LARGE MULTIMODAL MODELS EVALUATION: A SURVEY

Zicheng Zhang[*1], **Junying Wang**[*1,3], **Farong Wen**[*1,2], **Yijin Guo**[*1,2], **Xiangyu Zhao**[1,2], **Xinyu Fang**[1,4], **Shengyuan Ding**[1,3], **Ziheng Jia**[1,2], **Jiahao Xiao**[1], **Ye Shen**[1,2], **Yushuo Zheng**[1,2], **Xiaorong Zhu**[1,2], **Yalun Wu**[2], **Ziheng Jiao**[19], **Wei Sun**[17], **Zijian Chen**[1,2], **Kaiwei Zhang**[1,2], **Kang Fu**[2], **Yuqin Cao**[2], **Ming Hu**[18], **Yue Zhou**[17] **Xuemei Zhou**[8], **Juntai Cao**[9], **Wei Zhou**[10], **Jinyu Cao**[11], **Ronghui Li**[12], **Donghao Zhou**[15], **Yuan Tian**[1], **Xiangyang Zhu**[1], **Chunyi Li**[1,2,7], **Haoning Wu**[7], **Xiaohong Liu**[2], **Junjun He**[1], **Yu Zhou**[14], **Hui Liu**[14], **Lin Zhang**[14], **Zesheng Wang**[16], **Huiyu Duan**[2], **Yingjie Zhou**[2,6], **Xiongkuo Min**[2], **Qi Jia**[1], **Dongzhan Zhou**[1], **Wenlong Zhang**[1], **Jiezhang Cao**[5], **Xue Yang**[2], **Junzhi Yu**[13], **Songyang Zhang**[1], **Haodong Duan**[1], **Guangtao Zhai**[†1,2]

[1]*Shanghai AI Laboratory,* [2]*Shanghai Jiao Tong University,* [3]*Fudan University,*
[4]*Zhejiang University,* [5]*Harvard University,* [6]*PengCheng Laboratory,*
[7]*Nanyang Technological University,* [8]*Delft University of Technology,*
[9]*University of British Columbia,* [10]*Cardiff University,*
[11]*University of California, Berkeley,* [12]*Tsinghua University,*
[13]*Peking University,* [14]*China University of Mining and Technology,*
[15]*The Chinese University of Hong Kong,* [16]*Nantes Université,*
[17]*East China Normal University,* [18]*Monash University,* [19]*Huawei Technologies Co., Ltd.*

Project Page: `https://github.com/aiben-ch/LMM-Evaluation-Survey`
AIBench Homepage: `https://aiben.ch/`

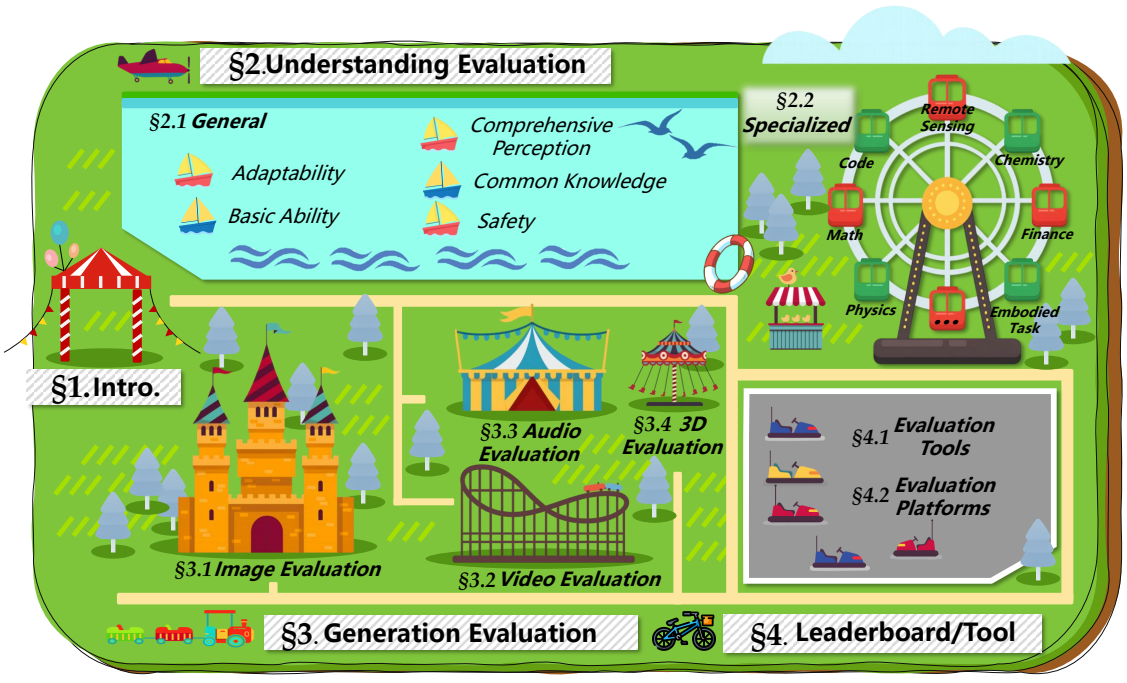

Figure 1: Content overview of the survey.

ABSTRACT

As Large Multimodal Models (LMMs) advance rapidly across diverse multimodal understanding and generation tasks, the need for systematic and reliable evaluation frameworks becomes increasingly critical. To address this need, this survey provides a structured overview of LMM evaluation, centered around two main axes: multimodal evaluation for understanding and generation. 1) For understanding, a dual-perspective framework is introduced to distinguish benchmarks between general capabilities, which emphasize common tasks, and specialized capabilities, which reflect expert-level competence in domain-specific fields. 2) For generation, evaluation is organized by output modality, including image, video, audio, and 3D content. 3) From a community perspective, this survey further highlights authoritative leaderboards and foundational tools that have been instrumental in establishing a comprehensive evaluation ecosystem for LMMs. By unifying general-specialized understanding and modality-specific generation evaluations, this survey clarifies the current landscape and provides guidance for future research in the LMM evaluation field.

1 INTRODUCTION

The emergence of Large Multimodal Models (LMMs), capable of processing and generating content across multiple modalities such as text, image, audio, video, and 3D models, mark a significant milestone in the development of artificial intelligence. These models (such as GPT [253; 252], Gemini [79], Grok [368]) unify vision, language, or other modalities under a shared framework, enabling a wide range of applications from multimodal understanding (image captioning [11], visual question answering [360], etc.) to multimodal generation (text-to-image/video generation [421], text-to-video [289], etc.) Therefore, LMM evaluation can be categorized into two pillars: **understanding evaluation**, which measures the ability of model to comprehend and reason over multimodal inputs, and **generation evaluation**, which assesses the quality of multimodal outputs conditioned on instructions.

In the domain of multimodal understanding, early efforts primarily targeted **general capabilities** [386], emphasizing versatility across modalities, tasks, and domains. These capabilities include common tasks like instruction-following, dialog comprehension, general visual grounding, etc. However, as real-world demands increase, there is a growing emphasis on **specialized capabilities** [445], which assess expert-level understanding in vertical domains such as medicine, law, or science. This evolution reflects a shift in focus from broad generalization to deep, domain-specific competence. On the other hand, multimodal generation involves producing content in **various output modalities** (including images, videos, audio, and 3D assets) based on user prompts. The evaluation of generative abilities presents unique challenges due to the subjectivity and modality-dependence of quality criteria [449]. Moreover, generation quality must be assessed with respect to both visual quality (e.g., techinical quality, aesthetics, realism) and alignment with user intent, often requiring modality-specific metrics or human evaluations.

Accordingly, the differing targets, formats, and metrics of multimodal understanding and generation tasks have led **to the emergence of their evaluations as two initially distinct paradigms**. Understanding evaluation is like a quiz: models tackle constrained tasks with clear correctness criteria, focusing on accuracy and reasoning. In contrast, generation evaluation resembles submitting a portfolio: models produce open-ended outputs such as images, videos, or 3D content assessed by assessed along multiple dimensions, including fidelity, relevance, and perceptual quality. Though objectives and metrics differ, the development of LMMs are bringing these paradigms closer. As shown in Figure 2, generation evaluation often relies on strong understanding for instruction following or reward modeling, while understanding can be inferred from generative output quality in unified tasks. This mutual influence suggests that understanding and generation evaluations are integrating, and this survey includes both to reflect this convergence.

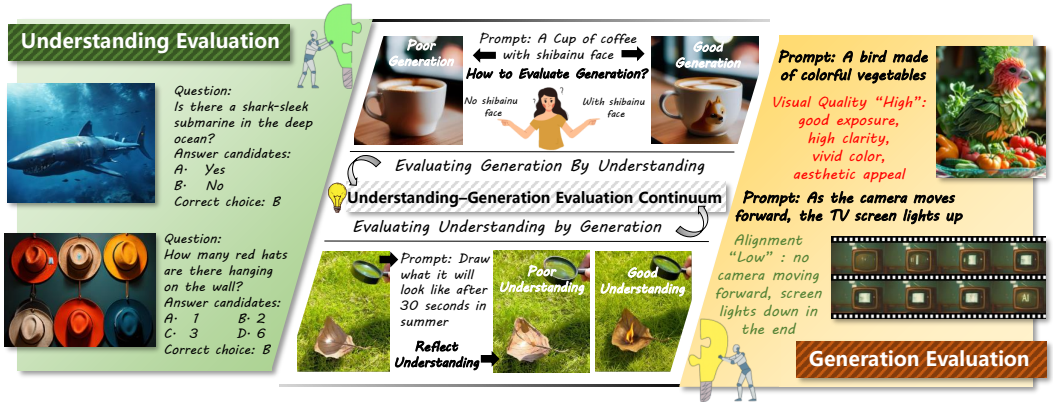

Figure 2: Understanding–Generation Evaluation Continuum. Understanding evaluation focuses on assessing LMM performance usually through question-answering accuracy, while generation evaluation emphasizes the quality of generated content. A growing trend reveals the convergence of these paradigms, where understanding can facilitate generation evaluation, and generation serves as a proxy for evaluating understanding.

In parallel with the development of benchmark and quality assessment, the broader ecosystem for LMM evaluation has been supported by the the rapid emergence of community-driven leaderboards and open-source evaluation tools [254; 445; 427; 459], which have evolved into central platforms for aggregating evaluation results, tracking performance across models and promoting reproducibility. Meanwhile, these evaluation tools help provide standardized interfaces for multimodal input-output evaluation, empowering researchers and developers to efficiently assess both general and modality-specific capabilities. These infrastructures not only facilitate transparent comparison among models but also accelerate the process of iteration and benchmarking at scale.

Despite rapid progress, existing surveys often focus narrowly on either model architecture or modality-specific tasks, lacking a unified view of how LMMs are evaluated across both understanding and generation dimensions[71; 170; 104; 246; 418]. Particularly, no comprehensive survey has addressed the evaluation of LMMs from a dual perspective of general versus specialized understanding, or from a modality-specific perspective in generation. To address this gap, this survey presents a comprehensive review of LMM evaluation frameworks along three axes: 1) understanding evaluation, structured by general and specialized capabilities, 2) generation evaluation, organized by output modality, and 3) community evaluation infrastructure, including leaderboards and tools. By unifying these perspectives, this survey aims to clarify the current landscape, identify emerging trends, and provide actionable guidance for future evaluation research in the era of foundation multimodal models. The content overview is shown in Fig. 1.

2 Evaluation for Understanding

Understanding evaluation measures the ability of LMMs to comprehend, interpret, and reason over multimodal inputs, forming the foundation for both downstream application and reliable generation. We categorize understanding evaluation into **general** and **specialized** capabilities. General evaluation focuses on versatility across tasks, domains, and modalities, highlighting adaptability, broad perceptual coverage, foundational skills, general knowledge, and safety. These capabilities reflect the model's capacity to operate robustly in diverse, non-expert scenarios and serve as prerequisites for high performance in specialized domains. In contrast, specialized evaluation targets expert-level competence within vertical fields where domain-specific reasoning and terminology are critical. [1]

[1]The 'VQA' in Sec. 2 and Sec. 3 refers to visual question answering and video quality assessment, respectively.

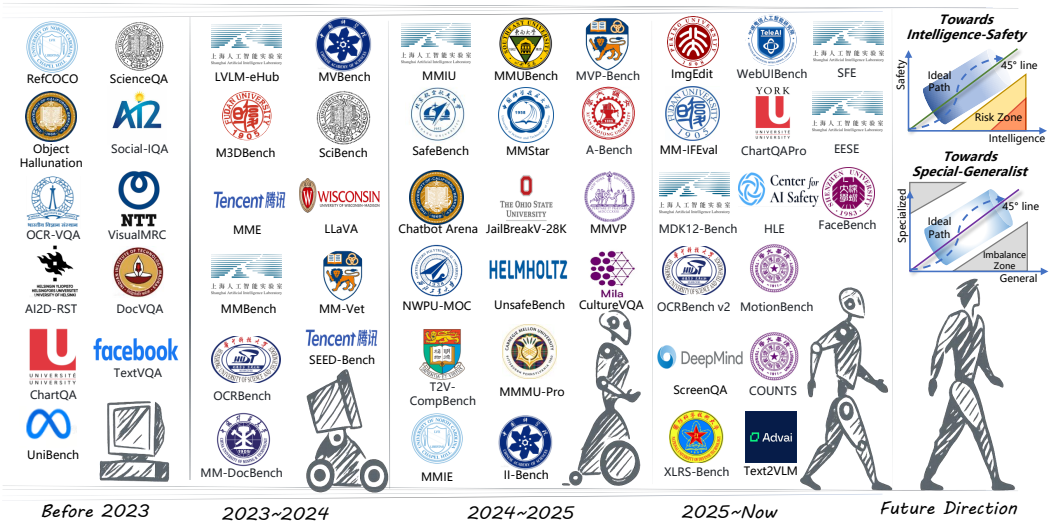

Figure 3: Representative benchmarks for understanding evaluation. As the field evolves, benchmark development is increasingly emphasizing the synchronized advancement of intelligence and safety, as well as comprehensive assessment spanning specialized and generalist capabilities.

2.1 General

General capability evaluation emphasizes the versatility of LMMs across tasks, domains, and modalities, reflecting its ability to adapt to diverse instructions, perceive varied multimodal inputs, and apply core reasoning skills in non-expert contexts. This category typically covers five dimensions: **adaptability** to different task formats, **basic abilities** as prerequisites for downstream tasks, **comprehensive perception** across various scenarios, **general knowledge** for multiple disciplines, and **safety** in handling tasks responsibly and robustly. These dimensions collectively form the basis for robust, transferable understanding performance and serve as prerequisites for specialized domain competence.

