A Comprehensive Survey of Evaluating Multimodal Foundation Models: Hierarchical Perspective and Extensive Applications

Anonymous ACL submission

Abstract

Multimodal foundation models have seen rapid progress, demonstrating impressive capabilities across vision, language, audio, and video tasks. However, evaluating these models remains challenging due to the diversity of modalities and the growing complexity of their applications. This survey proposes a hierarchical perspective for evaluating multimodal foundation models, structured across three levels: (1) core foundational abilities, such as unimodal understanding and cross-modal alignment; (2) higher-order intelligence, including multimodal reasoning, temporal understanding, and decision-making; and (3) real-world applications in domains such as healthcare, molecular science, industry, and society. We review representative benchmarks at each level and analyze their strengths and limitations. Our taxonomy aims to guide both model developers and benchmark designers by offering a clear capability hierarchy, a summary of recent progress in evaluating different aspects of multimodal models at different level, and application-driven evaluation settings. We also highlight key challenges in current evaluation practices and outline directions for developing more comprehensive and reliable benchmarks.

1 Introduction

Multimodal foundation models (MFMs) (Fei et al., 2022; Chen et al., 2024d), which integrate various input forms like text, images, audio, and video, are quickly reshaping fields ranging from robotics and scientific research to social sciences and medical diagnostics (Jin et al., 2024d; Liu et al., 2025b; Xiao et al., 2024). Despite their growing capabilities, the *evaluation* of MFMs remains a critical bottleneck (Liang et al., 2024a; Huang and Zhang, 2024). Existing evaluation metrics, methods, and benchmarks are plentiful, they remain fragmented (Fu et al., 2024; Li et al., 2024b). Many focus on narrow tasks under static conditions and fail to assess whether core perceptual abilities translate into reliable behavior in high-stakes, real-world scenarios (Li et al., 2024b; Huang et al., 2024b; Lu et al., 2024). As a result, even extensively benchmarked models leave fundamental questions unanswered: *How well do they perform? What are their failure modes? Are they ready for deployment?*

A Hierarchical Capability Lens. We propose a structured evaluation framework that mirrors the developmental trajectory of MFMs: ❶ ***Fundamental Competencies***: modality-specific understanding and cross-modal alignment; ❷ ***Higher-order Intelligence***: multimodal reasoning, streaming comprehension, decision making, and planning; ❸ ***Real-world Applications***: performance in safety-critical domains, value alignment, and societal impact. Figure 1 shows the structure for evaluating fundamental competencies and higher-order intelligence, while Figure 2 illustrates evaluations in real-world scenarios.

Difference from Previous Surveys. Most previous surveys (Zhang et al., 2024a; Caffagni et al., 2024) adopt flat taxonomies, organizing benchmarks by task (e.g., captioning and VQA) (Liang et al., 2024b) or by modality (e.g., vision-language) (Awais et al., 2025). While useful, such frameworks obscure the role of foundational skills in enabling complex capabilities and overlook how weaknesses in lower-level competencies cascade into real-world failures (Song et al., 2025). In contrast, this survey introduces: ❶ ***Hierarchical Structuring***: organizing evaluations by progressively complex capabilities, from basic perception to domain-grounded deployment; ❷ ***Cross-domain Integration***: treating real-world applications (e.g. medicine, biology, industry, and society) as core evaluation targets rather than end-stage add-ons; ❸ ***Practical Guidance***: providing actionable insights (e.g. design patterns and failure points) for selecting or designing benchmarks tailored to deployment contexts.

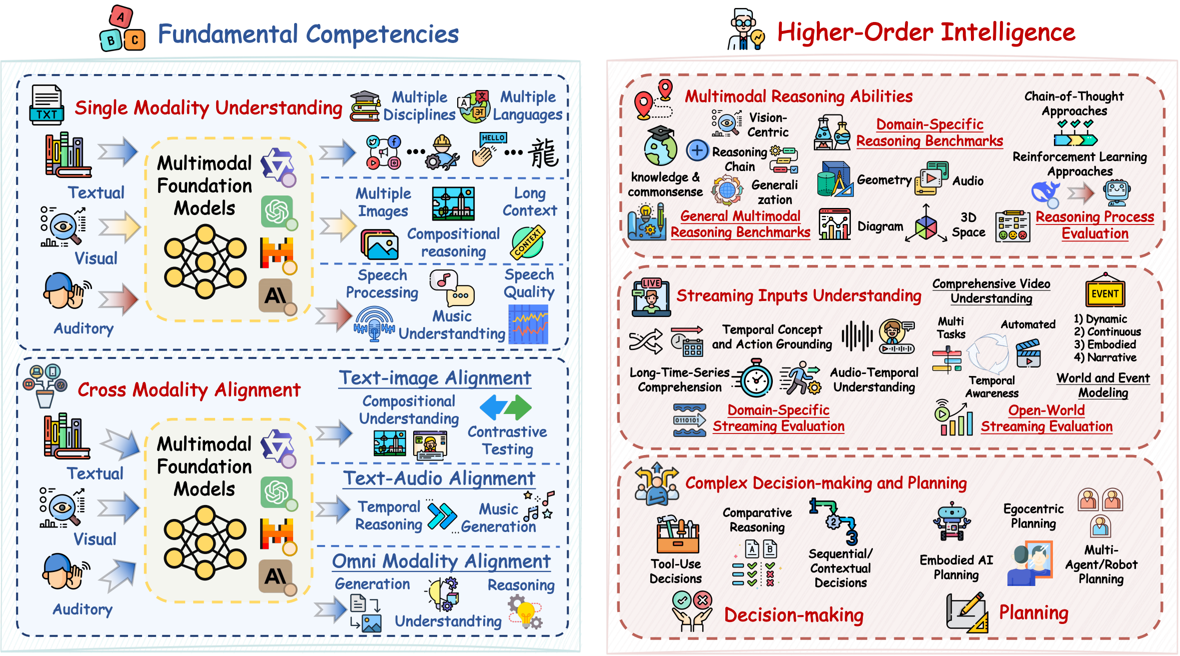


Figure 1: Hierarchical representation of abilities in multimodal foundation models.

Our hierarchical framework offers a principled roadmap for evaluating MFMs–clarifying what to measure, why it matters, and how to design benchmarks that reflect real-world deployment needs. This paper makes three main contributions:

- **Hierarchical Taxonomy**: A three-level perspective encompassing *Fundamental Competencies*, *Higher-Order Intelligence*, and *Real-World Applications* for structuring multimodal evaluations.
- **Synthesis of Evaluations**: A critical review of representative datasets, metrics, and protocols, highlighting common pitfalls, capability gaps, and modality-specific challenges.
- **Forward-Looking Agenda**: Identifying of future directions, including streaming benchmarks, reasoning-process evaluations, value alignment, and domain-specific robustness.

2 Taxonomy

This section categorizes evaluations for multimodal foundation models into three progressive capability aspects, as shown in Figure 3.

- ***Fundamental Competencies.*** *Foundational functionalities essential for multimodal comprehension.* ❶ Single Modality Understanding: Evaluating models' performance within individual modalities, including visual, textual, auditory, and tactile processing; ❷ Cross-Modality Alignment: Measuring the coherence and integration across different modalities, ensuring consistent semantic representation.
- ***Higher-Order Intelligence.*** *Advanced abilities enabling sophisticated multimodal information processing.* ❶ Multimodal Reasoning: Assessing logical inference, causal reasoning, and problem-solving involving multimodal data; ❷ Streaming Input Understanding: Evaluating continuous and sequential multimodal data processing capabilities in real-time scenarios; ❸ Complex Decision-Making and Planning: Determining the ability to synthesize multimodal inputs to formulate plans and decisions.
- ***Real-World Applications.*** *Capabilities that translate multimodal models into practical scenarios and industry-specific domains.* ❶ Value Alignment: Ensuring ethical alignment, cultural sensitivity, and appropriate responses in diverse social contexts; ❷ Medical Applications: Diagnostic assistance, medical image analysis, patient monitoring, and personalized medicine; ❸ Biological Applications: Understanding and interpreting multimodal biological data for genomics, proteomics, and bioinformatics research; ❹ Industry Applications: Application in sectors such as manufacturing, automation, robotics, logistics optimization, and smart systems; ❺ Society: Exploring broader impacts on social functioning and social good.

These aspects collectively enable a comprehensive and structured approach to evaluating multimodal foundation models, moving progressively

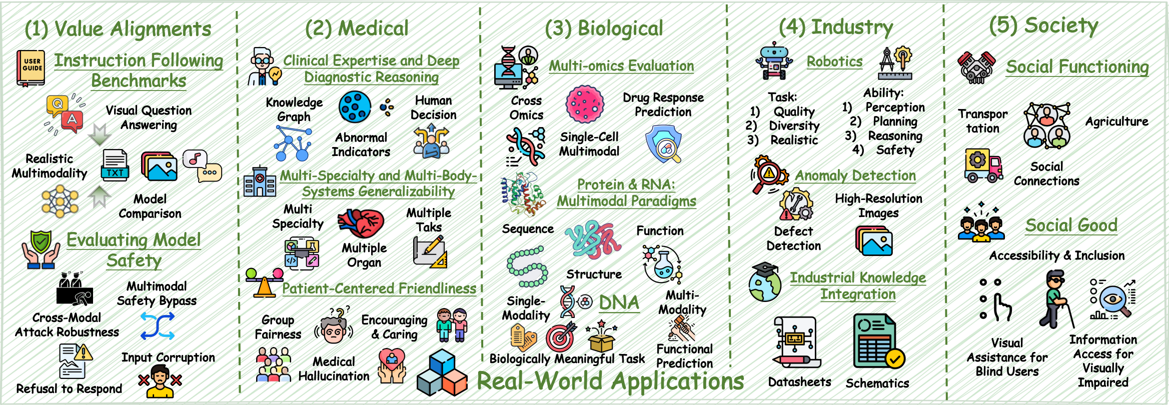


Figure 2: Illustration of strategies applied to real-world applications of multimodal foundation models.

from foundational abilities to advanced intelligence and real-world impacts.

3 Fundamental Competencies

3.1 Single Modality Understanding

Single modality understanding is the core foundational ability of large multimodal models.

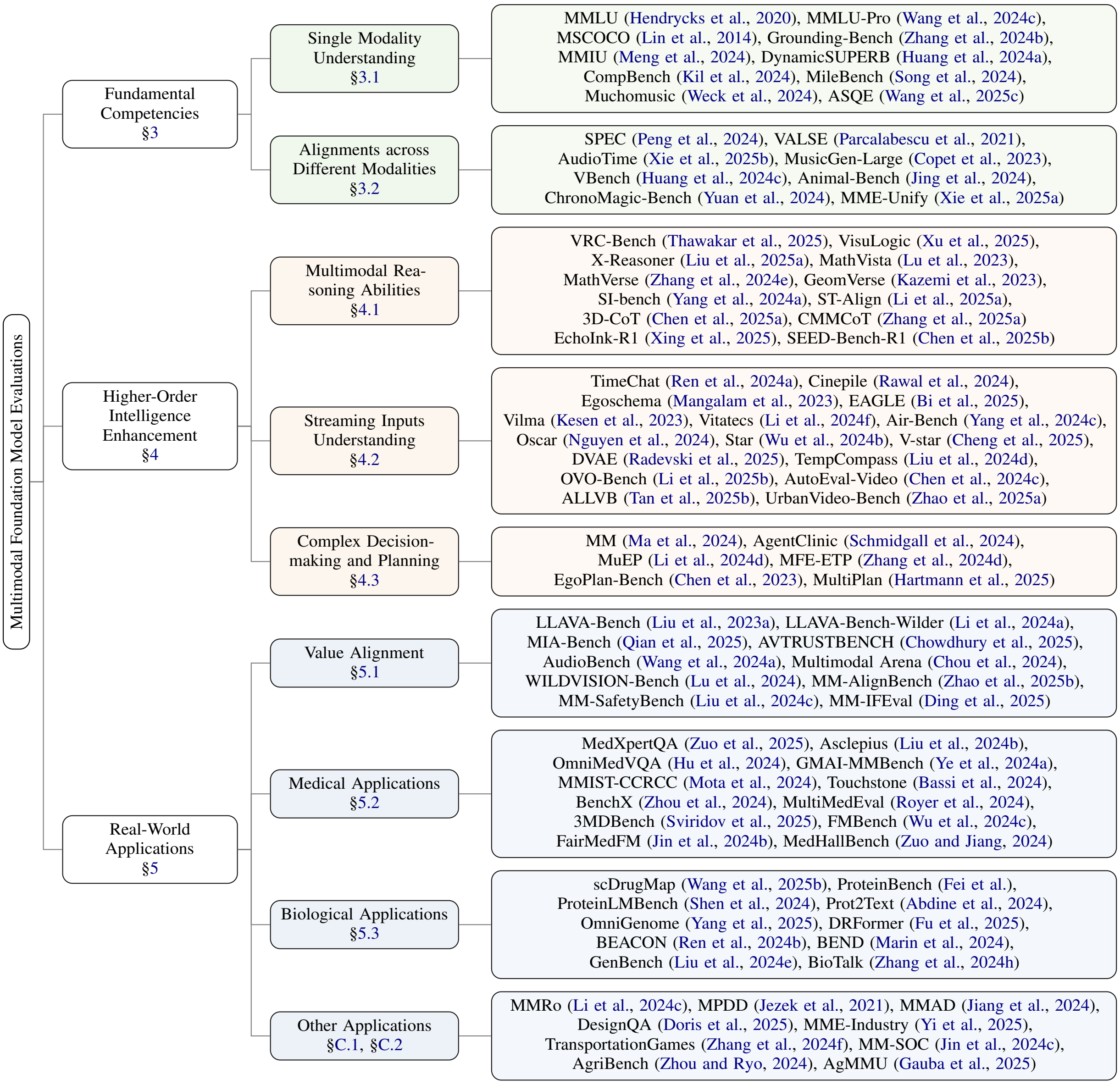
Textual Understanding Evaluation. Earlier systematical textual understanding benchmarks, such as MMLU (Hendrycks et al., 2020), provides a comprehensive evaluation with 57 disciplines. Based on MMLU, MMLU-Pro (Wang et al., 2024c) enhances the complexity and raises the scoring criteria towards more robust and challenging evaluation. Beyond general benchmarks, recent works are focused on multilingual understanding abilities. CMMLU (Li et al., 2023) and KMMLU (Son et al., 2024) provides Chinese and Korean understanding benchmarks. More advanced, Global-MMLU (Singh et al., 2024) extends to 42 languages with culturally sensitive subsets. MMLU-ProX (Xuan et al., 2025) further refines multilingual evaluation with expert-audited translations across 13 typologies.

