hui Wang, Xixin Wu, and Zhiyong Wu. Leveraging chain of thought towards empathetic spoken dialogue without corresponding question-answering data. arXiv preprint arXiv:2501.10937, 2025a.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models, 2025b. URL https://arxiv.org/abs/2504.03641.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. arXiv preprint arXiv:2410.11190, 2024.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. arXiv preprint arXiv:2408.11788, 2024b.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In International conference on machine learning, pp. 2397–2406. PMLR, 2016.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440, 2024a.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. CoRR, abs/2411.10440, 2024b. doi: 10.48550/ARXIV.2411.10440. URL https://doi.org/10.48550/arXiv.2411.10440.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks, 2023a. URL https://arxiv.org/abs/2306.04362.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? arXiv preprint arXiv:2501.11284, 2025a.

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

Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, Jun Song, and Bo Zheng. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning, 2025c. URL https://arxiv.org/abs/2504.12597.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025d. URL https://arxiv.org/abs/2504.15279.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. arXiv preprint arXiv:2406.08464, 2024c.

Zhaopan Xu, Pengfei Zhou, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification, 2025e. URL https://arxiv.org/abs/2503.12505.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. Towards reasoning in large language models via multi-agent peer review collaboration, 2023b. URL https://arxiv.org/abs/2311.08152.

Zhenran Xu, Senbao Shi, Baotian Hu, Longyue Wang, and Min Zhang. MultiSkill: Evaluating large multimodal models for fine-grained alignment skills. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 1506–1523, Miami, Florida, USA, November 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.81. URL https://aclanthology.org/2024.findings-emnlp.81/.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces, 2025f. URL https://arxiv.org/abs/2501.12909.

Yutaro Yamada, Khyathi Chandu, Yuchen Lin, Jack Hessel, Ilker Yildirim, and Yejin Choi. L3go: Language agents with chain-of-3d-thoughts for generating unconventional objects. arXiv preprint arXiv:2402.09052, 2024.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL https://arxiv.org/abs/2504.08066.

Dawei Yan, Yang Li, Qing-Guo Chen, Weihua Luo, Peng Wang, Haokui Zhang, and Chunhua Shen. Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning, 2025a. URL https://arxiv.org/abs/2503.18533.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. Uro-bench: A comprehensive benchmark for end-to-end spoken dialogue models. CoRR, abs/2502.17810, 2025b. doi: 10.48550/ARXIV.2502.17810. URL https://doi.org/10.48550/arXiv.2502.17810.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words, 2024a. URL https://arxiv.org/abs/2406.06040.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, Lars Liden, and Jianfeng Gao. Magma: A foundation model for multimodal AI agents. CoRR, abs/2502.13130, 2025a. doi: 10.48550/ARXIV.2502.13130. URL https://doi.org/10.48550/arXiv.2502.13130.

John Yang, Carlos E Jimenez, Alex L Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R Narasimhan, et al. Swe-bench multimodal: Do ai systems generalize to visual software domains? arXiv preprint arXiv:2410.03859, 2024b.

Juncheng Yang, Zuchao Li, Shuai Xie, Wei Yu, Shijun Li, and Bo Du. Soft-prompting with graph-of-thought for multi-modal representation learning. arXiv preprint arXiv:2404.04538, 2024c.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 1979–1998. Association for Computational Linguistics, 2024d. doi: 10.18653/V1/2024.ACL-LONG.109. URL https://doi.org/10.18653/v1/2024.acl-long.109.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560, 2025b.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Unisim: Learning interactive real-world simulators, 2024e. URL https://arxiv.org/abs/2310.06114.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. arXiv preprint arXiv:2407.08683, 2024f. URL https://arxiv.org/abs/2407.08683.

Shuo Yang, Siwen Luo, Soyeon Caren Han, and Eduard Hovy. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering, 2025c. URL https://arxiv.org/abs/2503.18491.