2.1.1 Adaptability

Adaptability indicates the ability of LMMs to generalize across heterogeneous task formats and varied interaction patterns, ranging from following fine-grained instructions to handling multi-turn conversations, multi-image reasoning, and interleaved multimodal contexts [30; 434]. This dimension is intended to highlight whether models are capable of robustly accommodating diverse input–output structures.

1) Instruction Following. Instruction following is the most direct form of adaptability, as it evaluates whether models can accurately align outputs with prompts of varying complexity and specificity [206]. LLaVA-Bench [192] represents an early effort, consisting of two subsets: LLaVA-Bench (COCO) with 90 GPT-4-generated tasks across 30 COCO images, and LLaVA-Bench (In-the-Wild) which introduces 60 tasks from 24 open-domain images. Together, these probe zero-shot generalization in both controlled and naturalistic settings. MIA-Bench [260] advances this line by requiring strict adherence to layered, structured instructions. Its 400 image–prompt pairs demand precisely formatted responses, enabling fine-grained diagnosis of whether models comply with compositional requirements. MM-IFEval [54] raises the bar further by integrating multimodal instructions with an average of five constraints per task across 32 categories, combining rule-based verification with model-judged assessment to ensure precision. Similarly, VisIT-Bench [14] spans 592 queries across 70 instruction families, ranging from descriptive recognition to creative generation. Its human-authored captions and instruction-conditioned outputs enable evaluation via both human judgment and LLM-based automatic scoring.

2) Multi-turn Dialogue. Adaptability is further reflected in a model's ability to sustain coherent, multi-round exchanges while integrating multimodal information. MMDU [213] is a representative benchmark, featuring dialogues of up to 27 turns grounded in 20 images, with conversations extending to 18k tokens. Its paired MMDU-45k dataset provides instruction-tuning material, bridging gaps between synthetic training and real-world conversations. ConvBench [200] takes a cognitive perspective, organizing 577 dialogues into a three-level hierarchy of perception, reasoning, and creativity. This pyramid structure allows precise attribution of model failures to specific cognitive layers, supporting detailed error analysis. SIMMC 2.0 [142] embeds conversations in photo-realistic shopping environments. It defines four standardized tasks, including dialog state tracking and multimodal response generation, and introduces human-paraphrased utterances for naturalism. Together, these benchmarks emphasize challenges in long-context tracking, multimodal grounding, and maintaining consistent conversational flow.

3) Multi-image Reasoning. Another central test of adaptability lies in reasoning across multiple images, where models must integrate information across distinct visual contexts. Mementos [339] targets sequential reasoning by introducing 4.7K dynamic image sequences, assessing whether models capture temporal changes and behavioral dynamics beyond static perception. MuirBench [310] expands coverage to 12 tasks and 10 relational types with over 11K images, pairing each instance with unanswerable variants to diagnose robustness. MMIU [245], the largest benchmark in this line, contains 77K images and 11K questions spanning 52 tasks, enabling comprehensive evaluation of cross-image perception and reasoning. MIRB [450] pioneers relational reasoning benchmarks across four dimensions, requiring comparative analysis of up to 42 images in a task. Subsequent efforts such as MIBench [193] and II-Bench [212] emphasize higher-order perception and knowledge-seeking, while Mantis-Eval [126] bridges single- and multi-image reasoning in one unified benchmark. MileBench [291] highlights performance degradation under long input sequences, and ReMI [134] stresses adaptability across domains including math, physics, and code, pushing beyond purely visual reasoning.

4) Interleaved Data. Adaptability also reflects a model's ability to process and generate interleaved sequences of text and images, reflecting real-world multimodal communication. Codis [220] evaluates context-dependent comprehension by pairing images with contradictory textual cues, requiring models to dynamically reinterpret visual content. Sparkles [112] introduces a word-level interleaved dialogue setting, comprising SparklesDialogue (machine-generated data), SparklesEval (GPT-assisted metrics), and SparklesChat (baseline models), thereby probing fine-grained integration across multiple images and texts. At larger scale, MMIE [370] offers 20K queries across 12 fields, unifying open-ended and multiple-choice tasks, while InterleavedBench [197] supports arbitrary multimodal input–output orders and introduces InterleavedEval, a reference-free metric for text quality, perceptual fidelity, and cross-modal coherence. OpenING [463] adds 5.4K human-annotated instances across 56 real-world tasks, focusing on interleaved image–text generation in creative scenarios such as travel, design, and brainstorming.

5) Human-centric Evaluation. Finally, adaptability extends to human-centeric scenarios, where benchmarks test whether models align with subjective judgments, values, and preferences. HumaniBench [267] and HERM-Bench [168] foreground fairness, inclusivity, and ethical reasoning in multimodal settings. UNIAA [474] and HumanBeauty [187] shift focus to aesthetics, drawing on large-scale human ratings to quantify visual appeal. Social-IQA [278] emphasizes social commonsense reasoning, requiring inference of motives, emotions, and event outcomes, while EmpathicStories [284] examines empathy through multimodal narratives with explicit emotional cues. In addition to task-specific datasets, community platforms provide direct human feedback. Chatbot Arena [39] implements double-blind pairwise comparison, where users vote on dialogue quality without model identity, thus capturing preference at scale. OpenAssistant Conversations [141] crowdsources human feedback from global volunteers in multi-turn dialogues, incorporating quality ratings and preference rankings. HCE [88] uses structured questionnaires to capture subjective assessments of problem-solving, information quality, and interaction experience.

Taken together, adaptability benchmarks illustrate that general-purpose multimodal models must cope with diverse input structures, ranging from strictly constrained instructions to dynamic dialogues, multi-image reasoning, interleaved modalities, and subjective human-centeric interactions. While these efforts demonstrate breadth and creativity in task design, open challenges remain in scaling to dynamic, real-world environments where modalities, instructions, and human values interact fluidly.

2.1.2 BASIC ABILITY

In this section, we categorize the fundamental capabilities of LMMs into three core types: **recognition**, **perception**, and **reasoning**. Recognition focuses on extracting structured or semi-structured information from visual inputs. Perception emphasizes understanding visual content at varying levels of granularity. Reasoning targets the capacity to perform higher-level inference based on visual and multimodal cues.

1) Recognition. Recognition encompasses the ability of LMMs to identify and extract structured or semi-structured information from visual inputs, covering object enumeration, text reading, document layout parsing, interface element localization, and chart/table interpretation. In contrast to general perception, these tasks demand precise grounding of discrete elements (e.g., objects, characters, cells, widgets) together with structure-aware parsing serving as prerequisites for robust reasoning.

1-A) COUNTING. Counting tasks measure the ability to accurately enumerate visual entities under varying conditions, often requiring resilience to distributional shifts or compositional complexity. NWPU-MOC [76] focuses on multi-category object counting in aerial images, offering 3.4K scenes across 14 fine-grained categories in both RGB and near-infrared modalities. Moving beyond static enumeration, T2V-CompBench [294] extends counting into the video domain, assessing compositional text-to-video generation abilities such as attribute binding and spatio-temporal consistency, while ConceptMIX [365] automates compositional evaluation for text-to-image models via generated prompts and VLM-based verification. PICD [456] adds a different dimension by evaluating recognition of photographic composition, comprising 36.8K images in 24 composition categories, and introducing a composition discrimination accuracy metric.

1-B) OCR. OCR-oriented evaluation has evolved from isolated text recognition to reasoning over text-rich scenes and videos. Early large-scale efforts such as TextVQA [290] frame recognition as question answering over images requiring correct interpretation of embedded text, and OCR-VQA [248] scales this paradigm to 207.5K book-cover images with over one million QA pairs. Consolidated suites like OCRBench [210] integrate 29 datasets covering text recognition, scene-text VQA, document VQA, key information extraction, and handwritten mathematical expression recognition, while OCRBench v2 [74] quadruples task diversity (31 scenarios) and adds 10K human-verified QA pairs, introducing challenges such as text localization and logical reasoning. ASCIIEval [122] extends the text recognition to arts formulated in text strings. Recent benchmarks focus on reasoning traces and modality interaction: OCR-Reasoning [107] annotates both answers and reasoning processes across six abilities, and M4-ViteVQA [452] extends evaluation to video contexts, testing spatio-temporal grounding. SEED-Bench-2-Plus [154] further broadens the scope to text-rich comprehension in charts, maps, and web pages.

1-C) DOCUMENT UNDERSTANDING. Document understanding probes the parsing of complex layouts, long contexts, and heterogeneous multimodal content [65]. MMDocBench [476] adopts OCR-free fine-grained tasks (15 in total, 4.3K QA pairs with 11.3K supporting regions) to test perception and reasoning without OCR pipeline shortcuts. MMLongBench-Doc [229] emphasizes multi-page reasoning over 135 lengthy PDFs (1K expert questions), with 33.7% cross-page queries and 20.6% hallucination-detection items. UDA [115] targets in-the-wild document complexity with 2.9K real-world documents and 29.5K expert Q&A pairs, including raw HTML/PDF tables. Layout-aware QA is also addressed by VisualMRC [297] (10K+ images and 30K+ QA) and DocVQA [243] (50K ques-

tions and 12K images). Scaling up, DocGenome [371] annotates 500K scientific documents with structured multimodal data, while GDI-Bench [174] decouples visual and reasoning complexity for nuanced difficulty analysis.

1-D) WEB/GUI UNDERSTANDING. Web and GUI understanding tests whether models can perceive interactive elements and follow instruction-grounded interactions in digital interfaces. AitW [266] offers 715K device-control episodes with 30K instructions, supporting multi-step gesture inference from demonstrations and language. ScreenSpot [34] evaluates GUI element grounding across mobile, desktop, and web platforms, while VisualWebBench [195] covers seven tasks with 1.5K curated instances from 139 websites (87 sub-domains). GUI-World [24] introduces dynamic and sequential content—conditions that often degrade LMM performance. Extending to programmatic competence, WebUIBench [190] assesses end-to-end capabilities for web application development (21K QA from 0.7K real sites) across perception, code generation, and HTML understanding. To evaluate current models' ability in understanding the content of mobile app screens, ScreenQA [100] constructs approximately 86K question–answer pairs over mobile app screenshots, enriched with short and full-sentence answers, UI element annotations and bounding box coordinates. It establishes a standard benchmark dedicated to advancing research in the field of screen content understanding.

1-E) CHARTS & TABLES UNDERSTANDING. Interpreting charts and tables requires mapping visual encodings and layouts to semantic and numerical meaning. ChartQA [241] establishes a large-scale foundation with human-written and generated questions targeting logical and arithmetic reasoning, while ChartQAPro [242] diversifies chart types (dashboards, infographics) and introduces unanswerable queries for robustness testing. Table-specific benchmarks include ComTQA [453] (about 9K QA) and TableVQA-Bench [139] (1.5K QA from generated tables), which reveal that visual processing remains more challenging than text-only inputs. CharXiv [343] (2.3K scientific charts) and SciFIBench [272] (2K figure questions) provide high-difficulty scientific contexts, while AI2D-RST [96] augments diagram QA with multi-layer discourse structure annotations. InfoChartQA [188] (5.6K infographic/plain chart pairs) and EvoChart-QA [108] (650 charts and 1.2K expert questions) expose robustness gaps under non-canonical designs. WikiMixQA [67] extends to cross-modal reasoning over tables and charts, combining 1K multiple-choice questions from 4K Wikipedia pages. ChartX [372] provides a comprehensive benchmark spanning diverse chart types and reasoning tasks. It enables systematic evaluation of models' skills in visual recognition, data extraction, and structured reasoning.

2) Perception. Perception refers to the ability of LMMs to extract and interpret visual information at varying levels of granularity, ranging from low-level sensory features to high-level semantic attributes. These benchmarks are primarily designed to focus on assessing raw perceptual capacity without requiring complex, multi-step inference, though they often form prerequisites for reasoning tasks.

2-A) LOW-LEVEL PERCEPTION. Low-level perception benchmarks evaluate models' sensitivity to fundamental visual properties such as color, texture, sharpness, and distortions. Q-Bench [352] provides a unified framework for testing low-level visual perception, description, and assessment abilities using both single-image and paired-comparison formats, integrating LLVisionQA+ (2.9K images + 1.9K pairs), LLDescribe+ (499 images + 450 pairs), and seven image quality assessment datasets. A-Bench [447] further pushes forward the low-level perception for LMMs on AIGC images. MVP-Bench [161] extends this to both low- and high-level perception tasks, incorporating synthetic distortions and natural images to evaluate object recognition and behavioral understanding. These resources highlight that while LMMs can generate plausible descriptions, their fine-grained visual sensitivity remains limited compared to specialized vision models.

2-B) HIGH-RESOLUTION PERCEPTION. High-resolution perception measures the ability to process and utilize detailed visual cues present in large-scale images. XLRS-Bench [312] targets ultra-high-resolution remote sensing scenarios, defining 16 sub-tasks across 10 per-

ceptual and six reasoning capabilities, with the largest average image size to date. HR-Bench [334] is the first benchmark to systematically evaluate 4K and 8K image understanding, demonstrating the performance loss caused by downsampling and exploring modality complementation with text. MME-RealWorld [435] pushes realism further, offering 29K+ manually annotated QA pairs over diverse high-resolution, real-world scenes. V* Bench [359] complements these by focusing on crowded, detail-rich images, emphasizing the need for visual search mechanisms in multimodal systems.

2-C) HIGHER-ORDER PERCEPTION. Higher-order perception extends beyond literal recognition to encompass aesthetic judgment, emotional understanding, and comprehension of abstract or implicit attributes. FaceBench [336] addresses comprehensive face perception, cataloging over 210 attributes and nearly 50K QA pairs. MMAFFBen [211] is the first multilingual, multimodal benchmark for affective analysis, covering sentiment polarity, intensity, and emotion classification across text, image, and video in 35 languages. FABA-Bench [178] jointly evaluates recognition and generation for fine-grained facial affective behaviors such as action unit recognition. Emotion-oriented datasets like MEMO-Bench [469], EmoBench [275], and EEmo-Bench [77] emphasize progressively fine-grained sentiment assessment. For aesthetic evaluation, AesBench [111] and UNIAA-Bench [475] offer structured frameworks across perception, empathy, assessment, and interpretation, while ImplicitAVE [479] and II-Bench [212] target implicit and abstract attribute extraction. CogBench [292] assesses comprehensive dimensions including time, location, character, event, mental state, etc., through images with rich reasoning chains. A4Bench [320] is the first comprehensive benchmark designed to assess the affordance perception capabilities of LMMs. Covering both constitutive and transformative affordance, it reveals critical challenges for LMMs in grasping contextual and dynamic affordance.