Visual Understanding Evaluation. Visual understanding evaluation spans both basic perception and advanced comprehension capabilities. Foundational benchmarks like MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) establish baselines for object-level detection and entity localization tasks. Grounding-Bench (Zhang et al., 2024b) is proposed to evaluate spatial visual understanding through phrase localization. Furthermore, more comprehensive visual benchmarks focus on advanced visual understanding. MMIU (Meng et al., 2024) presents a multimodal benchmark requiring multi-images understanding, while MileBench (Song et al., 2024) evaluates long-context visual comprehension, and CompBench (Kil et al., 2024) assess compositional visual reasoning capabilities.

Auditory Understanding Evaluation Audio understanding capability is important for omni large models. DynamicSUPERB (Huang et al., 2024a) is the first collaborative instruction-tuning benchmark for speech processing evaluation, while Muchomusic (Weck et al., 2024) measures music understanding through composer-style recognition and lyric alignment analysis. ASQE (Wang et al., 2025c) introduces auditory LLMs for automatic speech quality evaluation. Comprehensive frameworks like AudioBench (Wang et al., 2024a) unify evaluation across speech, music, and environmental sounds through multi-task assessment protocols.

3.2 Alignments across Different Modalities

The modality alignment benchmark evaluates the cross-modality alignment capabilities of omni large models. Text-image alignment benchmarks employ diverse evaluation strategies, SPEC (Peng et al., 2024) introduces a diagnostic framework for compositional understanding, while VALSE (Parcalabescu et al., 2021) evaluates fine-grained semantic alignment through contrastive testing. For text-audio alignment, AudioTime (Xie et al., 2025b) provides temporally synchronized audio-text pairs for temporal reasoning evaluation, while MusicGen-Large (Copet et al., 2023) benchmarks music generation through text-guided composition tasks. Beyond bio-modality alignments, many recent works focus on omni modality alignments evaluations. VBench (Huang et al., 2024c) provides a comprehensive suits for text to video generation, while Animal-Bench (Jing et al., 2024) and ChronoMagic-Bench (Yuan et al., 2024) specialize


Multimodal Foundation Model Evaluations

Fundamental Competencies §3

Single Modality Understanding §3.1

MMLU (Hendrycks et al., 2020), MMLU-Pro (Wang et al., 2024c), MSCOCO (Lin et al., 2014), Grounding-Bench (Zhang et al., 2024b), MMIU (Meng et al., 2024), DynamicSUPERB (Huang et al., 2024a), CompBench (Kil et al., 2024), MileBench (Song et al., 2024), Muchomusic (Weck et al., 2024), ASQE (Wang et al., 2025c)

Alignments across Different Modalities §3.2

SPEC (Peng et al., 2024), VALSE (Parcalabescu et al., 2021), AudioTime (Xie et al., 2025b), MusicGen-Large (Copet et al., 2023), VBench (Huang et al., 2024c), Animal-Bench (Jing et al., 2024), ChronoMagic-Bench (Yuan et al., 2024), MME-Unify (Xie et al., 2025a)

Higher-Order Intelligence Enhancement §4

Multimodal Reasoning Abilities §4.1

VRC-Bench (Thawakar et al., 2025), VisuLogic (Xu et al., 2025), X-Reasoner (Liu et al., 2025a), MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024e), GeomVerse (Kazemi et al., 2023), SI-bench (Yang et al., 2024a), ST-Align (Li et al., 2025a), 3D-CoT (Chen et al., 2025a), CMMCoT (Zhang et al., 2025a) EchoInk-R1 (Xing et al., 2025), SEED-Bench-R1 (Chen et al., 2025b)

Streaming Inputs Understanding §4.2

TimeChat (Ren et al., 2024a), Cinepile (Rawal et al., 2024), Egoschema (Mangalam et al., 2023), EAGLE (Bi et al., 2025), Vilma (Kesen et al., 2023), Vitatecs (Li et al., 2024f), Air-Bench (Yang et al., 2024c), Oscar (Nguyen et al., 2024), Star (Wu et al., 2024b), V-star (Cheng et al., 2025), DVAE (Radevski et al., 2025), TempCompass (Liu et al., 2024d), OVO-Bench (Li et al., 2025b), AutoEval-Video (Chen et al., 2024c), ALLVB (Tan et al., 2025b), UrbanVideo-Bench (Zhao et al., 2025a)

Complex Decision-making and Planning §4.3

MM (Ma et al., 2024), AgentClinic (Schmidgall et al., 2024), MuEP (Li et al., 2024d), MFE-ETP (Zhang et al., 2024d), EgoPlan-Bench (Chen et al., 2023), MultiPlan (Hartmann et al., 2025)

Real-World Applications §5

Value Alignment §5.1

LLAVA-Bench (Liu et al., 2023a), LLAVA-Bench-Wilder (Li et al., 2024a), MIA-Bench (Qian et al., 2025), AVTRUSTBENCH (Chowdhury et al., 2025), AudioBench (Wang et al., 2024a), Multimodal Arena (Chou et al., 2024), WILDVISION-Bench (Lu et al., 2024), MM-AlignBench (Zhao et al., 2025b), MM-SafetyBench (Liu et al., 2024c), MM-IFEval (Ding et al., 2025)

Medical Applications §5.2

MedXpertQA (Zuo et al., 2025), Asclepius (Liu et al., 2024b), OmniMedVQA (Hu et al., 2024), GMAI-MMBench (Ye et al., 2024a), MMIST-CCRCC (Mota et al., 2024), Touchstone (Bassi et al., 2024), BenchX (Zhou et al., 2024), MultiMedEval (Royer et al., 2024), 3MDBench (Sviridov et al., 2025), FMBench (Wu et al., 2024c), FairMedFM (Jin et al., 2024b), MedHallBench (Zuo and Jiang, 2024)

Biological Applications §5.3

scDrugMap (Wang et al., 2025b), ProteinBench (Fei et al.), ProteinLMBench (Shen et al., 2024), Prot2Text (Abdine et al., 2024), OmniGenome (Yang et al., 2025), DRFormer (Fu et al., 2025), BEACON (Ren et al., 2024b), BEND (Marin et al., 2024), GenBench (Liu et al., 2024e), BioTalk (Zhang et al., 2024h)

Other Applications §C.1, §C.2

MMRo (Li et al., 2024c), MPDD (Jezek et al., 2021), MMAD (Jiang et al., 2024), DesignQA (Doris et al., 2025), MME-Industry (Yi et al., 2025), TransportationGames (Zhang et al., 2024f), MM-SOC (Jin et al., 2024c), AgriBench (Zhou and Ryo, 2024), AgMMU (Gauba et al., 2025)

Figure 3: A taxonomy of recent progress on hierarchical evaluation of multimodal foundation models.

in animal behavior understanding and temporal reasoning generations. Unified evaluation frameworks like MME-Unify (Xie et al., 2025a) simultaneously assess understanding and generation capabilities across multiple tasks.

4 Higher-Order Intelligence

4.1 Multimodal Reasoning Abilities

Multimodal reasoning represents a core capability in the pursuit of advanced AI systems with omnimodal intelligence. Evaluating these capabilities requires specialized benchmarks that test models' ability to perform deep cross-modal information fusion and construct complex logical chains.

General Multimodal Reasoning Benchmarks. Early work like CLEVR (Johnson et al., 2017) established a diagnostic approach to visual reasoning, focusing on compositional language understanding and visual attribute reasoning with minimal bias. Later benchmarks evolved to incorporate world knowledge and commonsense, as in A-OKVQA (Schwenk et al., 2022), which requires reasoning beyond simple visual content recognition. ScienceQA (Lu et al., 2022) expanded this through multimodal science questions, spurring Chain-of-Thought research. And STEM (Shen et al., 2023) presents a more comprehensive evaluation among stem subjects. More recent benchmarks like VRC-Bench (Thawakar et al., 2025) focus on evaluating visual reasoning chains rather than just final answers, while VisuLogic (Xu et al., 2025) provides a comprehensive benchmark for vision-centric reasoning through six distinct reasoning types. X-Reasoner (Liu et al., 2025a) represents an important advancement by exploring reasoning generalization across both modalities and domains.

Domain-Specific Reasoning Benchmarks. These benchmarks evaluate specialized reasoning capabilities across various modalities.

❶ Mathematical and geometric reasoning benchmarks offer focused evaluation environments. MathVista (Lu et al., 2023) assesses mathematical reasoning in visual contexts, while MATH-Vision (Wang et al., 2024b) provides problems from actual math competitions incorporating visual contexts. MathVerse (Zhang et al., 2024e) specifically evaluates diagram understanding in mathematical problems, and GeomVerse (Kazemi et al., 2023) offers procedurally generated geometry problems with controllable difficulty. ❷ Audio-visual reasoning capabilities are assessed through benchmarks like AVQA (Yang et al., 2022). Spatiotemporal and 3D reasoning is evaluated through benchmarks such as VSI-bench (Yang et al., 2024a). ST-Align (Li et al., 2025a) focuses on fine-grained spatiotemporal understanding and localization. 3D-CoT (Chen et al., 2025a) extends reasoning annotations to 3D datasets. For diagram understanding, DPG (Kembhavi et al., 2016) focuses on diagram structure and semantics, while LISA (Yang et al., 2023) introduces *reasoning segmentation*, requiring models to output segmentation masks based on complex textual instructions.

Reasoning Process Evaluation. Beyond outcome-focused evaluation, several benchmarks target the reasoning process itself. Chain-of-Thought (CoT) approaches are evaluated through benchmarks like CMMCoT (Zhang et al., 2025a), which addresses complex multi-image comprehension, VisCoTVisual CoT (Shao et al., 2024), which provides annotations for intermediate reasoning, LLaVA-CoT (Xu et al., 2024), which assesses autonomous multi-stage reasoning, and M3CoT (Chen et al., 2024b), which evaluates multi-domain, multi-step, multi-modal reasoning. Reinforcement learning approaches to reasoning are evaluated through EchoInk-R1 (Xing et al., 2025) for audio-visual reasoning, SEED-Bench-R1 (Chen et al., 2025b) for video understanding, and Reason-RFT (Tan et al., 2025a) for general visual reasoning.

Discussion. Despite progress in multimodal reasoning evaluation, current benchmarks face several limitations. Most benchmarks evaluate explicit reasoning, whereas real-world scenarios often require implicit or counterintuitive reasoning.

4.2 Streaming Inputs Understanding

Understanding streaming data is vital for multimodal systems to model temporal dependencies and evolving contexts in the real world. Streaming data poses unique challenges like long-range temporal modeling, distinct from static inputs.

Domain-Specific Streaming Evaluation. Domain-specific benchmarks use narrowly scoped tasks to assess particular temporal modeling capabilities. ❶ *Long-Time-Series Comprehension.* Several benchmarks aim to assess how well models retain and reason over information in long temporal data. TimeChat (Ren et al., 2024a) probes key frame understanding for MLLM grounding in long videos. Cinepile (Rawal et al., 2024) uses open-ended movie QA to measure narrative coherence and long-range dependencies. Mmbench-video (Liu et al., 2023b) is a multi-shot benchmark for holistic video understanding of scene transitions and temporal shifts. Egoschema (Mangalam et al., 2023) (closed-ended) and EAGLE (Bi et al., 2025) (open-ended) provide diagnostic datasets for egocentric procedural activity understanding. ❷ *Temporal Concept and Action Grounding.* Another line of work focuses on the ability of models to ground abstract temporal concepts and human actions in visual data. Vilma (Kesen et al., 2023) offers zero-shot linguistic and temporal grounding by aligning text to video segments. Vitatecs (Li et al., 2024f) diagnoses temporal concept understanding through event order and duration recognition. Oscar (Nguyen et al., 2024) probes frame-level causal temporal reasoning via object state captioning and change representation. Star (Wu et al., 2024b) uses multi-choice human action tasks to assess situated reasoning in constrained setting. V-star (Cheng et al., 2025) simultaneously investigates the model's temporal and spatial reasoning capabilities by action grounding. ❸ *Audio-Temporal Understanding.* Other benchmarks investigate temporal reasoning in the audio modality. MuchoMusic (Weck et al., 2024) evaluates music understanding, focusing on temporal structure and themes. Air-Bench (Yang et al., 2024c) assesses generative audio comprehension via audio-text alignment and temporal inference. Audiotime (Xie et al., 2025b) provides a temporally aligned audio-text dataset for fine-grained timestamped understanding. DVAE (Radevski et al., 2025) decouples the evaluation of audio and visual capabilities to reduce the bias introduced by visual information.

Open-World Streaming Evaluation. Open-world benchmarks blend multiple tasks and modalities to simulate real-life streaming comprehension. ❶ *Multi-Task Video Understanding.* Recent benchmarks capture open-domain video comprehension through diverse temporal tasks and evaluation paradigms. MvBench (Li et al., 2024e) establishes multi-task closed-ended evaluation across 20 real-life domains, while TempCompass (Liu et al., 2024d) focuses on temporal reasoning over open-domain content. OVO-Bench (Li et al., 2025b) pioneers timestamp-aware evaluation for online video LLMs, assessing dynamic temporal awareness through backward tracing, real-time understanding, and forward active responding scenarios. Video-Bench (Ning et al., 2023) offers open-ended evaluation on narration/commonsense tasks, complemented by AutoEval-Video's (Chen et al., 2024c) automated QA assessment. ALLVB (Tan et al., 2025b) presents a large-scale, all-in-one long video understanding benchmark, converting 9 real-world tasks into a QA format and featuring an automated annotation pipeline. ❷ *World and Event Modeling.* Other efforts explore continuous modeling of dynamic environments and complex event sequences. MMWorld (He et al., 2024) evaluates world models on complex real-world video dynamics and cross-domain reasoning. WorldSense (Hong et al., 2025) simulates omni-modal scenes for open-ended understanding in continuous, interactive scenarios. UrbanVideo-Bench (Zhao et al., 2025a) assesses embodied intelligence in urban 3D spaces. EventBench (Du et al., 2024) targets event-oriented long-video understanding, requiring reasoning over intricate event narratives.