Yan Yang, Dongxu Li, Haoning Wu, Bei Chen, Liu Liu, Liyuan Pan, and Junnan Li. Probench: Judging multimodal foundation models on open-ended multi-domain expert tasks, 2025d. URL https://arxiv.org/abs/2503.06885.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025e. URL https://arxiv.org/abs/2503.10615.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Acvubench: Audio-centric video understanding benchmark, 2025f. URL https://arxiv.org/abs/2503.19951.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381, 2023.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 21–29. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.10. URL https://doi.org/10.1109/CVPR.2016.10.

Fanglong Yao, Changyuan Tian, Jintao Liu, Zequn Zhang, Qing Liu, Li Jin, Shuchao Li, Xiaoyu Li, and Xian Sun. Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. arXiv preprint arXiv:2308.06207, 2023a.

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

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training, 2021. URL https://arxiv.org/abs/2111.07783.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In NeurIPS, 2023b.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024b. URL https://arxiv.org/abs/2406.12045.

Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in language models. arXiv preprint arXiv:2305.16582, 2023c.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023.

Xinwu Ye, Chengfan Li, Siming Chen, Xiangru Tang, and Wei Wei. Mmscibench: Benchmarking language models on multimodal scientific problems, 2025. URL https://arxiv.org/abs/2503.01891.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Rouyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms, 2025. URL https://arxiv.org/abs/2504.15280.

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin S. Wang, Mukul Khanna, Théophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander Clegg, John M. Turner, Zsolt Kira, Manolis Savva, Angel X. Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open-vocabulary mobile manipulation. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA, volume 229 of Proceedings of Machine Learning Research, pp. 1975–2011. PMLR, 2023. URL https://proceedings.mlr.press/v229/yenamandra23a.html.

Sriram Yenamandra, Arun Ramachandran, Mukul Khanna, Karmesh Yadav, Jay Vakil, Andrew Melnik, Michael Büttner, Leon Harz, Lyon Brown, Gora Chand Nandi, Arjun P. S, Gaurav Kumar Yadav, Rahul Kala, Robert Haschke, Yang Luo, Jinxin Zhu, Yansen Han, Bingyi Lu, Xuan Gu, Qinyuan Liu, Yaping Zhao, Qiting Ye, Chenxiao Dou, Yansong Chua, Volodymyr Kuzma, Vladyslav Humennyy, Ruslan Partsey, Jonathan Francis, Devendra Singh Chaplot, Gunjan Chhablani, Alexander Clegg, Théophile Gervet, Vidhi Jain, Ram Ramrakhya, Andrew Szot, Austin S. Wang, Tsung-Yen Yang, Aaron Edsinger, Charles C. Kemp, Binit Shah, Zsolt Kira, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Towards open-world mobile manipulation in homes: Lessons from the neurips 2023 homerobot open vocabulary mobile manipulation challenge. CoRR, abs/2407.06939, 2024. doi: 10.48550/ARXIV.2407.06939. URL https://doi.org/10.48550/arXiv.2407.06939.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549, 2023.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=R4Ng8zYaiz.

Peng Yingzhe, Zhang Gongrui, Zhang Miaosen, You Zhiyuan, Liu Jie, Zhu Qipeng, Yang Kai, Xu Xingzhong, Geng Xin, and Yang Xu. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pp. 11289–11303. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.755. URL https://doi.org/10.18653/v1/2023.findings-emnlp.755.

Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui: Grounded mobile UI understanding with multimodal llms. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV, volume 15122 of Lecture Notes in Computer Science, pp. 240–255. Springer, 2024. doi: 10.1007/978-3-031-73039-9_14. URL https://doi.org/10.1007/978-3-031-73039-9_14.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pp. 3208–3216, 2021.

Hong-Tao Yu, Xiu-Shen Wei, Yuxin Peng, and Serge Belongie. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation, 2025. URL https://arxiv.org/abs/2504.14988.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. Trans. Mach. Learn. Res., 2022, 2022. URL https://openreview.net/forum?id=Ee277P3AYC.