2-D) FINE-GRAINED PERCEPTION. Fine-grained perception examines models' self-awareness and ability to detect subtle or systematic visual patterns. MM-SAP [341] introduces a knowledge quadrant framework to delineate what a model knows versus does not know, spanning three sub-datasets for different self-awareness levels. Cambrian-1 [303] provides a vision-centric evaluation over 15 visual representations, alongside CV-Bench (2.6K VQA questions) for fine-grained capability diagnosis. MMUBench [165] pioneers machine unlearning evaluation for LMMs, measuring forgetting efficacy, generality, specificity, fluency, and diversity. MMVP [304] reveals CLIP-blind pairs, where semantically distinct images are misperceived as similar, exposing systematic perceptual biases. To address the lack of standardized evaluation for multi-image quality comparison, MICBench [356] provides a diverse set of open-ended and multi-choice tasks that enable comprehensive, fine-grained assessment of the comparative perception capabilities of LMMs.

2-E) VISUAL GROUNDING. Visual grounding benchmarks assess the alignment between natural language expressions and their corresponding visual referents at object or part level. The RefCOCO family [404; 236] forms the classic REC suite, with RefCOCO restricting to short, interactive expressions, RefCOCO+ removing spatial terms to force reliance on visual cues, and RefCOCOg introducing longer, descriptive expressions. Ref-L4 [26] updates these with higher annotation accuracy, a larger vocabulary, and longer expressions (avg. 24.2 words). MRES-32M [335] expands to multi-granularity segmentation with over 32.2M masks, while UrBench [460] broadens grounding to complex multi-view urban scenarios with 11.6K questions across 14 task types. COUNTS [163] explicitly targets out-of-distribution generalization for object detectors and LMMs, introducing 14 natural distributional shifts with object-level annotations. MTVQA [299] is a multilingual benchmark for text-centric visual question answering, covering 9 languages and resolving the visual–textual misalignment limitations of prior machine-translated datasets. It establishes a robust standard for assessing and advancing multilingual scene-text understanding.

3) Reasoning. Reasoning evaluates the ability of LMMs to integrate perceptual cues with logical, spatial, comparative, and sequential inference to derive conclusions beyond direct recognition. These tasks often require models to combine multiple sources of information, maintain intermediate representations, and apply knowledge in context-sensitive ways.

3-A) RELATIONAL REASONING. Relational reasoning focuses on understanding spatial configurations, geometric relations, and comparative attributes among visual entities. GePBench [376] pioneers large-scale geometric perception evaluation with 80K figures and 285K multiple-choice questions, highlighting foundational gaps in shape and structure comprehension. SpatialMQA [194] introduces 5.3K human-annotated samples over COCO2017 to test spatial relation understanding, while SpatialRGPT-Bench [33] extends the challenge to 3D cognition using indoor, outdoor, and simulated environments with precise ground-truth annotations. CoSpace [478] probes continuous space perception from sequences of spatially consistent images, complementing static-scene evaluations. LMM-CompBench [137] addresses comparative reasoning by assembling 40K image pairs for relative judgments across eight dimensions. SOK-Bench [308] integrates situated and open-world knowledge for video reasoning, while GSR-Bench [264] and What'sUp [132] target specific spatial relations such as object positioning and orientation. Q-Spatial Bench [186] addresses the challenge of quantitative spatial reasoning in vision–language models by providing a human-annotated benchmark covering diverse object size and distance estimation tasks, and, together with the proposed zero-shot spatial-prompt strategy that elicits reasoning via reference objects, enables substantial performance improvements without additional training. AS-V2 [333] is a circular-based relation probing evaluation benchmark that advances models' capabilities in relation understanding, scene graph generation, and relation grounding, while effectively mitigating bias in relational comprehension.

3-B) MULTI-STEP REASONING. Multi-step reasoning benchmarks assess a model's ability to perform sequential inference, often requiring intermediate steps and multi-modal integration. Visual CoT [282] exemplifies this trend by annotating 98K questions with explicit reasoning steps and bounding boxes over key visual regions, enabling interpretability analysis. LogicVista [375] evaluates five logical reasoning tasks spanning nine capabilities using 448 multiple-choice questions, while VisuLogic [384] mitigates language shortcuts by designing 1K human-verified problems focusing on genuine vision-centric inference. CoMT [37] expands the scope to visual creation, deletion, update, and selection—four categories that simulate complex visual operations. PUZZLES [62], while originating in reinforcement learning, offers 40 adjustable-size logic puzzles to test algorithmic and generalization capabilities in structured problem-solving.

3-C) REFLECTIVE REASONING. Reflective reasoning involves self-assessment, error correction, and targeted knowledge editing in multimodal contexts. LOVA3 [451] equips models with the capacity to pose and evaluate questions in visual settings, with its EvalQABench containing 64K training and 5K testing samples. VLKEB [103] builds on multi-modal knowledge graphs to assess knowledge editing portability—whether changes apply consistently across relevant content. MMKE-Bench [56] offers 2.9K knowledge items and 8.3K images for evaluating visual entity, semantic, and user-specific editing. Fine-grained correction is addressed by MC-MKE [426; 164], which decomposes multimodal knowledge into visual and textual components to isolate and fix misreadings or misrecognitions. NegVQA [438] contributes 7.3K negated binary-choice questions, revealing substantial performance drops when models must process logical negation.

2.1.3 COMPREHENSIVE PERCEPTION

Comprehensive perception benchmarks aim to evaluate the breadth of multimodal understanding by systematically covering different sensory modalities. Unlike adaptability, which stresses the ability to handle diverse task formats, comprehensive perception emphasizes whether large multimodal models can capture a wide range of perceptual and cognitive skills across images, videos, audio, and 3D content.

1) Image Perception. A number of benchmarks target holistic evaluation of vision-language understanding through diverse image-based tasks. LVLM-eHub [382] offers a large-scale suite spanning six categories, from visual question answering to embodied AI, while its lightweight variant TinyLVLM-eHub [283] condenses evaluation into 2.1K image–text pairs for quick testing. Other large-scale frameworks expand coverage:

LAMM [399] bridges 2D and 3D with 12 tasks, MME [69] incorporates 14 subtasks from object recognition to code understanding, and MMBench [208] stabilizes grading by converting free-form responses into multiple-choice format. The SEED-Bench series [156; 155; 154] scales this further across multimodal QA and text-rich scenarios, while MMT-Bench [400] and LMMs-Eval [153; 427] integrate dozens of tasks across modalities for unified comparison. To address potential biases in such broad suites, MMStar [27] curates 1.5K vision-indispensable samples and further introduces evaluation metrics like multi-modal gain and leakage, while NaturalBench [152] seeks to minimizes language priors by constructing 10K adversarial human-verified samples. Finally, MM-Vet [406] and ChEF [286] move beyond aggregate scores by decomposing vision-language integration and calibration, providing finer diagnostic insights into model behavior.

2) Video Perception. Video benchmarks extend comprehensive perception into the temporal domain, testing whether models can integrate motion, temporal order, and multimodal signals. Video-MME [70] provides one of the earliest large-scale benchmarks, with 900 videos spanning 254 hours and 2.7K QA pairs, exposing deficiencies in temporal fusion. MMBench-Video [64] complements this with long-form YouTube videos and free-form QA, introducing ability-based categorization and GPT-4-based scoring. MVBench [169] emphasizes the evaluation of short-to-medium video reasoning, while LongVideoBench [351] pushes the scale to hour-long content with a novel referring reasoning task that requires models to locate and analyze specific temporal contexts. LVBench [332] further broadens this scope to extreme-length videos lasting several hours, thereby highlighting persistent weaknesses in long-term memory. MotionBench [97], in contrast, narrows the focus to fine-grained motion understanding, revealing that even state-of-the-art LMMs continue to struggle with dynamic perception.

3) Audio Perception. Audio-centric benchmarks evaluate whether models can capture speech, paralinguistic, and environmental sounds in a unified framework. AudioBench [309] aggregates 26 datasets, including seven newly compiled corpora, across eight task categories from speech recognition to acoustic scene understanding, and employs open-source model-as-judge protocols to reveal instruction-following gaps in AudioLLMs. AIR-Bench [391] scales this effort with 19K multiple-choice and 2K open-ended questions covering speech, environmental sound, and music, enabling both foundational and higher-level auditory evaluation. Dynamic-SUPERB [408] introduces a dynamic benchmark for multi-task and zero-shot generalization, spanning 33 speech tasks and 22 datasets, and supports continuous expansion through community contributions—offering a scalable platform for developing universal speech understanding systems.

4) 3D Perception. Comprehensive perception also extends into 3D understanding, which requires integrating visual, textual, and geometric information. M3DBench [171] introduces over 320K multimodal instruction–response pairs spanning text, images, and 3D data, making it the first large-scale foundation for 3D instruction tuning. In specialized domains, M3D [10] covers eight tasks in 3D medical imaging, enabling systematic evaluation of multimodal medical reasoning. Space3D-Bench [296] targets spatial reasoning with 3D question–answering tasks, offering rigorous tests of geometric and spatial cognition in multimodal settings. Together, these benchmarks push beyond 2D perception to evaluate whether models can adapt to spatial complexity and specialized modalities.

In summary, comprehensive perception benchmarks broaden evaluation from images to videos, audio, and 3D, offering systematic tests of LMMs' general perceptual and cognitive breadth. While progress has been made in building unified and large-scale suites, challenges remain in ensuring fairness, balancing modality coverage, and scaling to real-world multimodal data with long contexts and fine-grained signals. These benchmarks collectively provide indispensable baselines for diagnosing gaps in multimodal perception and guiding the development of more versatile LMMs.

2.1.4 General Knowledge

General knowledge benchmarks refer to evaluations covering broad, non-specialized subject areas, typically spanning multiple academic disciplines and assessing foundational to advanced reasoning abilities. They are designed to measure models' capacity for cross-domain understanding, from basic factual recall to complex problem-solving, without being limited to a single specialized field.

1) Primary Benchmarks. Early benchmarks primarily focus on science level below high school. ScienceQA [217] is a benchmark with 21k multimodal multiple-choice science questions from three different subjects, including annotations of answers with lectures and explanations, designed to explore language models' multi-hop reasoning by generating such content as chain of thought (CoT). CMMU [94]is a benchmark for Chinese multimodal and multi-type question comprehension and reasoning, consisting of 3.6K questions across seven subjects, in the form of multiple-choice, multiple-response, and fill-in-the-blank questions.

2) College-level Benchmarks. To systematically examine the reasoning capabilities required for solving complex scientific problems, several college-level benchmarks have emerged. Scibench [337] is an expansive benchmark suite, featuring a curated dataset of problems from mathematics, chemistry, and physics, with in-depth benchmarking of representative LLMs. EXAMS-V [50] is a challenging multi-discipline, multimodal, multilingual exam benchmark for evaluating vision-language models, with 20.9K multiple-choice questions across 20 school disciplines, 11 languages, diverse multimodal features, curated from global school exams requiring cross-language reasoning and text-visual joint reasoning. MMMU [413] is a benchmark evaluating multimodal models through 11.5K questions from six core fields (Art & Design, Business, etc.) across 30 subjects and 183 subfields with 30 diverse image types like charts and diagrams. MMMU-Pro [414] rigorously assesses the real-world understanding and reasoning capabilities of multimodal models through a three-step process based on MMMU that tests the model's fundamental cognitive skills of integrating visual and textual information.

3) Expert-level Benchmarks. As state-of-the-art multimodal large language models advance rapidly, many of them now achieve strong performance on these benchmarks above. This has spurred the creation of more expert-level benchmarks that feature high-difficulty questions. HLE [257] is a multimodal benchmark at the human knowledge frontier, designed as the final closed-ended academic benchmark with broad coverage, comprising 2.5K questions across dozens of subjects (e.g.mathematics, humanities and natural sciences) from global experts, including auto-gradable multiple-choice/short-answer questions with unambiguous, verifiable, non-internet-retrievable solutions. CURIE [45] is a scientific long-Context Understanding, Reasoning and Information Extraction benchmark designed to measure LLMs' potential in scientific problem-solving and assisting scientists, featuring ten challenging tasks with 580 expert-curated problem-solution pairs across six disciplines (materials science, condensed matter physics, quantum computing, geospatial analysis, biodiversity, and proteins) covering experimental and theoretical workflows. SFE [473] is a benchmark designed to evaluate LMMs' scientific cognitive capacities through three levels (scientific signal perception, attribute understanding, comparative reasoning), comprising 830 expert-verified VQA pairs across three question types and 66 multimodal tasks in five high-value disciplines. MMIE [370] is a large-scale, knowledge-intensive benchmark with reliable automated evaluation metrics to evaluate the interleaved multimodal understanding and generation capabilities of large vision-language models, containing 20K multimodal queries from multiple domains, supporting interleaved inputs and outputs and various question formats, which extends beyond basic perception by requiring models to engage in complex reasoning, leveraging subject-specific knowledge across different modalities.

To keep pace with evolving model capabilities and mitigate contamination, researchers have turned to dynamic evaluations. MDK12-Bench [464] is a multi-disciplinary benchmark assessing LMMs' multimodal reasoning via 140K real-world K-12 exam instances

across six disciplines, with diverse difficulty levels, 6.8K knowledge annotations, answer explanations, and a dynamic framework to mitigate data contamination. EESE [322] is a dynamic benchmark designed to reliably assess foundation models' scientific capabilities, consisting of a non-public 100K+ instance EESE-Pool (across 5 disciplines and 500+ subfields) and a periodically updated 500-instance subset, enabling leakage-resilient, low-overhead evaluations that effectively differentiate 32 models' scientific strengths and weaknesses, providing a robust, scalable, forward-compatible solution for science benchmarking.

2.1.5 SAFETY

The rapid advancement of LMMs has introduced unprecedented capabilities in both understanding and generating multimodal content. However, this progress has also raised significant safety concerns, necessitating systematic evaluation frameworks to assess potential risks and vulnerabilities [295; 422; 205; 395; 321; 398]. In this subsection, we provide a structured review of safety evaluation methodologies for LMMs, organizing existing efforts into four primary domains: jailbreak and adversarial robustness, comprehensive safety evaluations, hallucination and truthfulness, and fairness and bias. We also highlight emerging safety concerns such as privacy, deepfakes, and extremist content.