Discussion. Despite recent progress, streaming benchmarks still exhibit significant biases. As Park et al. (Park et al., 2025) point out, many streaming tasks can be solved using unimodal inputs, which leads to an overestimation of the true multimodal capabilities of current models. Future benchmarks should enforce genuine cross-modal integration, require implicit temporal inference and robustness to long-horizon and real-time dependencies.

4.3 Complex Decision-making and Planning

The ability to make accurate decisions and formulate executable plans in complex environments marks a critical frontier for multimodal systems. Recent benchmarks investigate such ability by simulating scenarios requiring tool usage and cross-modal coordination under dynamic constraints.

Benchmarking decision-making capabilities often focuses on scenarios requiring inference and choice under uncertainty. MM (Ma et al., 2024) serves as a benchmark for evaluating tool-use decision in multi-step multimodal tasks. MLLM-CompBench (Li et al., 2024d) evaluates the comparative reasoning of multimodal language models by presenting paired images and asking questions that require discerning relative characteristics.

For planning, benchmarks assess models' ability to generate coherent, executable sequences of actions in response to complex goals. In the domain of embodied AI, where agents must navigate and interact with physical environments, MuEP (Li et al., 2024d) is a comprehensive multimodal benchmark for embodied planning, evaluating multimodal and multi-turn interactions in complex scenes. Similarly, MFE-ETP (Zhang et al., 2024d) provides a comprehensive evaluation framework on embodied task planning, focusing on object understanding, spatiotemporal perception, task understanding, and embodied reasoning. For human-level planning from an egocentric perspective, Chen et al. (2023) and Qiu et al. (2024) have evaluated multimodal large language models in real-world tasks with action plans and intricate visual observations. Lastly, for multi-robot systems, MultiPlan (Hartmann et al., 2025) addresses optimal multimodal multi-robot multi-goal path planning.

5 Real-World Applications

5.1 Value Alignments

Aligning models with human values is a critical step toward production-ready, useful chatbot assistants that provide honest and safe responses (Ouyang et al., 2022). Value alignment encompasses faithfully following user instructions, minimizing hallucinations, and adhering to ethical principles. Due to its broad scope and importance, robust evaluation requires comprehensive yet targeted benchmarks (Askell et al., 2021).

Instruction Following Benchmarks. Early efforts to systematically evaluate instruction-following in the multimodal setting include LLAVA-Bench (Liu et al., 2023a) and its successor LLAVA-Bench-Wilder (Li et al., 2024a), which assess open-ended vision-language questions about image interpretation. Expanding this approach, MIA-Bench (Qian et al., 2025) and MM-IFEval (Ding et al., 2025) increase the diversity of questions and combine GPT-assisted evaluation with predefined rubrics for a more fine-grained and robust score. Concurrently, AudioBench (Wang et al., 2024a) adapts text-only datasets such as ALPACA (Taori et al., 2023) into the audio modality while retaining the same instruction-following paradigm. Meanwhile, another line of work targets model comparison. Chatbot Arena (Chiang et al., 2024) has been extended to multimodal inputs via platforms like WILDVISION-ARENA (Lu et al., 2024) and Multimodal Arena (Chou et al., 2024), which crowdsource vision-language model comparisons. The MM-AlignBench (Zhao et al., 2025b) is a similar endeavor but with additional filtering to improve the question and image diversity. Finally, Eval-Anything (Ji et al., 2024) offers a comprehensive benchmark suite for omni-models, evaluating both multimodal understanding and generation.

Evaluating Model Safety. In addition to general instruction-following abilities, other benchmarks target more specific scenarios. The inclusion of new modalities opens more ways to compromise safety. MM-SafetyBench (Liu et al., 2024c) and MM-RLHF-SafetyBench (Zhang et al., 2025b) demonstrate that supplying a content-relevant image can bypass safety filters. B-AVIBench (Zhang et al., 2024c) represents attacks that corrupt text or image input by paraphrasing queries or introducing noise and occlusions to images. AVTrust-Bench (Chowdhury et al., 2025) assesses whether omni-models appropriately refuse to respond under adversarial or incomplete conditions.

Discussion. Despite substantial progress, key limitations remain. Most benchmarks focus on modality-specific instructions, with few offering truly omni-modal evaluation. The exponential growth in input-output modality combinations complicates comprehensive evaluation and raises concerns about the transferability of language-specific abilities across modalities. This complexity further challenges the design of robust defenses against attacks that may exploit any modality pairing.

5.2 Medical Applications

Omni-modal medical data (Zuo et al., 2025; Jin et al., 2024a; Xia et al., 2024) includes textual descriptions, radiology reports, histopathology images, and statistical plots, etc. Multi-modal, LLM-driven medical platforms now play a crucial role in public health (Rao et al., 2025; Liu et al., 2024b; Zhu et al., 2024), acting as clinical assistants that helps to provide a comprehensive diagnostic knowledge base, personalized risk alerts, and tailored treatment suggestions.

Evaluating Clinical Expertise and Deep Diagnostic Reasoning. The real-world medical analysis is always complicated and challenging. Even experienced clinicians must engage in thorough analysis to reach reliable diagnostic conclusions. Simple medical QA or multiple-choice tasks cannot fully assess the suitability of these models as a clinical assistant. In response, Panagoulias et al. (Panagoulias et al., 2024) build medical knowledge graphs and design related questions to test expert-level reasoning, requiring answers that follow specific knowledge paths. RJUA-MedDQA (Jin et al., 2024a) evaluates the numerical reasoning ability by identifying abnormal indicators and clinical reasoning ability through case-based medical contexts. Robert et al. (Kaczmarczyk et al., 2024) compares the medical MLLMs with collective human decision-making outcomes. Together, these benchmarks consistently show that current medical MLLMs have not yet achieved expert-level.

Evaluating Multi-Specialty and Multi-Body-Systems Generalizability. Medical domains span a variety of specialties, body systems, and organs. Many works evaluate the generalizability of medical MLLMs across these tasks. Specifically, MedXpertQA (Zuo et al., 2025) covers 17 specialties and 11 body system–related tasks. Asclepius (Liu et al., 2024b) contains 15 medical specialties.OmniMedVQA (Hu et al., 2024) is a multi-organ benchmark that contains more than 20 anatomical regions. Several works (Ye et al., 2024a; Yan et al., 2024a; Mota et al., 2024; Bassi et al., 2024) have conducted experiments on mixed specialty and organ datasets. In the aspect of evaluation toolkits, BenchX (Zhou et al., 2024) proposes a unified finetuning protocol that enables medical MLLMs to adapt consistently across tasks. MultiMedEval (Royer et al., 2024) is a Python toolkit spanning over 11 medical domains.

Evaluating Patient-Centered Friendliness. As medical omni-modal models are integrated into patient care, many studies evaluate their patient-centered friendliness. To evaluate whether MLLMs are able to give real-world patients more encouraging and caring suggestions, 3MDBench (Sviridov et al., 2025) proposes simulated temperament-driven Agents for benchmarking. Because of the biases in medical training data, medical models often exhibit diagnostic bias when issuing instructions. FMBench (Wu et al., 2024c) and FairMedFM (Jin et al., 2024b) benchmark the fairness of medical

MLLMs across diverse patient groups, evaluating performance by race, ethnicity, language, and gender independently. MediConfusion (Sepehri et al., 2024) benchmarks the failure modes of medical MLLMs. Their results show that current models perform worse than random guessing, raising serious concerns about the reliability. Recent works (Yan et al., 2024b; Zuo and Jiang, 2024) evaluate the trustworthiness and hallucinations, i.e., inaccurate diagnosis or factual medical errors.

Discussion. Although omni-modal models have gained great progress in various medical domains and specialties, their reasoning ability has not reached expert-level. In addition, their inherent hallucinations or group bias may raise unreliable and inaccurate suggestions to patients.

5.3 Biological Applications

Recent advances in foundation models (Yang et al., 2024d) highlight the need for benchmarks reflecting biological data. Life science applications often involve reasoning over heterogeneous data (Boekel et al., 2015; Tang et al., 2023), such as DNA, RNA, proteins, phenotypes, and textual annotations.

Benchmarks Integrating Multiple Molecular Modalities and Omics. A significant trend is integrating multiple molecular or omics data types (Wang et al., 2025a). Recent benchmarks combine diverse biological data to foster more capable foundation models: ❶ *Multi-omics Evaluation.* COMET (Ren et al., 2024c) provides a comprehensive benchmarking framework for multi-omics tasks. scMultiBench (Liu et al., 2024a) systematically evaluates integration approaches for single-cell data. scDrugMap (Wang et al., 2025b) benchmarks large foundation models for drug response prediction using single-cell multimodal data. ❷ *Protein and RNA: Multimodal Benchmarking Paradigms.* Protein and RNA modeling have become fertile grounds for multimodal benchmark development (Shen et al., 2024). ProteinBench (Fei et al.) offers an evaluation framework that spans sequence, structure, and function, employing a multi-metric approach. Prot2Text (Abdine et al., 2024) represents a typical multimodal approach, integrating protein sequence, structure, and text annotation for function generation in free-form language. In the RNA field, OmniGenome (Yang et al., 2025) and DRFormer (Fu et al., 2025) both leverage RNA sequence and structural modalities for downstream tasks. BEACON (Ren et al., 2024b) provides a comprehensive suite of RNA tasks, requiring combined modeling of sequence and structure. ❸ *DNA: Predominantly Single-Modality with Multimodal Advances in Functional Prediction.* In contrast to proteins and RNAs, most benchmarks for DNA foundation models remain largely single-modality, focusing primarily on DNA sequence data. As a result, benchmarks such as BEND (Marin et al., 2024), and GenBench (Liu et al., 2024e) emphasize biologically meaningful tasks based on DNA sequence, such as classification, regulatory element discovery, and annotation. Nevertheless, there are emerging efforts to introduce multimodal benchmarks in the DNA field such as BioTalk (Zhang et al., 2024h), which presents a novel dataset that couples DNA sequences with free-text descriptions of enzymatic function.

Discussion. Multimodal and multi-omics benchmarks are reshaping foundation model evaluation in life sciences by integrating genomic, transcriptomic, proteomic, phenotypic, and textual data, enabling holistic assessments of cross-modal reasoning. However, challenges include limited standardization of data formats, insufficient datasets, and inconsistent benchmark task design and metrics.

6 Challenges and Future Directions

While multimodal foundation models have achieved promising performance in understanding multimodal data, current evaluation efforts of these foundation models still face several challenges.

▷ **Adversarial Perturbations.** Despite their success, deep learning models are prone to adversarial perturbations (Szegedy et al., 2013; Dong et al., 2018). Existing literature suggests that multimodal large language models can also be affected by adversarial inputs (Shayegani et al., 2023; Jha and Reddy, 2023), posing safety concerns on multimodal foundation models, whereas current evaluations fall short in this aspect.

▷ **Multimodal Distribution Shift.** The multimodal foundation models deployed in real-world applications are often challenged by data under multimodal distribution shifts (Yang et al., 2024b; Zhao et al., 2025c), damaging their performance (Zhao et al., 2025d), while a comprehensive evaluation of the models' robustness against multimodal distribution shifts is scarce.

▷ **Biases and Spurious Correlations.** While multimodal foundation models have been observed to be biased with spurious correlations (Adewumi et al., 2024; Zhang et al., 2024g; Ye et al., 2024b), many existing evaluation benchmarks do not take this into account. Therefore, systematic evaluations of these in multimodal foundation models are one of the important future directions.

Limitations

While this paper provides a structured taxonomy for evaluating multimodal foundation models, several limitations and avenues for future exploration remain. First, due to the rapid advancement in multimodal model research, some emerging techniques and novel evaluation methods may not be fully represented in our framework, requiring continuous updates to ensure relevance. Second, while our focus on fundamental competencies and higher-order intelligence is critical, additional challenges related to model robustness, interpretability, and real-world deployment still require further attention. Moreover, the interplay between various evaluation aspects, such as cross-modality alignment and complex decision-making, has yet to be fully understood. We hope these limitations will inspire future studies to refine the evaluation methodologies and address the remaining gaps in this dynamic field.

References

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765.

Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. 2024. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Pedro R. A. S. Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Saikat Roy, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, and 34 others. 2024. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? In *Advances in Neural Information Processing Systems*, volume 37, pages 15184–15201. Curran Associates, Inc.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600.

Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. 2025. Eagle: Egocentric aggregated language-video engine. *arXiv preprint arXiv:2409.17523*.

Jorrit Boekel, John M Chilton, Ira R Cooke, Peter L Horvatovich, Pratik D Jagtap, Lukas Käll, Janne Lehtiö, Pieter Lukasse, Perry D Moerland, and Timothy J Griffin. 2015. Multi-omic data analysis using galaxy. *Nature biotechnology*, 33(2):137–139.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *ACL*, pages 13590–13618.

Feng Chen, Botian Xu, Pu Hua, Peiqi Duan, Yanchao Yang, Yi Ma, and Huazhe Xu. 2024a. On the evaluation of generative robotic simulations. *arXiv preprint arXiv:2410.08172*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv:2405.16473*.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2024c. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer.

Yanjun Chen, Yirong Sun, Xinghao Chen, Jian Wang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. 2025a. Integrating chain-of-thought for multimodal alignment: A study on 3d vision-language learning. *arXiv:2503.06232*.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2023. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. 2025b. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv:2503.24376*.

Zheyi Chen, Liuchang Xu, Hongting Zheng, Luyao Chen, Amr Tolba, Liang Zhao, Keping Yu, and Hailin Feng. 2024d. Evolution and prospects of foundation models: From large language models to large multimodal models. *Computers, Materials & Continua*, 80(2).