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

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. Preprint, 2024b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024c. URL https://openreview.net/forum?id=KOTutrSR2y.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 9127–9134. AAAI Press, 2019a. doi: 10.1609/AAAI.V33I01.33019127. URL https://doi.org/10.1609/aaai.v33i01.33019127.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6281–6290, 2019b.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Liu Yuqi, Peng Bohao, Zhong Zhisheng, Yue Zihao, Lu Fanbin, Yu Bei, and Jia Jiaya. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement, 2025. URL https://arxiv.org/abs/2503.06520.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. arXiv preprint arXiv:2407.08693, 2024.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference, 2018. URL https://arxiv.org/abs/1808.05326.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 6720–6731. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00688. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html.

Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding, 2025. URL https://arxiv.org/abs/2504.14526.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. arXiv preprint arXiv:2402.12226, 2024.

Weichen Zhan, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space, 2025. URL https://arxiv.org/abs/2503.11094.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Vpgtrans: Transfer visual prompt generator across llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/407106f4b56040b2e8dcad75a6e461e5-Abstract-Conference.html.

Bofei Zhang, Zirui Shang, Zhi Gao, Wang Zhang, Rui Xie, Xiaojian Ma, Tao Yuan, Xinxiao Wu, Song-Chun Zhu, and Qing Li. Tongui: Building generalized gui agents by learning from multimodal web tutorials, 2025a. URL https://arxiv.org/abs/2504.12679.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. arXiv preprint arXiv:2401.02582, 2024a.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. arXiv preprint arXiv:2410.02884, 2024b.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000, 2023b.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, pp. 12401–12430. Association for Computational Linguistics, 2024c. doi: 10.18653/V1/2024.FINDINGS-ACL.738. URL https://doi.org/10.18653/v1/2024.findings-acl.738.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023c.

Hanlei Zhang, Zhuohang Li, Yeshuang Zhu, Hua Xu, Peiwu Wang, Haige Zhu, Jie Zhou, and Jinchao Zhang. Can large language models help multimodal language analysis? mmla: A comprehensive benchmark, 2025b. URL https://arxiv.org/abs/2504.16427.

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

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023d. URL http://papers.nips.cc/paper_files/paper/2023/hash/64008fa30cba9b4d1ab1bd3bd3d57d61-Abstract-Datasets_and_Benchmarks.html.

Li Zhang, Shihe Wang, Xianqing Jia, Zhihan Zheng, Yunhe Yan, Longxi Gao, Yuanchun Li, and Mengwei Xu. Llamatouch: A faithful and scalable testbed for mobile UI task automation. In Lining Yao, Mayank Goel, Alexandra Ion, and Pedro Lopes (eds.), Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16, 2024, pp. 132:1–132:13. ACM, 2024d. doi: 10.1145/3654777.3676382. URL https://doi.org/10.1145/3654777.3676382.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal analogical reasoning over knowledge graphs. arXiv preprint arXiv:2210.00312, 2022.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5579–5588, 2021a.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. CoRR, abs/2410.16198, 2024e. doi: 10.48550/ARXIV.2410.16198. URL https://doi.org/10.48550/arXiv.2410.16198.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. Advances in Neural Information Processing Systems, 36:5484–5505, 2023e.

Xi Zhang, Feifei Zhang, and Changsheng Xu. Explicit cross-modal representation learning for visual commonsense reasoning. IEEE Transactions on Multimedia, 24:2986–2997, 2021b.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361, 2023f.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Jingxuan Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users, 2025d. URL https://arxiv.org/abs/2504.10157.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? CoRR, abs/2408.13257, 2024f. doi: 10.48550/ARXIV.2408.13257. URL https://doi.org/10.48550/arXiv.2408.13257.

Zeyu Zhang, Zijian Chen, Zicheng Zhang, Yuze Sun, Yuan Tian, Ziheng Jia, Chunyi Li, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Puzzlebench: A fully dynamic evaluation framework for large multimodal models on puzzle solving, 2025e. URL https://arxiv.org/abs/2504.10885.