1) Jailbreak and Adversarial Robustness. One of the most pressing safety challenges for LMMs lies in their vulnerability to jailbreak attacks. By carefully crafting textual or multimodal prompts, adversaries circumvent guardrails and elicit unsafe, harmful, or disallowed content. This vulnerability raises concerns when deploying LMMs in open-world applications, where malicious users may deliberately attempt to bypass safety filters. Early benchmarks such as Unicorn [306] provide the first systematic evaluation of jailbreak risks in vision–language models. Building on this, JailbreakV-28K [223] introduces a large-scale dataset of 28K jailbreak samples covering diverse attack vectors, establishing a more comprehensive baseline for measuring susceptibility. MM-SafetyBench [204] extends the paradigm by showing how carefully selected images serve as enablers for jailbreak attacks, exploiting the multimodal nature of these systems. In parallel, Wang *et al.* [330] trace the landscape evolution from traditional text-only jailbreak strategies to multimodal ones, highlighting new risks introduced by visual prompts.

Recent works emphasize increasingly sophisticated attack strategies. AVIBench [425] investigates adversarial visual instructions, where subtle changes in images can successfully trigger unsafe responses. MMJ-Bench [348] provides a unified platform for evaluating both jailbreak attacks and defense mechanisms. Guo *et al.* [87] introduce the concept of the 'LMM safety paradox', observing that some models exhibit contradictory behaviors of being simultaneously more attackable yet also more easily defended. Empirical studies confirm that even cutting-edge systems such as GPT-4o remain at risk under these advanced jailbreak scenarios [402]. Collectively, this line of research shows that adversarial robustness is an evolving arms race, demanding continuous evaluation updates.

2) Comprehensive Safety Evaluations. Beyond individual attacks, researchers propose comprehensive benchmarks that integrate multiple safety dimensions, including harmful content generation, bias, robustness, and privacy. These frameworks provide a broader view of model safety in realistic settings. USB (Unified Safety Benchmark) [457] exemplifies this trend by consolidating diverse safety evaluation tasks into a single benchmark suite. Similarly, MLLMGuard [81] offers a bilingual multimodal evaluation suite, addressing the need for safety evaluation in multi-lingual contexts. Benchmarks such as SafeBench [401] and MM-SafetyBench [203] further extend the scope by including both text and image modalities in a unified testing pipeline.

At the same time, more targeted resources address specific real-world challenges. For instance, MemeSafetyBench [150] focuses on the unique risks of internet culture, curating over 50K memes to evaluate how LMMs handle socially charged multimodal content. UnsafeBench [263] evaluates unsafe image classification on both human-created and AI-generated images, recognizing the increasing role of synthetic data in safety evaluation.

These efforts underscore the importance of multi-dimensional, domain-specific evaluations that go beyond simple attack scenarios.

3) Hallucination and Truthfulness. Another central safety concern for LMMs is hallucination, where models produce outputs that are factually incorrect, ungrounded, or nonsensical. Unlike jailbreaks, hallucinations are not necessarily induced by adversaries, but emerge naturally in model responses, making them especially problematic in high-stakes applications such as education, healthcare, or law. Early studies such as Rohrbach *et al.* [274] analyze object hallucination in image captioning, laying the foundation for systematic evaluation. Subsequent work, including POPE [179] and M-HalDetect [83], introduce frameworks for detecting and mitigating hallucinations. More recent contributions expand diagnostic coverage: Hal-Eval [125] develops a fine-grained benchmark, Hallu-pi [53] examines robustness under perturbed inputs, and BEAF [396] introduces before–after comparison to measure response stability. Large-scale benchmarks such as HallusionBench [82] and AutoHallusion [366] emphasize scalability by automatically generating hallucination test cases.

Beyond hallucination detection, researchers have also explored trustworthiness in a broader sense. MultiTrust [436] benchmarks multiple aspects of model trustworthiness, while MMDT [379] investigates safety at the decoding level. Dataset adaptation methods such as Text2VLM [55] extend text-only safety evaluation resources to multimodal models, and MOSSBench [176] highlights oversensitivity issues, where models block benign queries due to over-aggressive safety mechanisms. Together, these studies emphasize that evaluating truthfulness requires both fine-grained diagnostic tools and holistic trustworthiness frameworks.

4) Fairness and Bias. Bias and fairness in LMMs have become key safety issues, especially as they are deployed in more sensitive domains. Unlike accuracy or robustness, fairness directly impacts social equity and can amplify existing stereotypes. Janghorbani and De Melo [117] propose one of the first comprehensive multimodal bias evaluation frameworks, expanding beyond traditional axes of gender and race. To address cultural diversity, CulturalVQA [250] curates over 2.3K image–question pairs from 11 countries across 5 continents, testing the cultural understanding of LMMs. ModScan [129] provides a structured approach to measure stereotype bias jointly across vision and language modalities.

In healthcare, where fairness is especially critical, FMBench [358] and FairMedFM [131] provide domain-specific benchmarks for evaluating fairness in medical LMMs. Beyond evaluation, fairness-aware training approaches such as FairCLIP [226] illustrate how fairness considerations can be integrated directly into model training. Collectively, these efforts highlight that fairness evaluation must be both context-sensitive and modality-aware.

5) Emerging Safety Concerns. Beyond well-studied domains such as jailbreaks, hallucinations, and bias, new forms of safety risks are emerging with the deployment of LMMs in the wild. These require specialized benchmarks and methodologies.

Privacy leakage is a growing concern: DoxingBench [224] demonstrates how multimodal agentic models may inadvertently reveal user locations from images, while PrivQA [28] investigates the ability of models to follow privacy-preserving instructions. In the context of misinformation and manipulation, SHIELD [285] evaluates forgery and face-spoofing detection, addressing the challenges of deepfakes. Meanwhile, ExtremeAIGC [22] benchmarks the vulnerability of LMMs to extremist AI-generated content, reflecting broader societal concerns about radicalization risks. These emerging benchmarks demonstrate that LMM safety must adapt dynamically to evolving threats.

2.2 SPECIALIZED

Specialized capability evaluation focuses on assessing the expert-level competence of LMMs in vertical domains, where domain-specific knowledge, reasoning paradigms, and modality combinations differ significantly from general-purpose tasks. Benchmarks in this

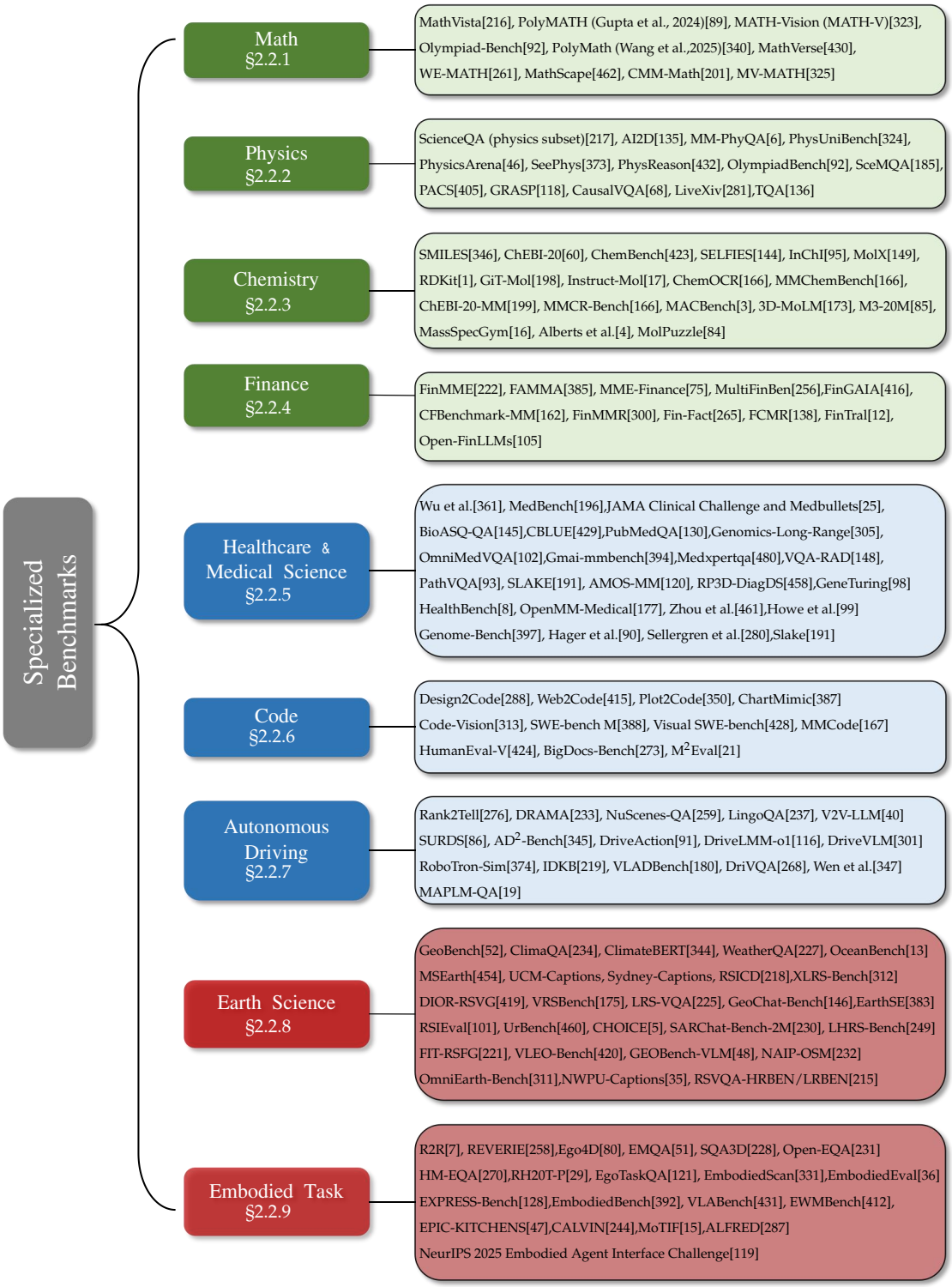

Figure 4: Quick references to the representative specialized benchmarks.

category are often constructed from professional datasets, competition problems, or real-world application scenarios, and thus present higher difficulty, stricter evaluation criteria, and richer context dependencies. Given the diversity of specialized domains, we organize the discussion by field, where the quick reference is presented in Fig. 4.

2.2.1 MATH

Mathematical multimodal benchmarks evaluate the ability of LMMs to couple visual understanding with formal reasoning. Early efforts focus on broad-coverage multi-source collections. For example, MathVista [216] integrates challenging problems from 31 multimodal datasets and further introduces curated subsets such as IQTest, FunctionQA, and PaperQA. PolyMATH (Gupta et al., 2024) [89] aggregates visually rich math problems across ten conceptual categories, including geometry, pattern recognition, and spatial or relational reasoning, providing a large-scale baseline for visual–logical integration.

Subsequent benchmarks raise the difficulty through competition-grade and discipline-diverse settings. MATH-Vision (MATH-V) [323] collects problems from real math competitions, spanning 16 mathematical disciplines and five difficulty levels to emphasize authenticity and fine-grained control. Olympiad-Bench [92] extends to Olympiad-level bilingual problems in mathematics and physics, targeting scientific reasoning beyond pure math. PolyMath (Wang et al., 2025) [340] is a multilingual math reasoning benchmark covering 18 languages and four difficulty levels, enabling cross-lingual comparisons (note the naming distinction from PolyMATH above).

More recent work emphasizes problem transformation and process-level evaluation. MathVerse [430] collects diagram-based problems and systematically transforms each into six versions, enabling modality ablation and robustness analysis. WE-MATH [261] decomposes composite problems into sub-problems based on knowledge concepts and proposes a four-dimensional diagnostic metric (Insufficient Knowledge, Inadequate Generalization, Complete Mastery, Rote Memorization) to hierarchically analyze reasoning failures. MathScape [462] focuses on photo-based scenarios that require connecting visual scenes to quantitative reasoning, assessing both theoretical understanding and application ability.

Further, some benchmarks target specialized multimodal settings. CMM-Math [201] is a Chinese multimodal benchmark spanning 12 grade levels, aligned with curriculum standards and common item types. MV-MATH [325] introduces multi-image problems (images interleaved with text) to test the synthesis of multi-visual evidence in mathematical reasoning. Mathematical multimodal benchmarks evolve from broad, multi-source collections toward competition-grade and multilingual challenges, raising both scope and difficulty. More recent efforts emphasize process-level diagnostics and specialized settings, marking a shift from general coverage to fine-grained, context-specific evaluation.

2.2.2 PHYSICS

In the domain of physics education and scientific reasoning, existing evaluation benchmarks primarily assess model performance through multimodal question answering tasks that combine textual descriptions with diagrams, images, or videos. Early multimodal physics benchmarks, such as ScienceQA [217] (physics subset), TQA [136], and AI2D [135], focus on interpreting textbook-style science questions with supporting diagrams or images. While these datasets feature carefully constructed question-answer pairs, their focus on K-12 or middle school content neglects higher-level reasoning complexity. Their scope is limited to straightforward conceptual recall and basic diagram interpretation, with minimal coverage of multi-step or problem-solving workflows.

Subsequent benchmarks introduce richer physics content and more challenging reasoning tasks. MM-PhyQA [6] targets high school physics problems with multi-image reasoning and chain-of-thought prompting, while PhysUniBench [324] advances to undergraduate-level problems across eight sub-disciplines, providing both multiple-choice and open-ended formats. PhysicsArena [46] further innovates by structuring its evaluation into three distinct stages: variable identification, process formulation, and solution derivation. This structure offers a closer simulation of authentic problem-solving processes. SeePhys [373]

spans middle school to PhD qualifying exams, featuring 21 types of diagrams and emphasizing vision-essential questions (75%) that require precise visual parsing for correct answers.

Benchmarks like PhysReason [432], OlympiadBench [92], and SceMQA [185] draw from physics competitions and entrance examinations, substantially increasing difficulty and diversity. PhysReason contains 81% diagram-based problems and provides theorem annotations with step-by-step derivations to evaluate joint visual–textual reasoning. OlympiadBench integrates Olympiad-level visual physics problems alongside mathematics, and SceMQA provides multi-science evaluation with a strong physics subset. These benchmarks begin to push models beyond rote formula application toward multi-step, domain-specific reasoning.