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. 2025. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Christopher Chou, Lisa Dunlap, Wei-Lin Chiang, Ying Sheng, Lianmin Zheng, Anastasios Angelopoulos, Trevor Darrell, Ion Stoica, and Joseph E. Gonzalez. 2024. The multimodal arena is here! `https://lmsys.org/blog/2024-06-27-multimodal/`. Accessed: 2025-05-18.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. 2025. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. *CoRR*, abs/2501.02135.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Mmifengine: Towards multimodal instruction following. *Preprint*, arXiv:2504.07957.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193.

Anna C Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. 2025. Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *Journal of Computing and Information Science in Engineering*, 25(2):021009.

Yifan Du, Kun Zhou, Yuqi Huo, Yifan Li, Wayne Xin Zhao, Haoyu Lu, Zijia Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. 2024. Towards event-oriented long video understanding. *arXiv preprint arXiv:2406.14129*.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, and 1 others. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094.

YE Fei, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. Proteinbench: A holistic evaluation of protein foundation models. In *The Thirteenth International Conference on Learning Representations*.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*.

Jianqi Fu, Haohao Li, Yanlei Kang, Hancan Zhu, Tiren Huang, and Zhong Li. 2025. Drformer: A benchmark model for rna sequence downstream tasks. *Genes*, 16(3):284.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S Adve, and Yu-Xiong Wang. 2025. Agmmu: A comprehensive agricultural multimodal understanding and reasoning benchmark. *arXiv:2504.10568*.

Markus Michael Geipel. 2024. Towards a benchmark of multimodal large language models for industrial engineering. In *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–4. IEEE.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Valentin N Hartmann, Tirza Heinle, and Stelian Coros. 2025. A benchmark for optimal multi-modal multi-robot multi-goal path planning with given robot assignment. *arXiv preprint arXiv:2503.03509*.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, and 1 others. 2024. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv:2009.03300*.

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

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024a. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv:2408.15769*.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, and 1 others. 2024b. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv:2407.00468*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024c. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE.

Akshita Jha and Chandan K Reddy. 2023. Codeattack: Code-based adversarial attacks for pre-trained programming language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14892–14900.

Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, Mohan Wang, Josef Dai, Tianyi Qiu, Hua Xu, Dong Li, Weipeng Chen, Jun Song, Bo Zheng, and Yaodong Yang. 2024. Align anything: Training all-modality models to follow instructions with language feedback. *CoRR*, abs/2412.15838.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. 2024. Mmad: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection. *arXiv:2410.09453*.

Congyun Jin, Ming Zhang, Weixiao Ma, Yujiao Li, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, and 1 others. 2024a. Rjuameddqa: A multimodal benchmark for medical document question answering and clinical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5218–5229.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S. Kevin Zhou, and Xiaoxiao Li. 2024b. Fairmedfm: Fairness benchmarking for medical imaging foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pages 111318–111357. Curran Associates, Inc.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024c. Mm-soc: Benchmarking multimodal large language models in social media platforms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6192–6210.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, and 1 others. 2024d. Efficient multimodal large language models: A survey. *arXiv:2405.10739*.

Yinuo Jing, Ruxu Zhang, Kongming Liang, Yongxiang Li, Zhongjiang He, Zhanyu Ma, and Jun Guo. 2024. Animal-bench: Benchmarking multimodal video models for animal-centric video understanding. *Advances in Neural Information Processing Systems*, 37:78766–78796.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Robert Kaczmarczyk, Theresa Isabelle Wilhelm, Ron Martin, and Jonas Roos. 2024. Evaluating multimodal ai in medical diagnostics. *npj Digital Medicine*.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv:2312.12241*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and 1 others. 2023. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *arXiv preprint arXiv:2311.07022*.

Jihyung Kil, Zheda Mai, Justin Lee, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, Arpita Chowdhury, and Wei-Lun Chao. 2024. Compbench: A comparative reasoning benchmark for multimodal llms. *arXiv:2407.16837*.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv:2306.09212*.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. 2025a. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. *arXiv:2501.08282*.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, and 1 others. 2024b. A survey on benchmarks of multimodal large language models. *arXiv:2408.08632*.

Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. 2024c. Mmro: Are multimodal llms eligible as the brain for in-home robotics? *arXiv:2406.19693*.

Kanxue Li, Baosheng Yu, Qi Zheng, Yibing Zhan, Yuhui Zhang, Tianle Zhang, Yijun Yang, Yue Chen, Lei Sun, Qiong Cao, and 1 others. 2024d. Muep: A multimodal benchmark for embodied planning with foundation models [c]. In *Intemational Joint Conferences on Artificial Intelligence. IJCAI*, pages 129–138.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024e. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. 2024f. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer.

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, and 1 others. 2025b. Ovo-bench: How far is your video-llms from realworld online video understanding? *arXiv preprint arXiv:2501.05510*.

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2024a. Hemm: Holistic evaluation of multimodal foundation models. *arXiv preprint arXiv:2407.03418*.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024b. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Chunlei Liu, Sichang Ding, Hani Jieun Kim, Siqu Long, Di Xiao, Shila Ghazanfar, and Pengyi Yang. 2024a. Multi-task benchmarking of single-cell multimodal omics integration methods. *bioRxiv*, pages 2024–09.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Jie Liu, Wenxuan Wang, Yihang Su, Jingyuan Huan, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, and Michael R. Lyu. 2024b. A spectrum evaluation benchmark for medical multi-modal large language models. *Preprint*, arXiv:2402.11217.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, and 1 others. 2025a. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv:2505.03981*.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025b. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv:2501.01282*.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024c. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, volume 15114 of *Lecture Notes in Computer Science*, pages 386–403. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024d. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Zicheng Liu, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z Li. 2024e. Genbench: A benchmarking suite for systematic evaluation of genomic foundation models. *arXiv preprint arXiv:2406.01627*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*.

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

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zixian Ma, Weikai Huang, Jieyu Zhang, Tanmay Gupta, and Ranjay Krishna. 2024. m & m's: A benchmark to evaluate tool-use for m ulti-step m ulti-modal tasks. In *European Conference on Computer Vision*, pages 18–34. Springer.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Frederikke I Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. 2024. Bend: Benchmarking dna language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, and 1 others. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv:2408.02718*.

Tiago Mota, M. Rita Verdelho, Diogo J. Araújo, Alceu Bissoto, Carlos Santiago, and Catarina Barata. 2024. Mmist-ccrcc: A real world medical dataset for the development of multi-modal systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2395–2403.

Nguyen Nguyen, Jing Bi, Ali Vosoughi, Yapeng Tian, Pooyan Fazli, and Chenliang Xu. 2024. Oscar: Object state captioning and state change representation. *arXiv preprint arXiv:2402.17128*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Dimitrios P. Panagoulias, Maria Virvou, and George A. Tsihrintzis. 2024. Evaluating llm – generated multimodal diagnosis from medical images and symptom analysis. *Preprint*, arXiv:2402.01730.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv:2112.07566*.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. 2025. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19821–19829.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2024. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13279–13288.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2025. Mia-bench: Towards better instruction following evaluation of multimodal llms. In *ICLR*. OpenReview.net.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. 2024. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*.

Gorjan Radevski, Teodora Popordanoska, Matthew B Blaschko, and Tinne Tuytelaars. 2025. Dave: Diagnostic benchmark for audio visual evaluation. *arXiv preprint arXiv:2503.09321*.

Vishwanatha M. Rao, Michael Hla, Michael Moor, Subathra Adithan, Stephen Kwak, Eric J. Topol, and Pranav Rajpurkar. 2025. Multimodal generative ai for medical image interpretation. *Nature*.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024a. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, and 1 others. 2024b. Beacon: Benchmark for comprehensive rna tasks and language models. *Advances in Neural Information Processing Systems*, 37:92891–92921.

Yuchen Ren, Wenwei Han, Qianyuan Zhang, Yining Tang, Weiqiang Bai, Yuchen Cai, Lifeng Qiao, Hao Jiang, Dong Yuan, Tao Chen, and 1 others. 2024c. Comet: Benchmark for comprehensive biological multi-omics evaluation tasks and language models. *arXiv preprint arXiv:2412.10347*.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *Preprint*, arXiv:2402.09262.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. 2024. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. *Preprint*, arXiv:2409.15477.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*.

Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. 2023. Measuring vision-language stem skills of neural models. In *ICLR*.

Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. 2024. A fine-tuning dataset and benchmark for large language models for protein understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2390–2395. IEEE.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv:2412.03304*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv:2402.11548*.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv:2404.18532*.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 2025. 3mdbench: Medical multimodal multi-agent dialogue benchmark. *Preprint*, arXiv:2504.13861.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025a. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv:2503.20752*.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. 2025b. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7211–7219.

Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, and 1 others. 2023. Explainable multi-task learning for multi-modality biological data analysis. *Nature communications*, 14(1):2546.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, and 1 others. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv:2501.06186*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024b. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Pengfei Wang, Wenhao Liu, Jiajia Wang, Yana Liu, Pengjiang Li, Ping Xu, Wentao Cui, Ran Zhang, Qingqing Long, Zhilong Hu, and 1 others. 2025a. sccompass: An integrated multi-species scrna-seq database for ai-ready. *Advanced Science*, page 2500870.

Qing Wang, Yining Pan, Minghao Zhou, Zijia Tang, Yanfei Wang, Guangyu Wang, and Qianqian Song. 2025b. scdrugmap: Benchmarking large foundation models for drug response prediction. *arXiv preprint arXiv:2505.05612*.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, and 1 others. 2025c. Enabling auditory large language models for automatic speech quality evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv:2408.01337*.

Alan Wu, Ye Yuan, and Ming Zhang. 2024a. Visionbraille: An end-to-end tool for chinese braille image-to-text translation. *arXiv:2407.06048*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024b. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Peiran Wu, Che Liu, Canyu Chen, Jun Li, Cosmin I. Bercea, and Rossella Arcucci. 2024c. Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. *Preprint*, arXiv:2410.01089.

Peter R Wurman, Raffaello D'Andrea, and Mick Mountz. 2008. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI magazine*, 29(1):9–9.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, and 1 others. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. 2024. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*.

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. 2025a. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv:2504.03641*.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. 2025b. Audiotime: A temporally-aligned audio-text benchmark dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. 2025. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv:2505.04623*.

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

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv:2504.15279*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, and 1 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv:2503.10497*.

Lawrence K. Q. Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, and Junyu Liu. 2024a. Large language model benchmarks in medical tasks. *Preprint*, arXiv:2410.21348.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024b. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *Preprint*, arXiv:2405.20421.

Heng Yang, Renzhi Chen, and Ke Li. 2025. Bridging sequence-structure alignment in rna foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21929–21937.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024a. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv:2412.14171*.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. 2024b. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024c. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. 2023. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv:2312.17240*.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, and 1 others. 2024d. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34(12):830–845.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, and 1 others. 2024a. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. 2024b. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*.

Dongyi Yi, Guibo Zhu, Chenglin Ding, Zongshu Li, Dong Yi, and Jinqiao Wang. 2025. Mme-industry: A cross-industry multimodal evaluation benchmark. *arXiv:2501.16688*.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. 2024. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430.

Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, and 1 others. 2025a. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. *arXiv:2503.05255*.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and 1 others. 2024b. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, Nanning Zheng, and Kaipeng Zhang. 2024c. B-avibench: Towards evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. *TIFS*.

Min Zhang, Xian Fu, Jianye Hao, Peilong Han, Hao Zhang, Lei Shi, Hongyao Tang, and Yan Zheng. 2024d. Mfe-etp: A comprehensive evaluation benchmark for multi-modal foundation models on embodied task planning. *arXiv preprint arXiv:2407.05047*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024e. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Xue Zhang, Xiangyu Shi, Xinyue Lou, Rui Qi, Yufeng Chen, Jinan Xu, and Wenjuan Han. 2024f. Transportationgames: Benchmarking transportation knowledge of (multimodal) large language models. *arXiv:2401.04471*.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024g. Debiasing multimodal large language models. *arXiv preprint arXiv:2403.05262*.

Yifan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025b. MM-RLHF: the next step forward in multimodal LLM alignment. *CoRR*, abs/2502.10391.

Yuchen Zhang, Ratish Kumar Chandrakant Jha, Soumya Bharadwaj, Vatsal Sanjaykumar Thakkar, Adrienne Hoarfrost, and Jin Sun. 2024h. A benchmark dataset for multimodal prediction of enzymatic function coupling dna sequences and natural language. *arXiv preprint arXiv:2407.15888*.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and 1 others. 2025a. Urbanvideo-bench: benchmarking vision-language models on embodied intelligence with video data in urban spaces. *arXiv preprint arXiv:2503.06157*.

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, Haodong Duan, Hua Yang, and Kai Chen. 2025b. Omnialign-v: Towards enhanced alignment of mllms with human preference. *CoRR*, abs/2502.18411.

Yusheng Zhao, Junyu Luo, Xiao Luo, Jinsheng Huang, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. 2025c. Attention bootstrapping for multi-modal test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22849–22857.

Yusheng Zhao, Junyu Luo, Xiao Luo, Weizhi Zhang, Zhiping Xiao, Wei Ju, Philip S Yu, and Ming Zhang. 2025d. Multifaceted evaluation of audio-visual capability for mllms: Effectiveness, efficiency, generalizability and robustness. *arXiv preprint arXiv:2504.16936*.

Yang Zhou, Tan Li Hui Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. 2024. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. In *Advances in Neural Information Processing Systems*, volume 37, pages 6625–6647. Curran Associates, Inc.

Yutong Zhou and Masahiro Ryo. 2024. Agribench: A hierarchical agriculture benchmark for multimodal large language models. *arXiv:2412.00465*.

Zhihong Zhu, Xuxin Cheng, Yunyan Zhang, Zhaorun Chen, Qingqing Long, Hongxiang Li, Zhiqi Huang, Xian Wu, and Yefeng Zheng. 2024. Multivariate cooperative game for image-report pairs: Hierarchical semantic alignment for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 303–313. Springer Nature Switzerland.