Zhuosheng Zhang and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. arXiv preprint arXiv:2309.11436, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023g.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces, 2025a. URL https://arxiv.org/abs/2503.06157.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. arXiv preprint arXiv:2407.05282, 2024a.

Jiahe Zhao, Ruibing Hou, Zejie Tian, Hong Chang, and Shiguang Shan. His-gpt: Towards 3d human-in-scene multimodal understanding, 2025b. URL https://arxiv.org/abs/2503.12955.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning, 2025c.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? arXiv preprint arXiv:2307.16368, 2023.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. Auto-arena: Automating llm evaluations with agent peer battles and committee discussions, 2024b. URL https://arxiv.org/abs/2405.20267.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing, 2025d. URL https://arxiv.org/abs/2504.02826.

Yiming Zhao, Yu Zeng, Yukun Qi, YaoYang Liu, Lin Chen, Zehui Chen, Xikun Bao, Jie Zhao, and Feng Zhao. V2p-bench: Evaluating video-language understanding with visual prompts for better human-model interaction, 2025e. URL https://arxiv.org/abs/2503.17736.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions. CoRR, abs/2411.14405, 2024c. doi: 10.48550/ARXIV.2411.14405. URL https://doi.org/10.48550/arXiv.2411.14405.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024a.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and Yu Su. Skillweaver: Web agents can self-improve by discovering and honing skills, 2025a. URL https://arxiv.org/abs/2504.07079.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 419–428, 2024b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755, 2025b.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 36:5168–5191, 2023.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. arXiv preprint arXiv:2411.12591, 2024c.

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Eric Wang. Vlmbench: A compositional benchmark for vision-and-language manipulation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/04543a88eae2683133c1acbef5a6bf77-Abstract-Datasets_and_Benchmarks.html.

Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, Jie Bai, Zhixiong Ma, and Xichan Zhu. Omnihd-scenes: A next-generation multimodal dataset for autonomous driving, 2025c. URL https://arxiv.org/abs/2412.10734.

Xiangxi Zheng, Linjie Li, Zhengyuan Yang, Ping Yu, Alex Jinpeng Wang, Rui Yan, Yuan Yao, and Lijuan Wang. V-mage: A game evaluation framework for assessing visual-centric capabilities in multimodal large language models, 2025d. URL https://arxiv.org/abs/2504.06148.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. arXiv preprint arXiv:2504.03160, 2025e.

Ling Zhong, Yujing Lu, Jing Yang, Weiming Li, Peng Wei, Yongheng Wang, Manni Duan, and Qing Zhang. Domaincqa: Crafting expert-level qa from domain-specific charts, 2025. URL https://arxiv.org/abs/2503.19498.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13246–13257, 2024a.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pp. 2299–2314. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.FINDINGS-NAACL.149. URL https://doi.org/10.18653/v1/2024.findings-naacl.149.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. arXiv preprint arXiv:1608.05442, 2016.

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

Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan Xie, Xiaojun Chang, Yu Qiao, Wenqi Shao, and Kaipeng Zhang. Gate opening: A comprehensive benchmark for judging open-ended interleaved image-text generation, 2024a.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. arXiv preprint arXiv:2405.13872, 2024b.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024c. URL https://openreview.net/forum?id=oKn9c6ytLx.

Chaoyu Zhu, Zhihao Yang, Xiaoqiong Xia, Nan Li, Fan Zhong, and Lei Liu. Multimodal reasoning based on knowledge graph embedding for specific diseases. Bioinformatics, 38(8):2235–2245, 2022.

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

Wenxuan Zhu, Bing Li, Cheng Zheng, Jinjie Mai, Jun Chen, Letian Jiang, Abdullah Hamdi, Sara Rojas Martinez, Chia-Wen Lin, Mohamed Elhoseiny, and Bernard Ghanem. 4d-bench: Benchmarking multi-modal large language models for 4d object understanding, 2025. URL https://arxiv.org/abs/2503.17827.