Beyond static diagrams, several benchmarks integrate temporal and multimodal sensory information. PACS [405] evaluates physical commonsense reasoning through audiovisual videos, while GRASP [118] uses simulation-based videos to test intuitive physics reasoning about object permanence and dynamics. CausalVQA [68] focuses on video-based multiple-choice questions requiring causal and physical reasoning. LiveXiv [281] represents a new direction—constructing visual question answering tasks from figures and charts in academic papers, including those from physics domains, thus bridging everyday reasoning benchmarks and specialized scientific literature comprehension.

Overall, these benchmarks impose progressively higher demands on LMMs, from integrating static and dynamic visual information to reasoning over complex diagrams and scientific figures. While advanced datasets like PhysicsArena, SeePhys, and PhysReason approach the rigor of high-level examinations, much of the field still target student-level difficulty, advancing toward expert-level content, authentic research-derived problems, and standardized multi-stage scoring remain a key future direction.

2.2.3 CHEMISTRY

Chemical information manifests across multiple modalities, which can be broadly grouped into 1D sequence, 2D structural, 3D spatial, and spectral representations. Each captures different aspects of molecular and chemical knowledge, forming the basis for probing LMMs in integrating symbolic, visual, as well as spatial cues with chemical reasoning.

1D representations encode molecules as strings or symbolic sequences. The most prevalent is SMILES (Simplified Molecular-Input Line-Entry System) [346], which shares a sequential format with natural language but follows unique chemical grammar. Benchmarks such as ChEBI-20 [60] and ChemBench [423] test sequence-sequence translation (e.g., SMILES $\leftrightarrow$ IUPAC names) and extend to reaction-centric prediction and cross-modal retrieval. Variants like SELFIES [144] and InChI [95] offer alternative encodings, while models such as MolX [149] employ specialized encoders to capture structural patterns.

2D representations include molecular graphs and rendered images, typically generated with RDKit [1]. Graph-based inputs are aligned with text via projectors, as in GiT-Mol [198] and Instruct-Mol [17], improving property prediction and molecular captioning through explicit spatial–topological encoding. Image-based molecule recognition is evaluated by ChemOCR [166], which spans styles from hand-drawn to scanned and photographed depictions. Multimodal benchmarks such as MMChemBench [166] and ChEBI-20-MM [199] extend earlier datasets with captioning and property prediction tasks. Beyond molecules, chemical vision benchmarks like MMCR-Bench [166] and MACBench [3] target diagram reasoning, scientific figure interpretation, and practical laboratory knowledge.

3D representations capture stereochemistry and spatial configurations critical for chemical function. 3D-MoLM [173] integrates a 3D molecular encoder into a LLM for structure-grounded QA and retrieval, while M3-20M [85] provides over 20 million molecules annotated with SMILES, 2D graphs, 3D coordinates, properties, and descriptions for large-scale pretraining.

Spectral data including mass spectrometry (MS), nuclear magnetic resonance (NMR), and infrared (IR) spectroscopy constitute another essential modality, requiring pattern recogni-

tion and cross-modal reasoning for structure elucidation. MassSpecGym [16] standardizes MS/MS-based evaluation for de novo generation, retrieval, and simulation. Alberts et al. [4] extend to multi-spectral datasets (IR, MS, NMR) for 790k molecules, enabling cross-spectra modeling. MolPuzzle [84] frames structure elucidation as sequential reasoning over IR, MS, and NMR, with tasks in molecule understanding, spectrum interpretation, and structure construction, revealing persistent performance gaps with expert chemists.

From an evolutionary perspective, chemical multimodal benchmarks have progressed from symbolic translation toward integrating structural, visual, and spectral modalities, with increasing emphasis on spatial realism, laboratory context, and diagnostic evaluation. Nevertheless, the complex semantics and modality diversity of this domain still pose significant challenges for unified modeling.

2.2.4 FINANCE

Financial multimodal benchmarks are designed to assess how well models can understand and reason over domain-specific information that combines unstructured financial text with structured and semi-structured visuals, such as tables, charts, and report figures. Compared to general-purpose multimodal benchmarks, they emphasize high-stakes, precision-critical tasks (ranging from interpreting market trends in graphs to extracting insights from earnings reports) where errors can carry significant real-world costs. Early resources establish fundamental QA and reasoning settings grounded in financial documents and visualizations. FinMME [222] provides a large-scale, high-quality dataset for chart- and table-based reasoning from financial reports, while FAMMA [385] introduces multilingual QA with textbook-derived and expert-authored questions in both basic and live professional settings. MME-Finance [75] targets bilingual, expert-level VQA with unique graphical content, revealing that top-performing general models struggle with specialized financial semantics and notation.

Later efforts broaden scope and complexity. MultiFinBen [256] spans multilingual, multimodal (text, vision, audio), and difficulty-aware evaluation, covering tasks from entry-level QA to expert financial reasoning. CFBenchmark-MM [162] focuses on Chinese multimodal QA over diverse visual formats, and FinMMR [300] emphasizes bilingual numerical reasoning across 14 financial subdomains. Fin-Fact [265] pivots to multimodal fact-checking, pairing claims with textual and visual evidence, while FCMR [138] advances to cross-modal multi-hop reasoning that requires integrating textual reports, tables, and charts for multi-step inference.

In parallel, several works introduce domain-specific models and agent-level evaluations. FinTral [12] presents a family of Mistral-based LMMs trained with financial data, outperforming GPT-4 in multiple benchmarks. Open-FinLLMs [105] offers open-source financial LMMs evaluated across 14 tasks, 30 datasets, and 4 multimodal settings, highlighting gains from targeted pre-training. FinGAIA [416] shifts toward end-to-end assessment of AI agents in finance, with 407 tasks across seven sub-domains and three difficulty levels. Financial multimodal benchmarks have evolved from static, document-centric QA to multilingual, difficulty-aware, and multi-hop reasoning tasks, and now toward agent-level workflows that simulate realistic analytical pipelines. Unlike general benchmarks, they demand capabilities far beyond the norm: precise numerical computation, robust interpretation of domain-specific charts and tables, and sensitivity to financial conventions and regulatory contexts. Current LMMs remain far from the accuracy, reliability, and interpretability required for professional use in these areas.

2.2.5 HEALTHCARE & MEDICAL SCIENCE

In the field of healthcare and medical sciences, existing evaluation benchmarks primarily assess AI model performance across different clinical scenarios through text question answering [361; 196; 25; 145; 429; 130] and visual question answering [102; 394; 480; 148; 191; 93] tasks. Early VQA benchmarks, such as VQA-RAD [148], PathVQA [93], and SLAKE [191], primarily focus on interpreting individual radiological or pathological images. While these feature carefully designed question-answer pairs, they lack complete

clinical context and have limited task complexity. Benchmarks like AMOS-MM [120], RP3D-DiagDS [458], and PubMedQA [130], although focus on diagnostic or domain-specific question-answering tasks, typically employ relatively simple modalities with insufficient fidelity to real clinical scenarios.

The new generation of comprehensive medical AI benchmarks has achieved significant breakthroughs in evaluation methodologies. HealthBench [8] is constructed with participation from 262 physicians across 60 countries, comprising 5K samples covering multiple dimensions including health dialogues, medical task requests, and medical record summaries, and employing customized evaluation criteria to replace traditional metrics. Large-scale datasets such as GMAI-MMBench [394] and OpenMM-Medical [177] contain vast collections of samples, ranging from tens to hundreds of thousands. This scale supports robust multilingual and multimodal evaluation. These benchmarks focus not only on technical performance but also emphasize clinical relevance, practicality, and safety. Evaluation metrics have been expanded from single accuracy measures to multidimensional assessments including factuality, completeness, and potential harm, reflecting the inevitable trend of medical AI transitioning from laboratory research to clinical application. Future developments will place greater emphasis on simulating real clinical scenarios, cross-modal information fusion, multilingual support, and ethical safety evaluation to advance the maturation and trustworthy clinical deployment of medical AI technologies.

In specialized medical and genomics fields, relevant datasets reflect the developmental needs of precision medicine [461; 99]. Regulatory sequence benchmarks such as Genomics-Long-Range [305] test models' abilities to identify motifs, predict chromatin states, and maintain attention across kilobase contexts. Due to highly imbalanced data, these benchmarks typically employ metrics such as AUROC or MCC to penalize false positives. In genomic knowledge retrieval, GeneTuring [98] evaluates large language models' ability to recall variant nomenclature, gene functions, and pathway contexts without hallucination through compact yet diverse question-answering modules. Genome-Bench [397] goes further by requiring models to perform multi-step reasoning in discussions related to CRISPR, thereby exposing current models' deficiencies in chain-of-thought depth.

These benchmarks impose higher requirements on models: the need to integrate real clinical information and generate accurate responses through complex reasoning. Existing benchmarks typically adopt multiple-choice questions or open-ended question formats, with evaluation metrics including accuracy, BLEU, ROUGE, and F1 scores. While multiple-choice questions facilitate precise evaluation, this format cannot authentically reflect clinical environments—in actual clinical practice, questions are often complex, may have multiple reasonable solutions, and sometimes lack known standard answers [8]. Although open-ended questions provide greater flexibility, they lack robust and reliable evaluation metrics to assess factual correctness and clinical appropriateness [90]. The difficulty of most benchmarks has not yet reached the level of professional clinicians. Currently, only a few datasets such as MedXpertQA and HealthBench approach the rigor of practicing physician examinations, while other benchmarks typically only test the level of medical students or junior physicians. Despite these limitations, these benchmarks continue to drive model improvements in diagnostic reasoning, image interpretation, and medical knowledge coherence [280]. For further development, evaluation datasets should be extracted from authentic hospital records, curated internet medical content, and academic case reports, establishing reliable gold standards through expert review and high inter-annotator agreement.

2.2.6 CODE

Multimodal code generation benchmarks evaluate the ability of models to generate executable and structurally correct code from diverse visual inputs, including user interface screenshots, algorithmic diagrams, scientific plots, and document layouts. Early efforts mainly focus on user interface rendering, such as Design2Code [288], which pairs 484 real webpage screenshots with HTML/CSS and evaluates output fidelity by rendering-based visual similarity and human judgment. Web2Code [415] further scales this paradigm with a large webpage understanding benchmark and a code generation benchmark using the

same screenshots, introducing GPT-4V–based visual comparison between generated and reference pages.

In addition to benchmarks for web UI, a separate series focuses on chart-to-code generation. Plot2Code [350] supports direct and conditional settings for producing Python/R plotting code from chart images, with evaluation combining code pass rate, text-match ratio, and GPT-4V image similarity. ChartMimic [387] expands to 4.8K curated (chart, instruction, code) triplets, defining both direct reproduction and customized modification tasks. These benchmarks assess the ability of a model to capture fine-grained visual details and translate them into precise data visualization code.

Other benchmarks center on diagram/algorithm-based programming tasks. HumanEval-V [424] extends the original HumanEval coding challenges with diagrams such as flowcharts, circuits, and UI layouts, measuring visual reasoning in pass@k terms on hidden tests. Code-Vision [313] builds on this by introducing flowchart- and math-based visual problem specifications, encompassing both basic and complex algorithmic logic.

At a larger scale, several benchmarks embed visual reasoning into realistic software engineering workflows. SWE-bench [388] extends repository-level issue solving to 619 multimodal tasks across 17 JavaScript repositories, where problem descriptions include screenshots or diagrams. Visual SWE-bench [428] similarly evaluates bug fixing with visual context, while MMCode [167] collects 3.5K competitive programming problems with 6.6K images of graphs, geometric figures, circuits, and boards, where the visual elements are essential to deriving correct solutions. M^2Eval [21] further targets multilingual, multimodal code generation, featuring 300 problems in 10 programming languages with UML diagrams and flowcharts, judged by average nine-unit-test pass rates. Document-to-structured-code generation forms another stream, as in BigDocs-Bench [273], which includes tasks like Screenshot2HTML, Table2LaTeX, Image2SVG, and Image2Flow, requiring precise translation of complex visual or tabular inputs into executable structured formats.

Overall, these benchmarks reveal a progression from pixel-to-code tasks for UI rendering, to plot and diagram grounding, and finally to repository-level and document-based software development scenarios. Evaluation protocols mix execution correctness (unit tests, pass@k), render-based similarity, and visual–text alignment, yet persistent challenges remain in extracting fine-grained semantics from complex visuals and aligning them with maintainable, functionally accurate code.

2.2.7 Autonomous Driving

Autonomous driving–oriented multimodal benchmarks evaluate the capacity of large vision–language or multimodal models to perceive complex traffic environments, reason over dynamic spatial–temporal cues, and produce accurate, context-aware responses. Early research concentrates on scene-level visual question answering (VQA) and joint perception–reasoning tasks using dashcam or simulation data. Rank2Tell [276] introduces an ego-centric dataset for importance ranking of traffic objects with natural language justifications, while DRAMA [233] collects over 17K interactive driving scenarios for joint risk localization and captioning. NuScenes-QA [259] scales to 34K multi-sensor scenes and 460K QA pairs from camera and LiDAR, capturing the challenges of multi-frame, multi-modal street-view reasoning. LingoQA [237] offers 28K video scenarios with 419K annotations and proposes Lingo-Judge to improve automated evaluation of driving-related VQA.

As the field progresses, tasks expand to cooperative driving and map understanding. V2V-LLM [40] builds a vehicle-to-vehicle QA benchmark for fusing distributed perception, while MAPLM-QA [19] focuses on traffic and HD map comprehension for domain-specific model fine-tuning. More specialized benchmarks such as SURDS [86] target spatial reasoning categories like orientation and depth, demonstrating gains from reinforcement learning–based alignment, and AD^2-Bench [345] adds adverse weather and complex traffic conditions with fine-grained Chain-of-Thought annotations to diagnose reasoning gaps.

In parallel, attention has shifted toward end-to-end decision-making and planning. DriveAction [91] is the first action-driven benchmark for Vision–Language–Action mod-

els, linking perception to high-level driving actions via an action-rooted evaluation tree. DriveLMM-o1 [116] introduces step-by-step reasoning annotations for perception–prediction–planning QA, while DriveVLM [301] integrates VLMs into real-time driving pipelines. RoboTron-Sim [374] leverages simulated hard cases to boost real-world performance, and IDKB [219] evaluates models on over one million handbook, theory, and simulated test items to probe licensing-level knowledge. Complementary resources such as VLADBench [180] enable fine-grained capability breakdowns, and DriVQA [268] adds gaze-tracking to align model attention with human driving behavior.