Kaiwen Zuo and Yirui Jiang. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv:2412.18947*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv:2501.18362*.

A Statistics

Figure 4 illustrates the temporal distribution of the surveyed papers, highlighting a significant and accelerating interest in the evaluation of multimodal foundation models. The data reveals a foundational period with a modest number of publications (6 papers) appearing before 2022, followed by a slight decrease in 2022 (4 papers). However, a marked increase is observed in 2023, with 10 papers published. This upward trend dramatically intensifies in 2024, which saw a striking peak of 41 publications. The momentum continues into 2025, with 22 papers already published by May. This pronounced surge underscores the rapidly expanding research landscape and the community's intensified focus on developing and assessing multimodal AI systems, thereby emphasizing the timeliness and necessity of a comprehensive survey such as this paper.

Figure 5 presents a word cloud generated from the titles of the surveyed research papers, offering a visual representation of the prominent themes and focal points within the field of multimodal foundation model evaluation. The most salient terms, such as "language," "model," "medical," "evaluation," "foundation," "reasoning," "large," "multimodal," and "understanding," collectively underscore the core research thrust. The prevalence of "evaluation" and "benchmarks," alongside specific capabilities like "reasoning," "instruction following," and "alignment," highlights the community's intense focus on developing robust methodologies for assessing these complex systems. Furthermore, the significant presence of terms related to diverse modalities—including "vision," "audio," and "video"—and application domains, particularly "medical," "clinical," and "genomic," reflects the broad impact and a strong leaning towards real-world applications, especially in healthcare. This thematic clustering emphasizes the research landscape's drive towards not only advancing the capabilities of large multimodal foundation models but also rigorously evaluating their performance and applicability across a spectrum of tasks and domains, aligning with the hierarchical perspective adopted in this survey.

B Takeaway Insights

B.1 Key Findings

The field is undergoing significant transformations in evaluation methodologies and frameworks:

(1) Shift from evaluating **single-modal** to **multimodal** that integrate images, video and audio. This transition emphasizes the need for comprehensive evaluation metrics that can capture the rich interactions between various modalities and their combined impact on performance.

(2) Transition from focusing on **single-task evaluations** to **multi-task evaluations**, where models are assessed on diverse tasks such as question answering, generation, and reasoning. This shift reflects the growing complexity of real-world applications that require a broader range of capabilities beyond task-specific performance.

(3) Growing interest in **omni-model evaluations**, where unified models are evaluated across a wide array of tasks and modalities. These evaluations not only test the robustness of models but also their adaptability to different use cases, driving the development of more generalizable and versatile models.

(4) Increasing emphasis on **real-world scenario evaluations** that move beyond isolated capability tests. These scenarios provide richer, more authentic contexts for assessing robustness, usability, and model alignment with human expectations, revealing potential that synthetic benchmarks may overlook.

B.2 Critical Limitations

Current approaches face several key challenges:

(1) **Adversarial vulnerability**, where multimodal foundation models are susceptible to subtle perturbations that undermine their reliability and safety.

(2) Insufficient understanding and evaluation of model performance under real-world **multimodal distribution shifts**, limiting their robustness and generalizability.

(3) Lack of systematic evaluation benchmarks addressing inherent **biases and spurious correlations**, potentially leading to unfair or misleading outcomes.

B.3 Future Directions

Addressing these limitations requires advances in:

(1) Developing specialized **adversarial robustness frameworks** tailored specifically for multimodal foundation models to ensure reliability against subtle attacks.

(2) Constructing comprehensive benchmarks and evaluation methodologies to accurately measure and enhance model resilience against **mul-**

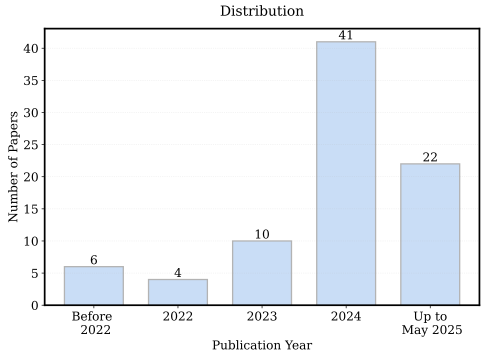


Figure 4: Distribution on publication year of surveyed papers.

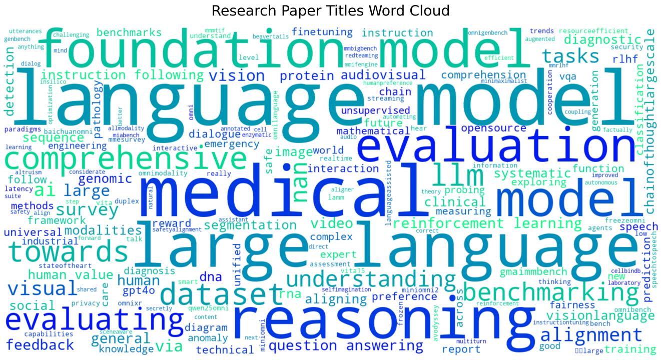


Figure 5: Word cloud of research paper titles.

timodal distribution shifts** observed in real-world applications.

(3) Establishing rigorous methods for systematically detecting, evaluating, and mitigating **biases and spurious correlations** inherent in multimodal datasets and models.

C Other Remarkable Real-World Applications

C.1 Industry Applications

The application of LLMs in industrial settings demonstrates remarkable technical adaptability and practical value, primarily attributed to the structured characteristics and relatively well-defined interaction paradigms inherent to industrial scenarios. Compared to general consumer-oriented applications, industrial environments feature clearly delineated task boundaries, highly standardized professional terminology, and rigorously defined operational workflows. These attributes significantly reduce the implementation barriers for MLLMs, positioning industrial applications as a highly promising domain at the current stage of technological development.

Robotics. Kiva warehouse-management system (Wurman et al., 2008) pioneered large-scale warehouse automation by deploying autonomous mobile robots to transport storage shelves, significantly improving productivity and operational flexibility in industrial settings. To address data scarcity in robotics, recent work (Chen et al., 2024a) introduces a simulation-based evaluation framework for foundation models, emphasizing task quality, diversity, and generalization for real-world applicability. MMRo (Li et al., 2024c) proposes a benchmark testing perception, planning, reasoning, and safety, finding no MLLM universally reliable as a robotic "brain".

Anomaly Detection. MPDD (Jezek et al., 2021) is a benchmark dataset for defect detection in industrial metal parts, featuring over 1,000 images with pixel-level annotations, while MVTec AD (Bergmann et al., 2019) is a more comprehensive multi-category anomaly detection dataset with 5,354 high-resolution images covering 70+ defect types, both providing anomaly-free training samples and annotated anomalies for testing. MMAD (Jiang et al., 2024) establishes a full-spectrum benchmark for industrial anomaly detection, evaluating MLLMs across seven subtasks with 39,672 questions, revealing a performance gap against industrial needs.

Industrial Knowledge Integration. Early work (Geipel, 2024) outlines requirements for industrial engineering benchmarks, advocating for practitioner-oriented evaluations of MLLMs on technical artifacts like datasheets and schematics. DesignQA (Doris et al., 2025) targets engineering documentation comprehension, combining textual requirements, CAD images, and drawings from Formula SAE; it exposes MLLMs' struggles in rule retrieval, technical component recognition, and drawing analysis. MME-Industry (Yi et al., 2025) broadens the scope with a cross-industry benchmark spanning 21 domains (1,050 QA pairs), emphasizing domain expertise and bilingual (Chinese/English) evaluation. Collectively, these studies highlight MLLMs' potential in industrial applications while underscoring critical limitations—domain-specific knowledge gaps, multimodal reasoning deficiencies, and safety concerns—that must be addressed to meet real-world demands.

Discussion. Current MLLMs excel in structured industrial settings due to standardized workflows and bounded tasks, enabling efficient automation in robotics, quality control, and documentation. However, challenges like domain-specific knowledge gaps, robustness in dynamic environments, and high computational costs persist. Future advancements should focus on lightweight deployment, real-time adaptability, and seamless human-AI collaboration to unlock broader industrial potential.

C.2 Society

Multimodal foundation models exhibit transformative potential in addressing societal challenges through cross-modal integration and contextual reasoning. Their societal applications span two critical dimensions: enhancing social functioning in operational systems like transportation and agriculture through environment-aware data processing, and advancing social good via assistive technologies that promote accessibility and inclusion. Crucially, systematic evaluation of such models' environmental adaptability, human-AI collaboration efficacy, and ethical alignment becomes imperative. Rigorous benchmarking of these capabilities not only validates technical robustness but also bridges the gap between laboratory performance and real-world deployment, thereby amplifying their societal impact through improved operational efficiency and equitable social benefits.

Social Functioning. Multimodal foundation models have already impacted various aspects of society, enhancing the efficiency of many social functions. TransportationGames (Zhang et al., 2024f) introduces a complete benchmark designed to reliably assess the ability of multimodal foundation models to handle transportation-related tasks. While MM-Soc (Jin et al., 2024c) focuses on evaluating social connections understanding. Recently, AgriBench (Zhou and Ryo, 2024) and AgMMU (Gauba et al., 2025) have been developed to measure the performance of multimodal foundation models in agriculture-specific contexts, aiming to capture their effectiveness in addressing domain-relevant tasks such as crop monitoring, disease detection, and environmental interpretation.

Social Good. Beyond optimizing societal efficiency, evaluating the contribution of multimodal foundation models to social good carries profound significance. These models have the potential to empower marginalized communities, enhance accessibility, and foster inclusive human-AI interaction. For instance, VizWiz (Gurari et al., 2018) provides a benchmark for assessing how effectively models assist blind users by answering questions about visual scenes captured through mobile devices. Meanwhile, Vision-Braille (Wu et al., 2024a) evaluates the performance of models in translating visual content into Braille, aiming to improve information access for individuals with visual impairments. These benchmarks highlight critical capabilities such as perceptual grounding, adaptive response generation, and sensitivity to human context. Systematic evaluation in these areas ensures that model development aligns with ethical goals and real human needs, advancing not only technological inclusivity but also societal equity.

Discussion. Current evaluations for societal applications are constrained by fragmented data collection methods and inconsistent benchmark designs, which fail to reflect the diverse and evolving needs of real-world environments. The lack of unified standards for measuring how models adapt to different social scenarios and collaborate with humans limits their broader adoption. Developing shared evaluation practices with real-world relevance is critical to guide multimodal foundation models toward practical, scalable solutions across social domains.

D Acknowledgment of AI Assistance in Writing and Revision

We utilized LLMs for the purposes of grammar correction and language refinement, in accordance with the ACL Policy on AI Writing Assistance. All ideas and technical content are the original work of the authors.

E Literature Review Summary

To offer a thorough overview of the literature surveyed, we provide a comprehensive summary table encompassing all cited works. Each entry in the table includes seven essential fields: **Title** (the complete title of the paper), **TLDR** (a concise synopsis of the paper's key contributions), **Category** (the main research focus within data-efficient LLM post-training), **Year** (the year of publication), **Venue** (the publishing venue), and **Link** (a direct URL to the paper). This organized presentation allows readers to efficiently access the original sources, trace the progression of research across categories, and utilize the compilation as a valuable reference point for future work in the field.

Table 1: Summary of Referenced Papers

Title	TLDR	Category	Year	Venue	Link
AV-Odyssey Bench: Can Your Multimodal LLMs Really Understand Audio-Visual Information?	The paper proposes AV - Odyssey Bench to assess MLLMs' understanding of audio - visual info, aiming to guide future dataset collection and model development.	Core Foundational Competencies	2024	Arxiv	link
Qwen2.5-Omni Technical Report	This paper presents Qwen2.5 - Omni, with block - wise encoders, TMRoPE, Thinker - Talker archi., and sliding - window DiT for multimodal streaming.	MLLM	2025	Arxiv	link
OmniBench: Towards The Future of Universal Omni-Language Models	The paper introduces OmniBench to evaluate omni - language models' tri - modal processing. It also develops OmniInstruct and advocates better techniques for OLM performance.	Core Foundational Competencies, Higher-Order Intelligence	2024	Arxiv	link
Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming	Paper introduces Mini - Omni, an end - to - end model for real - time speech interaction. Proposes methods, dataset; first open - source for such task.	Background	2024	Arxiv	link
Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities	The paper introduces Mini - Omni2, a visual - audio assistant akin to GPT - 4o. It has a three - stage training and interruption mechanism for flexible interaction.	Background	2024	Arxiv	link
MME-Survey: A Comprehensive Survey on Evaluation of Multimodal LLMs	This paper comprehensively surveys MLLM evaluation, covering benchmarks, construction, evaluation manner, and outlook, guiding researchers to evaluate MLLMs effectively.	Background	2024	Arxiv	link
A Survey on Benchmarks of Multimodal Large Language Models	This paper reviews 200 benchmarks for MLLMs from five aspects, discusses limitations and future directions, emphasizing evaluation's importance for MLLM development.	Background	2024	Arxiv	link
A Survey on Multimodal Large Language Models	This paper traces and summarizes MLLM progress, covering concepts, extensions, techniques, and points out challenges and directions.	Background	2023	NSR	link
A Survey on Evaluation of Large Language Models	This paper comprehensively reviews LLM evaluation methods from three dimensions, summarizes cases, and points out future challenges to aid LLM development.	Background	2023	TIST	link
A SURVEY OF RESOURCE-EFFICIENT LLM AND MULTIMODAL FOUNDATION MODELS	This survey analyzes resource - efficient strategies for large foundation models, covering algorithmic and systemic aspects to inspire future breakthroughs.	Background	2024	Arxiv	link
Assessment of Multimodal Large Language Models in Alignment with Human Values	The paper introduces Ch3Ef dataset and unified strategy to assess MLLMs' alignment with human values, summarizing key findings for field advancement.	Real-World Applications	2024	Arxiv	link
VITA: Towards Open-Source Interactive Omni Multimodal LLM	The paper introduces VITA, the first open - source MLLM for video, image, text, audio. It enhances interaction and paves the way for open - source multimodal research.	Real-World Applications	2024	Arxiv	link
Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM	This paper proposes Freeze - Omni, a novel speech - text multimodal LLM. It connects speech to frozen LLM, enabling speech - to - speech with low latency and duplex dialogue.	Real-World Applications	2024	Arxiv	link
VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction	The paper proposes a multi - stage training method for LLM to enable vision - speech interaction, bypassing ASR/TTS and accelerating end - to - end response.	Real-World Applications	2025	Arxiv	link
OmnixR: Evaluating Omni-modality Language Models on Reasoning across Modalities	The paper introduces OmnixR, an evaluation suite for OLMs. It offers two variants and provides a more rigorous cross - modal testbed than existing benchmarks.	Real-World Applications	2024	Arxiv	link
Baichuan-Omni Technical Report	The paper introduces Baichuan-omni, the first 7B open-source MLLM. It proposes a training schema, aiming to be a baseline for multimodal understanding.	Real-World Applications	2024	Arxiv	link
LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark	The paper presents LAMM, an open - source multi - modal framework. It offers a dataset, benchmark, methodology, and a training framework to boost MLLM research.	Real-World Applications	2023	NIPS	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
Audio Visual Scene-Aware Dialog	The paper introduces the AVSD Dataset for scene - aware dialog. It trains baselines, emphasizing models should use all inputs for best results.	Real-World Applications	2019	CVPR/ICCV/ECCV	link
MIA-Bench: Towards Better Instruction Following Evaluation of Multimodal LLMs	The paper introduces MIA - Bench for MLLM instruction - following evaluation, creates extra data, and aims to guide MLLM training methods.	Real-World Applications	2024	Arxiv	link
MM-BigBench: Evaluating Multimodal Models on Multimodal Content Comprehension Tasks	This paper introduces MM - BigBench to comprehensively evaluate MLLMs on multimodal content comprehension, proposes new metrics and offers novel insights.	Real-World Applications	2023	Arxiv	link
Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback	The paper first fine - tunes all - modality models with cross - modality data. It proposes align - anything and eval - anything, and open - sources relevant resources.	Real-World Applications	2024	Others	link
MMMT-IF: A Challenging Multi-Modal Multi-Turn Instruction Following Foundation Model Benchmark	The paper proposes MMMT - IF, a new evaluation set for multi - modal multi - turn instruction following, and PIF metrics, highlighting models' limitations.	Real-World Applications	2024	Arxiv	link
MM-RLHF: The Next Step Forward in Multimodal LLM Alignment	The paper introduces MM - RLHF dataset and proposes Critique - Based Reward Model and Dynamic Reward Scaling for better MLLM alignment.	Real-World Applications	2025	Arxiv	link
Aligning Large Multimodal Models with Factually Augmented RLHF	The paper adapts RLHF to vision - language alignment, proposes Factually Augmented RLHF, enhances training data, and develops an evaluation benchmark.	Real-World Applications	2024	*ACL	link
MM-IFEngine: Towards Multimodal Instruction Following	The paper presents MM - IFEngine to generate high - quality image - instruction pairs, introduces MM - IFEval benchmark, and fully open - sources related data and code.	Real-World Applications	2025	Arxiv	link
Aligning Multimodal LLM with Human Preference: A Survey	This paper surveys alignment algorithms for MLLMs, covering app scenarios, dataset factors, benchmarks, and future directions to inspire better methods.	Real-World Applications	2025	Arxiv	link
Autonomous Alignment with Human Value on Altruism through Considerate Self-imagination and Theory of Mind	This paper endows agents with self - imagination and ToM, enabling them to align with human altruistic values and make thoughtful decisions.	Real-World Applications	2024	Patterns	link
Aligning AI With Shared Human Values	The paper introduces the ETHICS dataset to assess LMs' moral knowledge, showing progress in machine ethics for AI aligned with human values.	Survey	2020	Arxiv	link
A General Language Assistant as a Laboratory for Alignment	The paper explores alignment of text - based assistants with human values, studies techniques, training objectives, and a pre - training stage for efficiency.	Real-World Applications	2021	Arxiv	link
Training language models to follow instructions with human feedback	The paper presents InstructGPT, fine - tuned with human feedback on user intent. It shows fine - tuning is promising for aligning models with humans.	Real-World Applications	2022	NIPS	link
Direct Preference Optimization: Your Language Model is Secretly a Reward Model	The paper presents Direct Preference Optimization (DPO), which simplifies preference alignment of LMs, eliminating complex steps of RLHF with a single - stage training.	Real-World Applications	2023	NIPS	link
Safe RLHF: Safe Reinforcement Learning from Human Feedback	The paper proposes Safe RLHF, decoupling helpfulness and harmlessness. It formalizes safety as an optimization task and adjusts objective balance during fine - tuning.	Real-World Applications	2024	NIPS	link
A Minimaximalist Approach to Reinforcement Learning from Human Feedback	The paper presents SPO, a minimalist-maximalist RL algorithm from human feedback. It uses MW concept, enabling single-agent self - play for efficient learning.	Real-World Applications	2024	ICLR	link
Aligner: Efficient Alignment by Learning to Correct	The paper introduces Aligner, a model-agnostic, plug-and-play alignment paradigm. It learns correctional residuals, applies to various models, and can bootstrap them.	Real-World Applications	2024	NIPS	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset	The paper introduces the BEAVERTAILS dataset for LLM safety alignment, separates annotations, and shows its applications, aiding safe LLM development.	Real-World Applications	2023	NIPS	link
Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment	The paper proposes RED - EVAL for safety evaluation, RED - INSTRUCT for alignment, and a fine - tuned model STARLING for LLM safety.	Real-World Applications	2023	Arxiv	link
CellBinDB: A Large-Scale Multimodal Annotated Dataset for Cell Segmentation with Benchmarking of Universal Models	The paper presents CellBinDB, a large - scale multimodal annotated dataset for cell segmentation, and benchmarks 7 methods and analyzes influence factors.	Real-World Applications, cell, science	2024	bioRxiv	link
OmniGenBench: Automating Large-scale in-silico Benchmarking for Genomic Foundation Models	The paper introduces GFMBench, an open - source framework for GFM benchmarking, standardizing suites and democratizing GFM applications.	Real-World Applications, science, dna, rna, omic	2024	Arxiv	link
Benchmarking DNA Foundation Models for Genomic Sequence Classification	A benchmarking study of three DNA models on 57 datasets. Mean token embedding boosts performance, offering a framework for model selection in genomics.	Real-World Applications	2024	Arxiv	link
A Benchmark Dataset for Multimodal Prediction of Enzymatic Function Coupling DNA Sequences and Natural Language	The paper proposes a novel dataset and benchmark suite for multimodal prediction of enzymatic function using DNA sequences and natural - language descriptions. [Dataset Link: https://hoarfrost - lab.github.io/BioTalk/.]	Real-World Applications	2024	Arxiv	link
GenBench: A Benchmarking Suite for Systematic Evaluation of Genomic Foundation Models	This paper introduces GenBench, a benchmarking suite for evaluating Genomic Foundation Models, and reveals model preference insights for future GFM design.	Real-World Applications	2024	Arxiv	link
BEND: Benchmarking DNA Language Models on biologically meaningful tasks	The paper introduces BEND, a benchmark for DNA LMs with realistic tasks on the human genome, available at https://github.com/frederikkemarin/BEND.	Real-World Applications	2024	Arxiv	link
Evaluating Vision and Pathology Foundation Models for Computational Pathology: A Comprehensive Benchmark Study	This paper benchmarks 31 AI foundation models for computational pathology, reveals factors influencing performance, and advocates further research.	Real-World Applications	2025	medRxiv	link
scDrugMap: Benchmarking Large Foundation Models for Drug Response Prediction	The paper develops scDrugMap, an integrated framework for drug response prediction, offering a large - scale benchmark for single - cell data in drug discovery.	Real-World Applications	2025	Arxiv	link
GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI	The paper develops GMAI - MMBench, a comprehensive medical AI benchmark, and identifies LVLMs' insufficiencies to boost next - gen LVLMs toward GMAI.	Real-World Applications	2024	NeurIPS	link
Multimodal Clinical Benchmark for Emergency Care (MC-BEC): A Comprehensive Benchmark for Evaluating Foundation Models in Emergency Medicine	The paper proposes MC - BEC, a benchmark for evaluating ED foundation models with 100K+ visits' data, aiming to encourage more effective multimodal model development.	Real-World Applications	2023	NeurIPS	link
COMET: Benchmark for Comprehensive Biological Multi-omics Evaluation Tasks and Language Models	The paper introduces COMET, the first multi - omics benchmark, curates tasks/datasets, evaluates models, and guides multi - omics research.	Real-World Applications	2024	Arxiv	link
A Fine-tuning Dataset and Benchmark for Large Language Models for Protein Understanding	This paper introduces ProteinLMDataset and ProteinLMBench for LLMs' protein understanding, addressing dataset gaps and setting a new eval standard.	Real-World Applications	2024	BIBM	link
ProteinBench: A Holistic Evaluation of Protein Foundation Models	The paper presents ProteinBench, a holistic evaluation framework for protein foundation models, releasing dataset, code, and leaderboard to drive field development.	Real-World Applications	2024	Arxiv	link
Prot2Text: Multimodal Protein's Function Generation with GNNs and Transformers	This paper proposes Prot2Text, integrating GNNs and LLMs to predict protein functions in free - text, enabling holistic protein representation.	Real-World Applications	2024	AAAI	link
Bridging Sequence-Structure Alignment in RNA Foundation Models	This paper introduces OmniGenome, an RNA FM that aligns sequences with secondary structures, enabling bidirectional mappings and outperforming existing FMs.	Real-World Applications	2025	AAAI	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
Redesigning the Eterna100 for the Vienna 2 folding engine	The paper introduces Eterna100 - V2 benchmark for RNA design, shows design difficulty links to models, and advocates model - agnostic algorithms.	Real-World Applications	2025	bioRxiv	link
DRFormer: A Benchmark Model for RNA Sequence Downstream Tasks	The paper proposes DRFormer, the first multimodal RNA downstream model integrating sequence & vision models via structural features, advancing RNA research.	Real-World Applications	2025	Genes (Basel)	link
BEACON: Benchmark for Comprehensive RNA Tasks and Language Models	The paper introduces BEACON, a comprehensive RNA benchmark with 13 tasks. It assesses models and proposes BEACON - B, open - sourced on GitHub.	Real-World Applications	2024	NeurIPS	link
Multi-task benchmarking of single-cell multimodal omics integration methods	The paper aims to provide a guideline for single - cell multimodal omics data analysis via systematic categorization and benchmarking of current methods.	Real-World Applications	2024	bioRxiv	link
M3SciQA: A Multi-Modal Multi-Document Scientific QA Benchmark for Evaluating Foundation Models	The paper introduces M3SciQA, a multi - modal multi - document benchmark for foundation model evaluation, and explores its implications for future application.	Real-World Applications	2024	Arxiv	link
MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI	The paper introduces MMMU, a benchmark for multimodal models on multi - discipline tasks, aiming to drive next - gen models towards expert AGI.	Real-World Applications	2024	CVPR/ICCV/ECCV	link
AVQA: A Dataset for Audio-Visual Question Answering on Videos	The paper introduces AVQA, a real - life video AVQA dataset with 57k+ videos and Q - A pairs, and a fusing module for better multimodal understanding.	reasoning	2022	ACM MM	link
EchoInk-R1: Exploring Audio-Visual Reasoning in Multimodal LLMs via Reinforcement Learning	The paper introduces EchoInk - R1, a RL framework on Qwen2.5 - Omni - 7B. It unifies audio, visual, and text for reasoning, enhancing MLLMs' cross - modal reasoning.	reasoning	2025	Arxiv	link
MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts	The paper presents MathVista benchmark for evaluating mathematical reasoning in visual contexts, highlights GPT - 4V's potential, and points out gaps for future AI development.	reasoning	2023	Arxiv	link
CMMCoT: Enhancing Complex Multi-Image Comprehension via Multi-Modal Chain-of-Thought and Memory Augmentation	The paper proposes CMMCoT, a framework mimicking human slow thinking for multi - image understanding, with key innovations and a new dataset.	reasoning	2025	Arxiv	link
VisCoTVisual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning	The paper presents a large-scale Visual CoT dataset, a multi-turn pipeline, and a benchmark to enhance MLLMs' interpretability and local region identification.	reasoning	2024	NIPS	link