Beyond dataset construction, some studies focus on probing the limits of existing large vision–language models in driving contexts without introducing new resources. For example, Wen et al. [347] evaluate GPT-4V's scene understanding, causal reasoning, and decision-making capabilities, highlighting persistent weaknesses in traffic light recognition and spatial reasoning.

Taken together, autonomous driving benchmarks have evolved from single-scene VQA toward multi-agent, spatial–temporal reasoning, and action-conditioned decision-making, often under adverse or cooperative settings. They demand that models integrate multimodal sensor inputs, interpret fine-grained spatial relations, and align with human-like risk assessment and planning—capabilities where even state-of-the-art LMMs still fall short of the robustness and reliability required for deployment in real-world traffic environments.

2.2.8 Earth Science

Earth science LMM benchmarks demonstrate a clear evolution in scope, modality, and task complexity, progressing from single-sphere textual QA to multi-sphere, multi-modal reasoning, particularly within the remote sensing (RS) ecosystem. For single-sphere QA, GeoBench [52] focuses on the lithosphere, offering over 2.5K multiple-choice and open-ended questions collected via a semi-automated pipeline from web and academic resources, targeting factual knowledge and scientific reasoning. In the atmospheric domain, ClimaQA [234], ClimateBERT [344], and WeatherQA [227] integrate textbook, scientific, and RS data to evaluate both comprehension and visual reasoning. For the hydrosphere, OceanBench [13] compiles 13K+ open-ended QA pairs from marine science literature, advancing free-text scientific reasoning.

Multi-sphere benchmarks further extend this coverage. OmniEarth-Bench [311] integrates heterogeneous databases and expert review to produce nearly 30K multiple-choice questions with rich metric annotations. MSEarth [454] emphasizes image-based VQA with chain-of-thought reasoning, featuring 11K+ expert-verified items across all spheres. EarthSE [383] similarly collects 10K mixed-format QA pairs from academic literature, prioritizing structured reasoning over diverse scientific contexts.

In RS, early paired image–text captioning datasets such as UCM-Captions, Sydney-Captions, and RSICD [218] provide hundreds to 10K labeled images, later expanded by NWPU-Captions [35] to 31K images for broader scene coverage. VQA-centric datasets include: RSVQA-HRBEN/LRBEN [215] scales to one million automatically generated QA pairs, and DIOR-RSVG [419] enables object grounding evaluation. More recent datasets diversify task coverage, with VRSBench [175] supporting captioning, VQA, and localization, LRS-VQA [225] targeting large-scale RS VQA, and GeoChat-Bench [146] extending to multimodal instruction following.

Recent benchmarks further raise standards of scale, annotation fidelity, and resolution. XLRS-Bench [312] offers ultra-high-resolution imagery (average 8,500×8,500 pixels) across 16 perception and reasoning tasks. RSIEval [101] provides 2.5K manually annotated descriptions for detail-oriented assessment. UrBench [460] integrates street-view and satellite imagery for urban perception tasks, while CHOICE [5] mitigates data leakage through independently collected imagery across 23 fine-grained tasks. SARChat-Bench-2M [230] standardizes Synthetic Aperture Radar evaluation with 2M samples covering six target categories.

Fine-grained RS understanding has also become a focus. LHRS-Bench [249] combines label filtering, balanced sampling, and GPT-4–generated instructions to evaluate recognition, spatial perception, and reasoning. FIT-RSFG [221] leverages high-resolution global RS imagery with scene graphs for relational understanding. VLEO-Bench [420] supports scene understanding, localization, and change detection, while GEOBench-VLM [48] spans disaster monitoring, crop classification, and marine debris detection. NAIP-OSM [232] further aligns high- and low-resolution satellite imagery for pixel- and image-level tasks, enriching foundational RS evaluation.

Overall, the trajectory of Earth science LMM benchmarks reflects a systematic shift from domain-specific textual QA toward large-scale, multi-sphere, multi-modal, and ultra-high-resolution RS benchmarks. This progression not only diversifies evaluation scenarios but also imposes increasingly stringent requirements for geospatial reasoning, cross-modal alignment, and fine-grained perception in foundation models.

2.2.9 EMBODIED TASK

The evaluation of embodied intelligence initially focuses on single tasks involving visual navigation and language-based question answering. Embodied Questioning Answering (EQA) [49] is the first to combine visual navigation with language question answering, assessing an agent's ability to actively explore and acquire information within 3D environments. Shortly thereafter, R2R (Room-to-Room) [7] expands the task to natural language navigation within real building environments, enhancing the realism and complexity of navigation scenarios, while Reverie [258] contributes further challenging navigation tasks in this domain. To address more complex environmental interactions, Alfred [287] designs multi-step, compositional household tasks, and Calvin [244] develops long-horizon, language-conditioned robotic manipulation tasks, advancing research on multimodal perception and complex action sequences. At the same time, large-scale first-person video datasets such as EPIC-KITCHENS [47] and Ego4D [80] enrich task content by including hand-object interactions, daily activities, social interactions, as well as long-term memory and future prediction, promoting the evaluation of temporal and semantic reasoning capabilities.

To deepen the study of spatial and temporal reasoning abilities, EMQA [51] and SQA3D [228] introduce multi-hop reasoning tasks across space and time, strengthening the understanding of 3D spatial relationships. Open-EQA [231] innovatively supports open-vocabulary and episodic memory-based question answering, combining real-world environments with human-authored questions and leveraging LLMs for evaluation. HM-EQA [270] explores the use of visual-language models' semantic knowledge and visual prompting to improve exploration efficiency, while applying confidence calibration to mitigate model memory limitations and confidence miscalibration. Further benchmarks such as MoTIF [15] and EgoTaskQA [121] focus on task execution and causal reasoning in GUI environments and first-person perspectives, enhancing diagnostics of temporal perception, spatial awareness, and causal inference. Later, larger and more complex benchmarks like EmbodiedScan [331] and RH20T-P [29] emerged, combining 3D object detection, environment grounding, and foundational robotic operations to better approximate real-world robotic applications. As a large-scale embodied question answering benchmark, EXPRESS-Bench [128] integrates exploration and reasoning behaviors, proposing hybrid navigation models and novel metrics to ensure alignment between exploratory behavior and answer accuracy.

Recently, embodied intelligence evaluation has advanced into a unified system covering multiple tasks, dimensions, and scenarios. EmbodiedEval [36] aggregates 328 tasks and 125 3D scenes, encompassing navigation, interaction, social engagement, and multidimensional question answering. EmbodiedBench [392] establishes 1.1K test tasks across four environments, assessing commonsense reasoning, complex instruction understanding, spatial cognition, and long-term planning, revealing shortcomings in fine-grained control by current state-of-the-art models. VLABench [431] focuses on general-purpose, language-conditioned manipulation tasks, emphasizing real-world knowledge, multi-step reasoning, and joint action-language comprehension, promoting the development of general em-

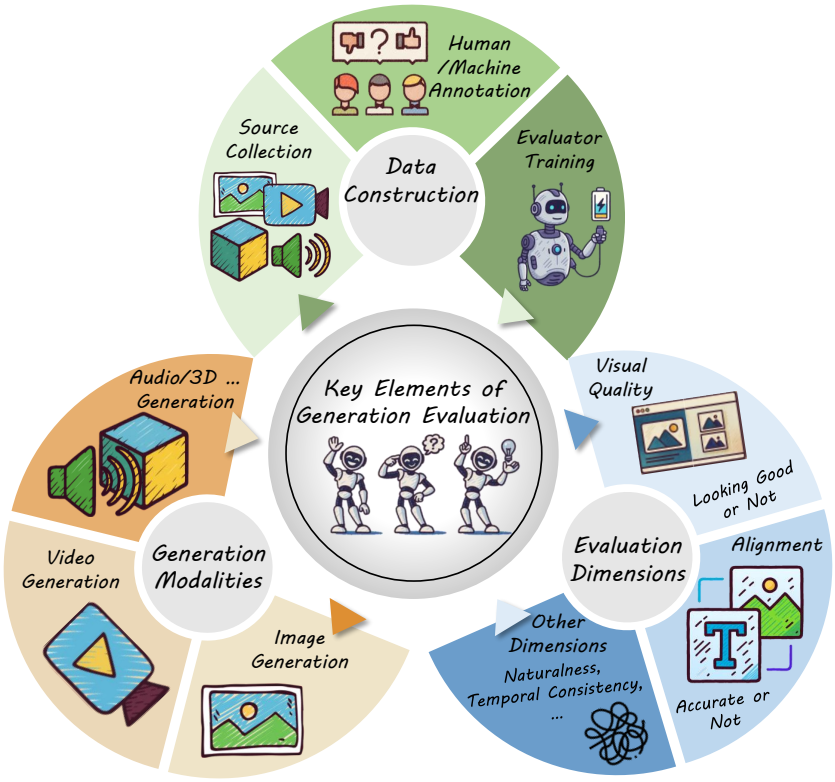

Figure 5: Key elements of generation evaluation for large multimodal models. The process involves three complementary aspects: 1) **Data Construction**, which covers source collection, human/machine annotation, and evaluator training; 2) **Generation Modalities**, spanning image, video, audio, and 3D content creation; and 3) **Evaluation Dimensions**, which assess visual quality (looking good or not), alignment with instructions (accurate or not), and other factors such as naturalness, temporal consistency, and coherence. Together, these elements provide a systematic framework for benchmarking multimodal generation.

bodied intelligence.Additionally, EWMBench [412] assesses generative planning abilities of embodied world models through scene consistency, action correctness, and semantic alignment. Meanwhile, the NeurIPS 2025 Embodied Agent Interface Challenge [119] advances the creation of unified benchmarking frameworks, facilitating standardized evaluation and reproducibility of large language models in embodied decision-making, driving embodied intelligence evaluation towards more open and systematic directions.

3 Evaluation for Generation

Generation evaluation focuses on assessing the quality, alignment, and diversity of multimodal outputs produced by LMMs in response to instructions. Key elements of this process are summarized in Figure 5. Unlike understanding evaluation, which typically measures accuracy on constrained tasks, generation evaluation must account for subjective and modality-dependent quality criteria, often requiring a combination of human judgment and specialized automatic metrics. Key challenges include balancing technical quality with semantic alignment, handling the open-ended nature of outputs, and ensuring fair, reproducible scoring across models.

Table 1: Brief comparison of quality assessment datasets for visual AIGC. **Scale** denotes the number of AIGC samples in the dataset, and **Ratings** denotes the number of subjective or objective quality annotations. Both are reported in **K** (thousand) or **M** (million) with one decimal place for consistency. "-" indicates that the information is not explicitly provided in the original paper and cannot be reliably inferred.

Dataset	Year	Scale	Ratings	Models	Quality Assessment Aspects
*Quality assessment for **AIGIs***					
DiffusionDB [342]	2022	14.0M	—	1	None
HPD [364]	2023	98.8K	98.8K	1	Human preference
ImageReward [380]	2023	136.9K	136.9K	3	Human preference
Pick-A-Pic [140]	2023	500.0K	500.0K	6	Human preference
AGIQA-1K [442]	2023	1.1K	23.7K	2	Overall perceptual quality
AGIQA-3K [160]	2023	3.0K	125.0K	6	Perceptual quality, text alignment
AIGCIQA2023 [316]	2023	2.4K	48.0K	6	Perceptual quality, authenticity, correspondence
AGIN [32]	2023	6.1K	181.0K	18	Overall naturalness
AIGIQA-20K [157]	2024	20.0K	420.0K	15	Perceptual quality, text alignment
AIGCOIQA2024 [390]	2024	0.3K	6.0K	5	Perceptual quality, comfortability, text alignment
CMC-Bench [159]	2024	58.0K	160.0K	6	Ultra-low bitrate compression quality
PKU-I2IQA [411]	2023	3.2K	96.0K	6	Perceptual quality in NR and FR settings
SeeTRUE [393]	2023	31.9K	31.9K	—	Text–image semantic alignment verification
AIGCIQA2023+ [317]	2025	2.4K	48.0K	6	Perceptual quality, authenticity, correspondence
Q-Eval-100K [441]	2025	100.0K	960.0K	23	Visual quality, long-text alignment (images/videos)
HPD v2 [363]	2023	433.8K	798.1K	—	Human preference on diverse AIGIs
*Quality assessment for **AIGVs***					
Chivileva *et al.* [41]	2023	1.0K	48.2K	5	Perceptual quality, text alignment
EvalCrafter [207]	2023	3.5K	8.6K	7	Perceptual quality, text alignment, temporal quality
FETV [209]	2023	2.5K	29.7K	3	Perceptual quality, text alignment, temporal quality
VBench [114]	2023	7.0K	—	4	Video quality, consistency
T2VQA-DB [143]	2024	10.0K	540.0K	9	Perceptual quality, text alignment
GAIA [31]	2024	9.2K	971.0K	18	Video action quality
AIGVQA-DB [318]	2025	36.6K	122.0K	15	Perceptual quality, temporal smoothness, dynamic degree, alignment
AIGVE-60K [315]	2025	58.5K	180.0K	30	Perceptual preference, text–video correspondence, task-specific accuracy
Human-AGVQA-DB [440]	2025	6.0K	630.0K	22	Human activity quality
TDVE-DB [319]	2025	3.9K	173.6K	12	Edited quality, editing alignment, structural consistency
AGAVQA-3K [20]	2025	3.1K	9.3K	8	Audio–visual quality, content consistency, overall quality
*Quality assessment for **AIGAs***					
Qwen-ALLD [23]	2025	20.0K	60.0K	—	MOS, SIM, A/B preference (speech quality)
BASE-TTS [147]	2024	—	—	—	Fine-grained semantic capture, emotion, prosody
ATT [338]	2025	—	—	—	Human-likeness, multi-dimensional TTS quality
TTSDS2 [247]	2025	0.3K	—	20	Prosody, intelligibility, multi-lingual TTS quality
*Quality assessment for **AIG3Ds***					
MATE-3D [439]	2024	1.3K	107.5K	8	Alignment, Geometry, Texture, Overall quality
3DGCQA [472]	2025	0.3K	12.5K	7	Alignment, Overall quality
AIGC-T23DAQA [73]	2025	1.0K	1.0K	6	Quality, Authenticity, Text–Content correspondence
SI23DCQA [72]	2025	1.5K	—	5	Overall, Color, Shape quality
3DGS-IEval-15K [377]	2025	15.2K	228.0K	6	Image quality for compressed 3D Gaussian Splatting

3.1 Image Generation Evaluation

This section reviews evaluation methodologies for image generation, which represents the most mature yet still rapidly evolving area of AIGC assessment. With the proliferation of text-to-image (T2I) systems and related pipelines, image generation evaluation has become a cornerstone of multimodal benchmarking, providing both large-scale corpora and fine-grained quality annotations to support human-aligned and model-based assessment.