Unified Reward Model for Multimodal Understanding and Generation	This paper proposes UnifiedReward, a unified reward model for multimodal assessment, enabling preference alignment via joint task learning and DPO.	reasoning	2025	Arxiv	link
Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset	The paper presents the MATH - V dataset from real math competitions to evaluate LMMs' math reasoning, aiding future R & D with categorization.	reasoning	2024	NIPS	link
MATHVERSE: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?	The paper introduces MathVerse, a visual math benchmark, and a CoT evaluation strategy to assess MLLMs' diagram - understanding ability for future development.	reasoning	2024	CVPR/ICCV/ECCV	link
LLaVA-CoT: Let Vision Language Models Reason Step-by-Step	The paper introduces LLaVA-CoT, a VLM for autonomous multi - stage reasoning. It compiles a dataset and proposes a search method, enhancing reasoning abilities.	reasoning	2025	Arxiv	link
A Diagram is Worth a Dozen Images	This paper studies diagram interpretation, introduces DPGs, devises parsing methods, compiles a dataset, showing DPGs' significance for diagram tasks.	reasoning	2016	CVPR/ICCV/ECCV	link
VRC-Bench	VRC-Bench offers a comprehensive visual reasoning benchmark assessing reasoning chains and final outcomes in complex scenarios with semi-automated annotation and manual verification.	reasoning	2025	Arxiv	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning	The paper presents a diagnostic dataset for visual reasoning with minimal biases and detailed annotations to analyze models' abilities and limitations.	reasoning	2017	CVPR/ICCV/ECCV	link
Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering	The paper presents ScienceQA benchmark with multimodal questions and answer annotations, and designs models to generate CoT, showing CoT's utility in QA.	reasoning	2022	NIPS	link
Exploring the Effect of Reinforcement Learning on Video Understanding: Insights from SEED-Bench-R1	The paper introduces SEED - Bench - R1 to evaluate MLLMs in video understanding. It compares RL and SFT, reveals RL's pros and cons, and suggests future improvements.	reasoning	2025	Arxiv	link
A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge	The paper introduces A - OKVQA, a crowdsourced VQA dataset with 25K diverse questions needing world knowledge and commonsense reasoning.	reasoning	2022	Arxiv	link
LISA: Reasoning Segmentation via Large Language Model	The paper proposes reasoning segmentation, creates a benchmark, and presents LISA, unlocking new reasoning abilities for multimodal LLMs.	reasoning	2024	CVPR/ICCV/ECCV	link
M3CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought	The paper introduces a novel M3CoT benchmark to address MCoT challenges, advancing multi - domain, multi - step, multi - modal CoT research.	reasoning	2024	Arxiv	link
3D-CoT	The paper extends 3D datasets with reasoning annotations in multiple styles, constructs 3D - CoT Benchmark from diverse sources for CoT study.	reasoning	2025	Arxiv	link
VisuLogic: A Benchmark for Evaluating Visual Reasoning in Multi-modal Large Language Models	The paper introduces VisuLogic, a benchmark with 1K verified problems, to assess MLLMs' visual reasoning. It also provides training data and RL baseline.	reasoning	2025	Arxiv	link
X-Reasoner: Towards Generalizable Reasoning Across Modalities and Domains	The paper explores cross - modality and - domain reasoning generalizability. It introduces X - Reasoner and X - Reasoner - Med via text post - training, enabling wider reasoning transfer.	reasoning	2025	Arxiv	link
VSI-bench	The paper presents VSI - Bench with over 5,000 QA pairs, studies MLLMs' visual - spatial intelligence, and finds cognitive maps enhance spatial distance ability.	reasoning	2024	Arxiv	link
ST-Align benchmark	The paper proposes LLaVA - ST for fine - grained spatio - temporal understanding, presents ST - Align dataset and benchmark, with novel methods for alignment and compression.	reasoning	2025	CVPR/ICCV/ECCV	link
GeomVerse: A Systematic Evaluation of Large Models for Geometric Reasoning	This paper evaluates VLMs' geometric reasoning via a synthetic geometry dataset, revealing their limitations and releasing data for further research.	reasoning	2023	Arxiv	link
InteractVLM: 3D Interaction Reasoning from 2D Foundational Models	The paper introduces InteractVLM to estimate 3D contact points from 2D images, using a novel module. It also proposes a new semantic contact estimation task.	reasoning	2025	CVPR/ICCV/ECCV	link
Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning	The paper proposes Reason - RFT, a two - phase reinforcement fine - tuning framework for visual reasoning, enhancing generalization and advancing multimodal research.	reasoning	2025	Arxiv	link
LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding	The paper introduces LongVideoBench, a QA benchmark for long video-lang. interleaved inputs, with novel tasks and diverse questions to assess LMMs.	streaming-input	2024	NIPS	link
EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding	The paper introduces EgoSchema, a long - form video QA dataset. It also proposes temporal certificate sets, valuable for long - term video understanding system evaluation.	streaming-input	2023	Arxiv	link
ViSMaP: Unsupervised Hour-long Video Summarisation by Meta-Prompting	ViSMaP offers unsupervised hour - long video summarization. It uses meta - prompting and LLMs with short - video data, bypassing long - video annotations.	streaming-input	2025	Arxiv	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI	This paper proposes the GMAI-MMBench, a comprehensive medical AI benchmark with multi-granular data, supporting customized evaluation and contributing to the development of medical AI.	Real-World Applications	2024	NeurIPS	link
OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM	This paper introduces the OmniMedVQA medical VQA benchmark, which covers real-world medical images with multiple modalities and anatomical regions, highlighting the need to build a more powerful large model in the biomedical field.	Real-World Applications	2024	CVPR	link
MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding	This paper proposes the MedXpertQA benchmark, which contains diverse questions. After rigorous processing, a reasoning subset is set up to evaluate the capabilities of medical models.	Real-World Applications	2025	arxiv	link
SURE-VQA: SYSTEMATIC UNDERSTANDING OF ROBUSTNESS EVALUATION IN MEDICAL VQA TASKS	The paper proposes the SURE-VQA framework to evaluate the robustness of VLMs in medical VQA tasks and overcomes the existing deficiencies from three aspects, being systematic.	Real-World Applications	2024	arxiv	link
CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models	This paper proposes CARES to comprehensively evaluate the trustworthiness of Medical Vision Language Models, assessing them from five dimensions and containing about 41K question-answer pairs.	Real-World Applications	2024	NeurIPS	link
RJUA-MedDQA: A Multimodal Benchmark for Medical Document Question Answering and Clinical Reasoning	This paper establishes the medical multimodal benchmark RJUA - MedDQA and proposes the ESRA method to improve annotation efficiency and accuracy, which helps to promote the application of medical document understanding.	Real-World Applications	2024	KDD	link
MedHallBench: A New Benchmark for Assessing Hallucination in Medical Large Language Models	This paper proposes the MedHallBench evaluation framework, which combines multiple methods to evaluate hallucinations in MLLMs, laying a foundation for enhancing their reliability in healthcare.	Real-World Applications	2024	arxiv	link
BenchX: A Unified Benchmark Framework for Medical Vision-Language Pretraining on Chest X-Rays	The paper proposes the BenchX unified benchmark framework, which contains multiple components and is used for the comparative analysis of MedVLP methods, driving the re - examination of achievements in the field.	Real-World Applications	2024	NeurIPS	link
Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA	The paper introduces the ProbMed dataset to evaluate the performance of large models in medical VQA, reveals the limitations of the models, studies the deficiencies of open - source models, and proposes improvement solutions.	Real-World Applications	2024	arxiv	link
Touchstone Benchmark: Are We on the Right Way for Evaluating AI Algorithms for Medical Segmentation	The paper presents the Touchstone benchmark for abdominal organ segmentation, evaluates AI using diverse test sets, and assesses existing frameworks to promote innovation in medical AI algorithms.	Real-World Applications	2024	NeurIPS	link
FMBENCH: BENCHMARKING FAIRNESS IN MULTIMODAL LARGE LANGUAGE MODELS ON MEDICAL TASKS	Proposed FMBench to evaluate the fairness of Multimodal Large Language Models (MLLMs) on medical tasks, including multiple attributes and a new metric, to assist in model evaluation and advance the field.	Real-World Applications	2024	arxiv	link
Evaluating LLM - Generated Multimodal Diagnosis from Medical Images and Symptom Analysis	We propose an LLM evaluation paradigm, evaluate the accuracy of diagnoses using publicly available MCQs, analyze the results to identify deficiencies, and this approach can be used to evaluate other LLMs.	Real-World Applications	2024	arxiv	link
Evaluating multimodal AI in medical diagnostics	This study assesses the accuracy and responsiveness of multimodal AI in medical diagnostic questions and compares it with human intelligence, revealing the potential and limitations of AI.	Real-World Applications	2024	npj Digital Medicine	link
FairMedFM: Fairness Benchmarking for Medical Imaging Foundation Model	The paper proposes FairMedFM, a fairness benchmark for medical imaging foundation models. It integrates multiple datasets and models, evaluates fairness from multiple perspectives, and its code is open - source.	Real-World Applications	2024	NeurIPS	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
Large Language Model Benchmarks in Medical Tasks	This paper comprehensively surveys benchmark datasets for medical large language models, covering multiple modalities. It summarizes the challenges and opportunities, laying a foundation for research on the application of large language models in medicine.	Real-World Applications	2024	arxiv	link
3MDBench: Medical Multimodal Multi-agent Dialogue Benchmark	Propose the 3MDBench evaluation framework, simulate patient diversity, combine multimodal data, and provide a scalable solution for AI medical assistant evaluation.	Real-World Applications	2025	arxiv	link
Asclepius: A Spectrum Evaluation Benchmark for Medical Multi-Modal Large Language Models	This paper proposes the Asclepius evaluation benchmark to comprehensively assess medical multi-modal large language models, analyze their advantages and disadvantages, and lay the foundation for subsequent evaluations and applications.	Real-World Applications	2024	arxiv	link
WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation	The paper presents the multilingual and multimodal medical dataset WorldMedQA - V, which includes data from multiple countries. It aims to adapt to diverse medical environments and promote the fair and effective application of AI.	Real-World Applications	2024	arxiv	link
MEDICONFUSION: CAN YOU TRUST YOUR AI RADIOLOGIST? PROBING THE RELIABILITY OF MULTIMODAL MEDICAL FOUNDATION MODELS	This paper introduces the MediConfusion medical VQA benchmark dataset to explore the failure modes of multimodal large models and facilitate the design of more reliable models.	Real-World Applications	2024	arxiv	link
MMIST-ccRCC: A Real World Medical Dataset for the Development of Multi-Modal Systems	The paper introduces the MMIST-CCRCC multi-modal dataset and sets multi-modal benchmarks. Even when the data missing rate is severe, fusing modalities can improve the survival prediction results.	Real-World Applications	2024	CVPR	link
Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions	This paper introduces a new dataset for painted metal parts defect detection and evaluates SOTA AD methods, highlighting real - world dataset testing need.	Real-World Applications	2021	IEEE	link
MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection	The paper introduces MVTec AD, a comprehensive real - world dataset for anomaly detection with pixel - precise ground truth, focusing on real - world use.	Real-World Applications	2019	CVPR/ICCV/ECCV	link
MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection	The paper presents MMAD, the first MLLMs benchmark in industrial anomaly detection, defines subtasks, generates a dataset, and explores enhancement strategies.	Real-World Applications	2024	Arxiv	link
DesignQA: A Multimodal Benchmark for Evaluating Large Language Models' Understanding of Engineering Documentation	The paper introduces DesignQA, a multimodal benchmark for MLLMs on engineering docs. It reveals gaps and paves the way for AI - supported design.	Real-World Applications	2025	Others	link
MME-Industry: A Cross-Industry Multimodal Evaluation Benchmark	This paper introduces MME - Industry, a novel benchmark for evaluating MLLMs in industrial scenarios. It has unique QA pairs and bilingual versions, guiding future research.	Real-World Applications	2025	Arxiv	link
MMRo: Are Multimodal LLMs Eligible as the Brain for In-Home Robotics?	This paper introduces the first MMRo benchmark to evaluate MLLMs for robotics, identifying 4 key capabilities, and finds current MLLMs unfit as robot cores.	Real-World Applications	2024	Arxiv	link
Coordinating Hundreds of Cooperative, Autonomous Vehicles in Warehouses	The paper presents the Kiva warehouse - management system, a large - scale autonomous robot system that enhances worker productivity, accountability, and flexibility.	Real-World Applications	2008	AAAI	link
On the Evaluation of Generative Robotic Simulations	The paper proposes a comprehensive framework for evaluating generative robotic simulations in quality, diversity, and generalization, highlighting balance needs.	Real-World Applications	2024	Arxiv	link
Towards a Benchmark of Multimodal Large Language Models for Industrial Engineering	This paper outlines requirements and proposes a starting - point for MLLM benchmark in industrial applications, targeting industry practitioners.	Real-World Applications	2024	IEEE	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
Measuring Social Norms of Large Language Models	The paper presents a new dataset for testing LLMs' social norm understanding, proposes a multi - agent framework, and offers a direction for future improvements.	Real-World Applications	2024	*ACL	link
VizWiz Grand Challenge: Answering Visual Questions from Blind People	The paper proposes VizWiz, the first natural VQA dataset from blind people. It aims to encourage generalized algorithms to assist them.	Real-World Applications	2018	CVPR/ICCV/ECCV	link
MM-Soc: Benchmarking Multimodal Large Language Models in Social Media Platforms	The paper introduces MM - Soc to evaluate MLLMs on social media content, targeting multiple tasks, and suggests improvement pathways for models.	Real-World Applications	2024	*ACL	link
TransportationGames: Benchmarking Transportation Knowledge of (Multimodal) Large Language Models	The paper proposes TransportationGames, a benchmark for assessing (M)LLMs' transportation knowledge, aiming to boost their implementation in this domain.	Real-World Applications	2024	Arxiv	link
Beyond Good Intentions: Reporting the Research Landscape of NLP for Social Good	The paper introduces NLP4SG Papers, a dataset with 3 tasks, and visualizes NLP4SG landscape using ACL Anthology, aiding research in this field.	Real-World Applications	2023	*ACL	link
AgriBench: A Hierarchical Agriculture Benchmark for Multimodal Large Language Models	The paper introduces AgriBench for evaluating agricultural MM - LLMs and proposes MM - LUCAS dataset, offering insights for expert - knowledge MM - LLMs.	Real-World Applications	2024	Arxiv	link
AgMMU: A Comprehensive Agricultural Multimodal Understanding and Reasoning Benchmark	The paper presents AgMMU, an agri VLM benchmark with data from real convos. It aims to evaluate & develop VLMs and incorporate agri expertise.	Real-World Applications	2025	Arxiv	link
TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding	The paper proposes TimeChat for long video understanding, with key architectures and an instruction-tuning dataset, aiming to serve as a video assistant.	Higher-Order Intelligence	2024	CVPR/ICCV/ECCV	link
m &m's: A Benchmark to Evaluate Tool-Use for multi-step multi-modal Tasks	The paper introduces m&m's benchmark for multi - step multi - modal tasks, offers task plans, evaluates LLMs, and gives planner design recommendations.	Higher-Order Intelligence	2024	CVPR/ICCV/ECCV	link
AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments	The paper introduces AgentClinic, a multimodal benchmark for LLM evaluation in simulated clinical settings, and explores new ways to scrutinize clinical simulations.	Higher-Order Intelligence	2024	Arxiv	link
Enhancing clinical decision support with physiological waveforms — A multimodal benchmark in emergency care	The paper presents a dataset and protocol for multimodal decision support in emergency care, showing waveform data improves prediction, offering a basis for AI progress.	Higher-Order Intelligence	2025	Others	link
AGAV-Rater: Adapting Large Multimodal Model for AI-Generated Audio-Visual Quality Assessment	The paper introduces AGAVQA dataset and AGAV - Rater, a LMM - based model to score AGAVs multi - dimensionally, enhancing VTA performance and user experience.	Higher-Order Intelligence	2025	Arxiv	link
Mllm-compbench: A comparative reasoning benchmark for multimodal llms	The paper introduces MLLM - CompBench to evaluate MLLMs' comparative reasoning, covering 8 dimensions with 40K image - pairs, to guide future improvements.	Higher-Order Intelligence	2024	NIPS	link
A Benchmark for Optimal Multi-Modal Multi-Robot Multi-Goal Path Planning with Given Robot Assignment	The paper formalizes multi - modal multi - robot multi - goal path planning as a single problem, introduces a benchmark, and adapts planners for diverse settings.	Higher-Order Intelligence	2025	Arxiv	link
EgoPlan-Bench: Benchmarking Multimodal Large Language Models for Human-Level Planning	The paper introduces EgoPlan - Bench to evaluate MLLMs' planning abilities and EgoPlan - IT for improvement, sharing codes, data and a leaderboard.	Higher-Order Intelligence	2023	Others	link
EgoPlan-Bench2: A Benchmark for Multimodal Large Language Model Planning in Real-World Scenarios	The paper introduces EgoPlan - Bench2 to assess MLLMs' planning in real - world scenarios, reveals limitations, and proposes a training - free CoT prompting approach.	Higher-Order Intelligence	2024	Arxiv	link
MuEP: A Multimodal Benchmark for Embodied Planning with Foundation Models	The paper presents MuEP, a multimodal benchmark for embodied planning. It evaluates agents with foundation models, identifies gaps, and hopes to advance embodied AI.	Higher-Order Intelligence	2024	IJCAI	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
MFE-ETP: A Comprehensive Evaluation Benchmark for Multi-modal Foundation Models on Embodied Task Planning	The paper presents MFE - ETP, a benchmark for evaluating MFMs on embodied task planning, with a framework, diverse tasks, and an auto - evaluation platform.	Higher-Order Intelligence	2024	Arxiv	link
A comprehensive multi-modal video understanding benchmark	The paper introduces MVBench for multi-modal video understanding and develops VideoChat2, offering a novel way to define and evaluate temporal tasks.	Higher-Order Intelligence	2024	CVPR/ICCV/ECCV	link
DAVE: Diagnostic benchmark for Audio Visual Evaluation	The paper introduces DAVE, a novel benchmark for AV models. It addresses existing issues and offers insights for robust AV model development.	Higher-Order Intelligence	2025	Arxiv	link
ALLVB: All-in-One Long Video Understanding Benchmark	The paper proposes ALLVB, a long - video understanding benchmark integrating 9 tasks, with auto - annotation, large - scale data, revealing MLLMs' improvement potential.	Higher-Order Intelligence	2025	AAAI	link
UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces	The paper introduces UrbanVideo - Bench to evaluate Video - LLMs in urban embodied cognition, uses collected data to generate QAs, and validates Sim - to - Real transfer potential.	Higher-Order Intelligence	2025	Arxiv	link
ViLMA: A Zero-Shot Benchmark for Linguistic and Temporal Grounding in Video-Language Models	The paper presents ViLMA, a task - agnostic benchmark for VidLMs. It assesses models' fine - grained capabilities and catalyzes future VidLM research.	Higher-Order Intelligence	2024	ICLR	link
A benchmark for situated reasoning in real-world videos	The paper presents the STAR Benchmark for evaluating situated reasoning in real - world videos and proposes a diagnostic neuro - symbolic model to address challenges.	Higher-Order Intelligence	2024	NIPS	link
VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models	The paper presents VITATECS, a video - text dataset. It uses a taxonomy and counterfactual descriptions to evaluate VidLMs' temporal understanding, revealing research needs.	Higher-Order Intelligence	2024	CVPR/ICCV/ECCV	link
TempCompass: Do Video LLMs Really Understand Videos?	The paper proposes TempCompass benchmark with diverse temporal aspects and task formats, novel data collection strategies, and an evaluation approach, revealing poor video LLMs' temporal perception.	Higher-Order Intelligence	2024	*ACL	link
OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding?	The paper presents OVO - Bench, a novel video benchmark for online video LLMs, focusing on timestamp - based temporal awareness to drive research in video reasoning.	Higher-Order Intelligence	2025	Arxiv	link
OSCaR: Object State Captioning and State Change Representation	The paper introduces OSCaR dataset and benchmark for evaluating MLLMs, addressing limitations of traditional methods in object state change understanding.	Higher-Order Intelligence	2024	*ACL	link
Measuring Massive Multitask Language Understanding	Paper proposes a new test for a text model's multitask accuracy, evaluates understanding breadth/depth, and identifies model shortcomings.	Core Foundational Competencies	2021	ICLR	link
MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark	This paper introduces MMLU-Pro, enhancing MMLU with harder reasoning Qs, more options and less noise, a better benchmark for tracking field progress.	Core Foundational Competencies	2024	NeurIPS	link
CMMLU: Measuring massive multitask language understanding in Chinese	The paper introduces CMMLU, a Chinese benchmark for LLMs. It evaluates 18 models, finds room for improvement, and suggests enhancement directions.	Core Foundational Competencies	2024	*ACL	link
KMMLU: Measuring Massive Multitask Language Understanding in Korean	The paper proposes KMMLU, a Korean benchmark from original exams. It aims to improve Korean LLMs, and the dataset is publicly available for progress tracking.	Core Foundational Competencies	2024	Arxiv	link
Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation	This paper traces biases in multilingual MMLU, presents their impact on evaluations, and releases Global MMLU with wider coverage and bias checks.	Core Foundational Competencies	2024	Arxiv	link
MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation	The paper introduces MMLU - ProX, a multilingual benchmark for LLM evaluation, using semi - auto translation and expanding to assess multilingual capabilities.	Core Foundational Competencies	2025	Arxiv	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
MovieChat: From Dense Token to Sparse Memory for Long Video Understanding	The paper proposes MovieChat with a designed memory mechanism using the memory model to handle long - video understanding challenges, releasing a benchmark.	Higher-Order Intelligence	2023	CVPR/ICCV/ECCV	link
Microsoft COCO: Common Objects in Context	The paper presents a new dataset for object recognition in scene - understanding context, with per - instance labels and extensive crowd - sourced data.	Core Foundational Competencies	2014	Arxiv	link
LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models	This paper creates GVC data, introduces Grounding - Bench, proposes a GVC - supporting model design, making contributions to grounded visual chat.	Core Foundational Competencies	2024	CVPR/ICCV/ECCV	link
MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models	The paper introduces MMIU benchmark for evaluating LVLMs on multi - image tasks, identifying gaps to advance LVLM R & D.	Core Foundational Competencies	2025	ICLR	link
EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding	The paper introduces EgoSchema, a long - form video QA dataset. It defines temporal cert. sets and shows models' poor long - term video understanding.	Higher-Order Intelligence	2023	NIPS	link
MileBench: Benchmarking MLLMs in Long Context	The paper introduces MileBench to test MLLMs' multimodal long - context capabilities with multiple tasks, encouraging research on such abilities.	Core Foundational Competencies	2024	Arxiv	link
EAGLE: Egocentric AGgregated Language-video Engine	The paper introduces EAGLE model and EAGLE - 400K dataset, unifying egocentric video tasks and proposing evaluation metrics for MLLMs.	Higher-Order Intelligence	2024	ACM MM	link
MLLM-CompBench: A Comparative Reasoning Benchmark for Multimodal LLMs	The paper introduces MLLM - CompBench to evaluate MLLMs' comparative reasoning, collects image pairs, and uncovers their comparative ability limitations.	Core Foundational Competencies	2024	NeurIPS	link
MLVU: Benchmarking Multi-task Long Video Understanding	The paper proposes MLVU benchmark for LVU evaluation, featuring extended video lengths, diverse genres, and tasks, advancing long - video understanding research.	Higher-Order Intelligence	2024	Arxiv	link
Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech	The paper presents Dynamic - SUPERB, a speech benchmark for zero - shot multi - task models. It encourages community contribution and proposes baseline approaches.	Core Foundational Competencies	2024	ICASSP	link
Towards Event-oriented Long Video Understanding	The paper introduces Event - Bench for long - video event understanding and VIM method, addressing dataset issues and data scarcity, all resources are open - sourced.	Higher-Order Intelligence	2024	Arxiv	link
MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models	The paper introduces MuChoMusic, a benchmark for evaluating music understanding in audio - focused multimodal LMs, and points out issues in multimodal integration.	Core Foundational Competencies	2024	ISMIR	link
Video-Bench: A Comprehensive Benchmark and Toolkit for Evaluating Video-based Large Language Models	The paper proposes Video - Bench, a benchmark and toolkit for evaluating Video - LLMs with 10 tasks, offering insights for future research.	Higher-Order Intelligence	2023	Arxiv	link
AutoEval-Video: An Automatic Benchmark for Assessing Large Vision Language Models in Open-Ended Video Question Answering	The paper proposes AutoEval - Video, a benchmark for video Q&A of large vision - language models, with unique rules and adversarial annotation for evaluation.	Higher-Order Intelligence	2024	CVPR/ICCV/ECCV	link
Enabling Auditory Large Language Models for Automatic Speech Quality Evaluation	This paper proposes using auditory LLMs for automatic speech quality assessment via task - specific prompts, with interpretable outputs. Code available at URL.	Core Foundational Competencies	2025	ICASSP	link
AudioBench: A Universal Benchmark for Audio Large Language Models	The paper introduces AudioBench, a benchmark for AudioLLMs with 8 tasks and 26 datasets. It fills a gap and provides resources for future model development.	Core Foundational Competencies	2025	*ACL	link
Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis	This paper introduces Video - MME, a comprehensive MLLM video analysis benchmark with diverse features, and calls for better handling of long - sequences and multi - modal data.	Higher-Order Intelligence	2024	Arxiv	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding	The paper introduces MMBench - Video for evaluating LVLMs in video understanding, using long YouTube videos and free - form questions, with human - annotated questions and GPT - 4 assessment.	Higher-Order Intelligence	2024	NIPS	link
Synthesize, Diagnose, and Optimize: Towards Fine-Grained Vision-Language Understanding	The paper emphasizes evaluating VLMs from text - visual perspectives, introduces SPEC benchmark, and proposes an approach to optimize fine - grained understanding.	Core Foundational Competencies	2024	CVPR/ICCV/ECCV	link
VALSE : A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena	The paper proposes VALSE, a novel benchmark for V&L models on linguistic phenomena to enable fine - grained evaluation and measure future progress.	Core Foundational Competencies	2022	*ACL	link
MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos	The paper introduces MMWorld, a new benchmark for multi - discipline, multi - faceted multimodal video understanding, aiming for world model evaluation in videos.	Higher-Order Intelligence	2024	CVPR/ICCV/ECCV	link
Audiotime: A temporally-aligned audio-text benchmark dataset	The paper presents AudioTime, a strongly aligned audio - text dataset with rich temporal annotations, and provides test set and metric for model temporal control.	Core Foundational Competencies	2025	ICASSP	link
Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models	This paper introduces Modality Importance Score (MIS) to identify single - modality bias in VidQA benchmarks, guiding creation of balanced multimodal datasets.	Higher-Order Intelligence	2025	AAAI	link
Simple and Controllable Music Generation	The paper introduces MusicGen, a single LM for conditional music generation. It eliminates multi - model cascading and enables better output control.	Core Foundational Competencies	2023	NeurIPS	link
WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs	The paper introduces WorldSense, the 1st benchmark for multimodal video understanding. It has unique features and aims to evaluate models' omnimodal context - constructing ability.	Higher-Order Intelligence	2025	Arxiv	link
Vbench: Comprehensive benchmark suite for video generative models	The paper presents VBench, a benchmark suite for video generative models. It dissects video quality, has human - aligned metrics, and aims to drive video generation research.	Core Foundational Competencies	2024	CVPR/ICCV/ECCV	link
AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models	The paper establishes a benchmark for MLLM - based emotion understanding with MER - Caption, AffectGPT and MER - UniBench, releasing code and data.	Higher-Order Intelligence	2025	Arxiv	link
Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech	The paper presents Dynamic - SUPERB, a speech benchmark for zero - shot multi - task models. It encourages community contribution and proposes baseline approaches.	Higher-Order Intelligence	2024	ICASSP	link
AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension	The paper introduces AIR - Bench, the first benchmark for evaluating LALMs' audio understanding and interaction. It reveals limitations and guides future research.	Higher-Order Intelligence	2024	ACL	link
MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models	The paper introduces MuChoMusic, a benchmark for evaluating music understanding in audio - focused multimodal LMs, identifying pitfalls in model integration and open - sourcing data/code.	Higher-Order Intelligence	2024	ISMIR	link
Animal-bench: Benchmarking multimodal video models for animal-centric video understanding	This paper establishes Animal - Bench, an animal - centric benchmark with defined task systems and data pipelines to evaluate multimodal models, releasing data and code.	Core Foundational Competencies	2025	NeurIPS	link
ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation	The paper proposes ChronoMagic - Bench to evaluate T2V models' temporal & metamorphic abilities, introduces metrics, and creates ChronoMagic - Pro dataset.	Core Foundational Competencies	2024	NeurIPS	link
AudioTime: A Temporally-aligned Audio-text Benchmark Dataset	This paper proposes a temporally - aligned audio - text dataset AudioTime, with rich temporal annotations, and provides test set and metric for text - to - audio models.	Higher-Order Intelligence	2025	ICASSP	link
MME-Unify: A Comprehensive Benchmark for Unified Multimodal Understanding and Generation Models	The paper presents MME - Unify, a comprehensive benchmark for U - MLLMs, with standardized tasks, novel ones, and model benchmarking, to address evaluation gaps.	Core Foundational Competencies	2025	Arxiv	link

Continued on next page

Table 1 – Continued

Title	TLDR	Category	Year	Venue	Link
V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning	The paper introduces V - STaR benchmark, decomposes video understanding into RSTR task, constructs dataset to evaluate Video - LLMs' spatio - temporal reasoning.	Higher-Order Intelligence	2025	Arxiv	link