3.1.1 Datasets for AI-Generated Image Quality Assessment

This section surveys datasets for generation-oriented evaluation of AI-generated images (AIGIs) from text-to-image (T2I) and related pipelines, emphasizing perceptual quality, semantic alignment, authenticity, and aesthetics. We consolidate prior categories into four groups: foundational corpora, human-aligned supervision, multi-dimensional perceptual & alignment datasets, and integrated benchmarks & task-oriented evaluation.

1) Foundational Corpora. The large-scale collections of prompt-image pairs are essential for downstream evaluation and analysis. DiffusionDB [342] is the first large-scale prompt–image corpus, comprising 14M AIGIs from 1.8M unique prompts with associated hyperparameters, enabling research on prompt engineering, model behavior, and misuse detection.

2) Human-Aligned Supervision (Preferences & Rewards). Pairwise preference annotations provide a scalable approach to capturing human judgments, while reward models translate these comparisons into scalar signals for automated evaluation. The Human Preference Dataset (HPD) [364] and HPD v2 [363] collect 98.8K and 798K human choices, respectively, serving as foundational resources for training HPS-style reward models. Pick-a-Pic [140] contains over 500K instances across 35K prompts, where each instance comprises a pair of generated images with a human preference label. This dataset facilitates the development of CLIP-based scoring models such as PickScore, which are aligned with human perceptual preferences. ImageReward [380] leverages 137K expert-curated comparisons from DiffusionDB to train a general-purpose reward model. The resulting signal shows strong correlation with human ratings and demonstrates superior performance over aesthetic-based and CLIP-based proxy metrics.

3) Multi-Dimensional Perceptual and Alignment Datasets. This category includes resources that go beyond a single mean opinion score by capturing multiple quality facets, such as naturalness, aesthetics, and text–image consistency. AGIQA-1K [442] contains 1K images annotated for technical quality, aesthetics, and text alignment. AGIQA-3K [160] expands this to 2.9K images, while AIGIQA-20K [157] scales up to 20K images from 15 text-to-image models with 420K human ratings. PKU-I2IQA [411] focuses on image-to-image generation under both no-reference and full-reference settings. For assessing naturalness, AGIN [32] includes 6K images labeled for overall naturalness, technical plausibility, and rationality, and proposes JOINT and JOINT++ evaluators that jointly model technical and semantic cues. For evaluating alignment, SeeTRUE [393] offers 31.8K labeled text–image pairs and introduces VQ^2 and VNLI metrics that outperform CLIP and BLIP on challenging alignment tasks. GenAI-Bench [189] comprises 1.6K prompts generating 8K images across six models, each rated on a five-point Likert scale, supporting quantitative analysis of text–image alignment quality.

4) Integrated Benchmarks & Task-Oriented Evaluation. Unified resources aim to bridge isolated criteria and demonstrate practical utility by progressively expanding coverage and integration. AIGCIQA2023 [316] takes the first step by assessing 2.4K AI-generated images from six text-to-image models across three core dimensions (quality, authenticity, and correspondence), while AIGCIQA2023+ [317] enriches these assessments with human preference scores and explanatory annotations, adding interpretability. Building on this foundation, Q-Eval-100K [441] dramatically scales both scope and modality, covering 100K image and video instances with 960K human ratings, and introduces Q-Eval-Score, a unified evaluator capable of generalizing across visual domains. Finally, moving beyond static evaluation, CMC-Bench [159] explores task-oriented cooperation, analyzing how image-to-text and text-to-image models perform under ultra-low bitrate compression, and revealing that some pairings can even outperform advanced visual codecs.

The field has evolved from foundational prompt–image galleries to human-aligned supervision, followed by multi-dimensional perceptual and alignment datasets, and finally to integrated benchmarks. This progression establishes standardized protocols for end-to-end generation evaluation and lays the groundwork for multi-modal, interactive, and sequential assessment scenarios.

3.1.2 Models for AI-Generated Image Quality Assessment

This section reviews representative approaches for evaluating perceptual quality and text–content alignment of AIGIs. We organize methods into distributional proxy metrics, alignment and preference modeling, LMM-driven evaluation with instruction-tuned assessors, specialized IQA architectures for AIGIs, and Reasoning-Driven Generation and Editing Evaluation.

1) Distributional Proxy Metrics. Inception Score (IS) [277] is an early proxy intended to reflect both quality and diversity, but it has been criticized for imprecision and weak correlation with human judgment in many scenarios. To better capture temporal dynamics, FVD [307] measures the distance between feature distributions of generated vs. real videos

using I3D embeddings. Lower FVD indicates more natural-looking videos, though its correlation still depends on content/domain.

2) Alignment and Preference Modeling. Moving beyond distributional surrogates, CLIP-based preference models such as HPS [364] and PickScore [140] learn to mimic human choices on AIGIs. To improve semantic robustness, BLIP-based ImageReward/ReFL [380] predicts human-aligned quality with richer language grounding. For explicit verification, VQ^2 reframes text–image alignment as answering verifiable questions about the prompt, while VNLI performs direct natural language inference on image–text pairs [393]. A complementary formulation, VQAScore [189], uses a VQA model to score the probability of a "Yes" answer to "Does this figure show {text}?". In practice, public challenges (e.g., NTIRE 2024 AIGC QA [202]) provide open testbeds and leaderboards that calibrate these alignment/quality assessors and track progress across methods.

3) LMM-driven Evaluation and Instruction-tuned Assessors. LMMs can directly emit quantitative quality judgments as well. Q-Bench [353] first enables LMMs to output calibrated scores via a softmax strategy evaluated on standard IQA sets. Building on this, Q-instruct, Q-align, Q-boost, and Co-Instruct [354; 355; 446; 357] introduce instruction/training procedures that align LMMs to IQA objectives and improve rating consistency. DepictQA [403] elicits fine-grained, language-based rationales for human-like judgments, while M3-AGIQA [44] conducts multi-round, multi-aspect prompting for holistic, human-aligned assessment across visual and textual facets. A unified evaluator, Q-Eval-Score [441], scores both perceptual quality and long-text alignment, improving robustness on complex prompts. Crucially, evaluation signals can close the loop: Q-Refine [158] uses quality-aware feedback to refine T2I generation, demonstrating how learned assessors guide generators toward higher perceptual fidelity and alignment.

4) Specialized IQA Architectures for AIGIs. Recent studies have proposed dedicated architectures to handle the unique artifacts and semantic requirements of AIGIs. MA-AGIQA [326] incorporates semantic guidance by injecting prompt cues and combining them through a mixture-of-experts framework, while SF-IQA [407] focuses on fusing quality and similarity features using a multi-layer extractor built on a fine-tuned vision–language backbone. SC-AGIQA [172] explicitly enforces text–visual semantic constraints to jointly evaluate alignment and perceptual distortion, and MINT-IQA [317] extends this direction by learning multi-perspective human preferences and leveraging instruction tuning to enhance explainability. TSP-MGS [369] further separates perceptual quality and alignment using task-specific prompts and combines information across multiple granularities, whereas MoE-AGIQA [389] integrates degradation-aware and semantic-aware experts via cross-attention. From a modeling perspective, several methods focus on improving feature representation and prediction stability. AMFF-Net [465] fuses global and local features with alignment information, PSCR [409] uses patch-sampling contrastive regression to improve robustness and reduce geometric bias, and NR/FR-AIGCIQA [411] compares no-reference (NR) regression with full-reference (FR) feature fusion. Other approaches combine multimodal information more explicitly: TIER [410] regresses quality jointly from text and image encoders, JOINT [32] models technical and rationality cues to capture naturalness, and IPCE [255] transforms calibrated classification probabilities into continuous regression targets to improve CLIP-based quality assessment.

5) Reasoning-Driven Generation and Editing Evaluation. A complementary subdirection explicitly evaluates understanding–generation integration by requiring models to reason (decompose instructions, ground targets, plan edits) before producing or modifying images, and by scoring both instruction satisfaction and perceptual plausibility. RISEBench [455] targets reasoning-informed visual editing with four families of reasoning (temporal, causal, spatial, logical) and three evaluation axes (instruction reasoning, appearance consistency, and visual plausibility), using both human judgment and LMM-as-a-judge. GoT [63] operationalizes a "first reason, then generate/edit" paradigm at scale, coupling chain-of-thought with diffusion-based editing to test whether explicit reasoning improves fidelity and controllability. SmartEdit [113] explores com-

plex instruction–based editing with multimodal LLMs, introducing a Reason-Edit protocol that disentangles understanding, grounding, and editing to diagnose failure modes. Knowledge-centric evaluation further stresses semantic adequacy: WISE [251] probes world-knowledge–informed semantic correctness for T2I, complementing editing benchmarks along the knowledge axis, while KRIS-Bench [367] organizes editing tasks by factual, conceptual, and procedural knowledge with multi-dimensional reasoning diagnostics. Methodologically, CoT-editing [133] demonstrates that injecting chain-of-thought into the editing loop (plan–act–verify) yields measurable gains on multi-constraint, multi-step edits. Together, these works push AIGC evaluation beyond pixel-level scores toward reasoning-aware, plan-then-edit assessment, and can be used in tandem with alignment/preference models to provide calibrated, human-aligned judgments for complex editing scenarios.

In all, the field progresses from proxy scores to human-aligned alignment/preference modeling and instruction-tuned LMM assessors, while specialized architectures capture AIGI-specific semantics and artifacts; public challenges offer external calibration, and quality-aware refinement shows a practical path to evaluation-for-generation.

3.2 Video Generation Evaluation

This section reviews evaluation methodologies for video generation, which represent one of the most challenging modalities due to their inherent spatiotemporal complexity. Compared to image generation, video outputs require models not only to maintain frame-level visual quality but also to capture motion consistency, temporal coherence, and multimodal alignment with prompts or instructions.

3.2.1 LMM-Based VQA for User-Generated Content

Recent advances have seen the emergence of LMM-powered methods for video quality assessment (VQA), particularly targeting user-generated content (UGC). LMM-VQA [78] integrates a motion processor (SlowFast) into an LMM backbone and adopts a multi-prompt, multi-stage fine-tuning strategy to achieve high-performance VQA. FineVQ [58] introduces fine-grained MOS annotations across six dimensions and applies efficient LoRA-based fine-tuning for multi-dimensional UGC video quality prediction. VQA^2 [123] jointly addresses quality scoring and fine-grained description generation by curating over 110K instruction-tuning samples and conducting multi-stage supervised fine-tuning (SFT), resulting in an LMM capable of producing both quantitative and qualitative assessments. Extending this work, Omni-VQA [124] adopts a human-in-the-loop approach driven largely by machine rejection-sampling, producing one of the largest VQA-instruction datasets to date (over 80K UGC videos and 400K instruction–tuning pairs) substantially reducing manual annotation requirements while improving versatility. LMM-PVQA [18] draws from the Compare2Score paradigm [477], replacing labor-intensive MOS labels with pairwise preference annotations to enhance both scalability and out-of-distribution (OOD) generalization.

3.2.2 Datasets for AI-Generated Video Quality Assessment

The landscape of AI-generated video (AIGV) quality assessment datasets has evolved from small-scale, single-aspect resources to large-scale, multi-dimensional, and task-specific benchmarks. Early works such as Chivileva et al. [41] and EvalCrafter [207] focus on perceptual and alignment quality with 1K–2.5K videos, while FETV [209] and VBench [114] introduce fine-grained attribute annotations and preference-based evaluation at moderate scales. Recent general-purpose benchmarks have scaled substantially, including AIGVQA-DB[318] (36.6K videos, 122K MOS/pairwise annotations), AIGVE-60K [315] (60K videos with $\sim$12K MOS and 60K instruction pairs), and T2VQA-DB [143] (10K videos with MOS from 27 subjects). Meanwhile, specialized datasets address more focused aspects of video and audio–visual evaluation. GAIA [31] emphasizes motion realism, providing 9K video–action pairs with 971K human ratings. Human-AGVQA-DB [440] targets the quality of human activity in videos, while TDVE-DB [319] concentrates on assessing text-driven video editing. Extending beyond vision, AGAVQA-3K [20] incorporates audio–visual con-

tent to evaluate cross-modal consistency and quality. Collectively, these datasets illustrate a clear trajectory toward greater scale, diversity, and multidimensionality, integrating perceptual, semantic, temporal, and cross-modal quality dimensions to support both benchmarking and training of LMM-based evaluators.

3.2.3 Models for AI-Generated Video Quality Assessment

Architectures for AIGV quality evaluation often combine spatio-temporal modeling with multimodal alignment. AIGV-Assessor [318] fuses 2D (InternViT) and 3D (SlowFast) features within a multi-stage LoRA framework to provide both single- and pairwise quality predictions across multiple dimensions. LGVQ [439] leverages foreground-background prompting to probe spatial-temporal disentanglement. GHVQ [440] addresses human activity quality assessment using a dual-branch architecture (spatial quality analyzer and action quality analyzer), integrated with CLIP-based regression. LOVE [315] systematically evaluates both AIGV quality and LMM understanding across three dimensions and twenty subtasks. TDVE-Assessor [319] benchmarks TDVE-specific tasks, focusing on editing-relevant quality dimensions. VQ-Insight [433] introduces a rule-based reinforcement learning framework with multi-task, reward-driven training, achieving strong AIGV quality assessment performance and enabling a closed-loop generation–evaluation paradigm where assessment feedback improves generation. AGAV-Rater [20] pioneers AGAV-specific quality assessment by integrating a dedicated audio-processing module within a multi-stage fine-tuning pipeline, achieving efficient and modality-aware multimodal evaluation.

4) Quality Assessment for AI-Generated Talking Heads and Digital Humans. Beyond general-purpose AIGVQA, a dedicated line of work targets human-centric generative content, especially talking heads and digital humans. Who is a Better Talker [466] and THQA [471] establish large-scale perceptual quality databases for AI-generated talking heads, with multi-dimensional MOS annotations capturing lip–audio synchronization, naturalness, and overall quality. Extending beyond speech-driven avatars, Who is a Better Imitator [470] addresses animated human imitation, combining subjective studies and objective metrics to assess realism and fidelity. Methodologically, MI3S [468] introduces an LMM-assisted framework that integrates multimodal reasoning for evaluating talking head quality, highlighting the potential of foundation models in perceptual assessment. From a system perspective, An Implementation of Multimodal Fusion System [467] explores practical pipelines for digital human generation via multimodal fusion, bridging algorithmic evaluation with engineering deployment. Together, these works form a complementary branch of AIGVQA, emphasizing perceptual alignment, expressiveness, and multimodal consistency in human-centric content.

3.3 Audio Generation Evaluation

This section reviews representative approaches for evaluating the quality of audio generation, particularly in speech synthesis and related tasks. Methods can be broadly categorized into deep learning–based, LLM/ALM–based, and benchmark–driven evaluations.

1) Deep Learning–based Evaluation. Learning the mapping from speech signals to human subjective scores via deep neural networks has become a mainstream paradigm in speech quality assessment. MOSNet [214] pioneered deep learning–based MOS prediction for converted speech. MOSA-Net+ [417] extended this approach by leveraging acoustic features extracted from Whisper. MOSLight [181] pursued a lightweight design using 1D convolutions for faster inference. MBNet [151], DeePMOS [184], and LDNet [110] incorporated listener-specific perceived quality scores in addition to the mean opinion scores. ADTMOS [183] further enhanced robustness with a frame-wise MOS generator and an audio distortion token extractor. UAMOS [314] proposed an uncertainty-aware MOS framework to improve reliability in open-world applications. Audiobox Aesthetics [302] decomposed human listening perspectives into four perceptual axes for aesthetic evaluation. While most methods resample audio to a fixed rate, HighRateMOS [271] was the first to explicitly model sampling rate as an evaluation factor.

2) LLM/ALM–based Evaluation. Recent studies have explored LLMs and audio language models (ALMs) for audio generation evaluation. In zero-shot settings, Chiang et al. [38] enabled spoken language models such as ChatGPT-4o-audio and Gemini-2.5-pro to role-play as judges, assessing speaking style appropriateness and human-likeness. Fine-tuning approaches adapt ALMs (e.g., SALMONN [298], Qwen-Audio [43], Qwen2-Audio [42]) for multiple assessment tasks, including MOS and speaker similarity (SIM) prediction, A/B preference testing, and natural language quality descriptions [329]. Qwen-ALLD [23] introduced the first natural language–based speech quality dataset and employed LLM distillation to align ALMs for MOS, SIM, and A/B testing. QualiSpeech [328] further advanced natural language–based evaluation by integrating reasoning and contextual cues to improve accuracy and interpretability.

3) Benchmark–based Evaluation. To systematically evaluate (Text-to-Speech) TTS and speech synthesis systems, several multi-dimensional benchmark frameworks have been proposed. BASE-TTS [147] presented an English emergent abilities suite covering seven categories of text (e.g., emotions, paralinguistics, syntactic complexities) to test fine-grained semantic capture. DiscreteEval [327] assessed five dimensions: speaking style, intelligibility, speaker consistency, prosodic variation, and spontaneity. EmergentTTS-Eval [235] addressed six challenging scenarios (e.g., emotions, foreign words, complex pronunciation) with ALM-generated test cases evaluated by an ALM. ATT [338] introduced the Audio Turing Test, combining a multidimensional Chinese corpus with a Turing Test–inspired protocol to assess human-likeness in LLM-based TTS. TTSDS2 [247] benchmarked 20 open-source TTS systems across 14 languages along four axes, including prosody and intelligibility. InstructTTSEval [106] evaluated instruction-driven TTS models on acoustic-parameter control, descriptive-style directives, and role-play scenarios. Mos-Bench [109] contributed SHEET, a toolkit supporting MOS prediction for single- and multi-dataset training, along with diagnostic tools such as best score difference/ratio and latent-space visualization.

3.4 3D Content Generation Evaluation

This section reviews representative approaches for evaluating the quality of 3D content generation, particularly in text-to-3D and image-to-3D synthesis. Existing studies span the construction of large-scale annotated datasets, the development of subjective and objective evaluation methodologies, and the exploration of automated, human-aligned scoring pipelines. Methods can be broadly categorized into text-to-3D evaluation, image-to-3D evaluation (single-image and multi-image), and automatic human-aligned evaluation.

1) Text-to-3D Generation Quality Assessment. At the early stages, 3D quality assessment methods [443; 444] are developed based on limited-scale datasets. Later, several benchmarks target the perceptual evaluation of 3D assets generated from textual descriptions. MATE-3D [439] contains 1,280 textured meshes prompted by eight diverse categories, with 107.5K human annotations across four dimensions: Alignment, Geometry, Texture, and Overall. The authors also propose HyperScore, a hypernetwork-based evaluator for multi-dimensional quality prediction. 3DGCQA [472] aggregates 313 textured meshes from seven representative text-to-3D models, collecting subjective ratings on Alignment and Overall Quality while benchmarking existing objective metrics. AIGC-T23DAQA [73] comprises 969 validated 3D assets from 170 prompts via six models, with ratings for Quality, Authenticity, and Text–Content Correspondence. The associated T23DAQA model is tailored for these dimensions. GT23D-Bench [293] provides $\sim$400k multimodal annotations (multi-view renderings, depth, normals, and hierarchical text) to evaluate both text–3D alignment and visual quality in a unified framework. CAP (unpublished) proposes a no-reference quality assessment method focusing on geometry–texture coherence.

2) Image-to-3D Generation Quality Assessment. Image-to-3D evaluation can be divided into single-image and multi-image settings. SI23DCQA [72] assesses 1.5K 3D assets generated from 300 input images (spanning real photographs, AI-generated content, and model-rendered inputs), using five SI23D algorithms. Subjective ratings are collected for Overall,

Color, and Shape. NeRF-based works include NeRF-NQA [262], the first NR-QA method for densely observed NeRF/NVS scenes, combining viewwise and pointwise evaluations to capture inter-view consistency and surface angular quality. Explicit-NeRF-QA [378] focuses on compression, providing subjective scores for multiple parameter levels of 22 source objects. Martin et al. [238; 239] conduct extensive subjective studies across scene types and benchmark FR/NR metrics against human scores. For Gaussian splatting, GS-QA [240] evaluates static GS methods under 360° and forward-facing trajectories with subjective ratings and 18 objective metrics. 3DGS-IEval-15K [377] is the first large-scale IQA dataset for compressed 3DGS, containing 15.2K rendered images from 10 real scenes, with MOS from 60 viewers and 30 IQA metrics benchmarked.

3) Automatic, Human-Aligned Evaluation. To overcome the scalability limits of subjective studies, recent works develop automatic evaluation pipelines aligned with human judgments. GPT-4V Evaluator [362] uses GPT-4V to perform pairwise comparisons of 3D assets (via multi-view projections) along dimensions such as Alignment, Plausibility, and Texture–Geometry Coherence, deriving Elo ratings for model ranking. Eval3D [59] integrates pretrained models and LMMs to score Geometric Consistency, Semantic Consistency, Structural Consistency, Text–3D Alignment, and Aesthetics. 3DGen-Bench [437] combines CLIP-based 3DGen-Score and MLLM-based 3DGen-Eval to better correlate with human preferences through multimodal reasoning. LMM-PCQA [448] targets point cloud quality assessment by projecting point clouds into multiple 2D views, enabling LMMs to generate textual descriptions fused with structural features for final scoring.

4 Evaluation Tools and Platforms

A growing ecosystem of evaluation toolkits and benchmarking platforms has emerged to support the standardized, reproducible, and community-driven assessment of LMMs and foundation models more broadly. These resources differ in scope and design philosophy: toolkits often emphasize lightweight, reproducible pipelines, whereas platforms provide dynamic, interactive leaderboards and broader community participation.

4.1 Evaluation Tools

Early efforts have focused on providing unified interfaces for running and analyzing multimodal benchmarks. VLMEvalKit [57] is a widely adopted open-source toolkit that streamlines evaluation workflows by offering automated pipelines, results aggregation, and standardized model comparisons across diverse multimodal tasks. Building upon VLMEvalKit outputs, the OpenVLM Leaderboard [57] integrates results into an interactive Hugging Face platform hosted by OpenCompass, enabling public model submissions and fine-grained performance comparisons over 70+ image and video benchmarks.

Beyond toolkit-style designs, LMMS-Eval [427] provides a more comprehensive framework covering over 50 tasks and multiple models, with transparent pipelines and extensions such as Lite (efficiency-focused) and LiveBench (dynamic tracking). In contrast, GenAI-Arena [127] adopts a community-driven paradigm: users vote on head-to-head generations for text-to-image, image editing, and text-to-video tasks, with Elo ratings aggregated into GenAI-Bench, thereby reflecting collective user preferences.

4.2 Evaluation Platforms

A parallel line of development has emphasized large-scale, dynamic benchmarking platforms, which aggregate heterogeneous tasks into evolving leaderboards (screenshots are shown in Fig 6). OpenCompass [254] exemplifies this trend by consolidating evaluations across reasoning, LLMs, vision-language, and spatial domains, providing multi-capability insights under standardized workflows. Similarly, HELM [182] from Stanford CRFM offers a modular framework that compares models under diverse real-world scenarios, extending metrics beyond accuracy to fairness, robustness, efficiency, and safety. LiveBench [349]

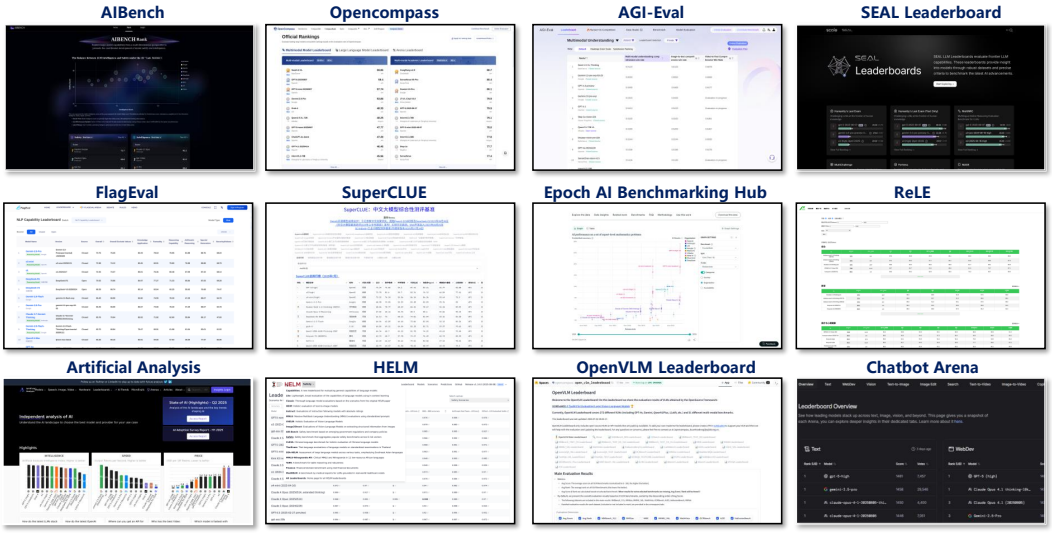

Figure 6: Screenshots of the representative evaluation platforms.

pushes this further by refreshing tasks monthly to mitigate contamination, while supporting automated scoring across multiple categories.

Commercial and independent initiatives also play a growing role. Epoch AI's Benchmarking Hub [61] visualizes historical benchmark performance trends, and Artificial Analysis [9] provides comparative intelligence across providers with insights into latency, price, and safety. Scale's SEAL Leaderboards [279] adopt expert-curated, high-complexity evaluations, ensuring robustness by testing models only on unseen prompts. FlagEval [66] introduces customizable tasks and debate-style comparisons, while AGI-Eval [2] integrates automatic and human reviews under a general scheme, supporting both official and user-submitted test suites.

In the Chinese ecosystem, several domain-focused platforms have emerged. ReLE [269] provides a live-updating leaderboard covering hundreds of fine-grained dimensions in education, finance, healthcare, and law. SuperCLUE [381] extends traditional Chinese LLM benchmarks with three complementary tracks (CArena, OPEN, CLOSE), each targeting distinct evaluation modes from user preferences to open- and closed-form tasks. AIBench [445] emphasizes fast iteration across both intelligence and safety dimensions, while incorporating cost-effectiveness as an explicit axis.

Taken together, these tools and platforms highlight the diversification of evaluation ecosystems: from lightweight toolkits facilitating reproducible pipelines, to large-scale leaderboards capturing community preferences, to fast-updating hubs tracking safety, efficiency, and cost. Their complementary designs collectively advance transparent, standardized, and user-centered evaluation of foundation models.

5 Future Look & Conclusion

Large Multimodal Models (LMMs) have achieved remarkable progress in both understanding and generation, supported by the rapid development of benchmarks, leaderboards, and evaluation tools. Yet, our survey highlights that current evaluation practices remain fragmented and face several open challenges. Looking ahead, we outline several promising directions:

1) Toward unified evaluation paradigms. While understanding and generation evaluations are often treated separately, their increasing convergence suggests the need for integrated frameworks that capture their interdependence. Unified benchmarks should simultaneously test perception, reasoning, and generation, enabling more holistic assessments.

2) Dynamic and continuously updated benchmarks. Static datasets risk contamination and rapid saturation as models improve. Future evaluation must incorporate dynamic benchmarks that evolve over time, incorporating adversarially designed samples, human feedback, and domain-specific updates to maintain reliability and forward compatibility.

3) Human-centered and trustworthy evaluation. Beyond technical accuracy, evaluation must emphasize alignment with human values, fairness, interpretability, and safety. Incorporating large-scale human feedback, preference modeling, and ethical auditing will be critical for assessing whether LMMs can be responsibly deployed in high-stakes domains.

4) Community-driven infrastructure and ecosystem. Open leaderboards, transparent evaluation protocols, and collaborative platforms are indispensable for reproducibility and progress tracking. Strengthening these infrastructures with standardized metrics and open-source tools will accelerate both research and industrial adoption.

In conclusion, this survey provides the first comprehensive review of LMM evaluation across general–specialized understanding, modality-specific generation, and community infrastructures. By synthesizing current progress and challenges, we aim to chart a roadmap for building systematic, reliable, and trustworthy evaluation ecosystems in the era of foundation multimodal models.

REFERENCES

[1] RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

[2] AGI-Eval. Large language model leaderboard. https://agi-eval.cn/mvp/listSummaryIndex, May 2025.

[3] Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research, 2024.

[4] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808, 2024.

[5] Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. Choice: Benchmarking the remote sensing capabilities of large vision-language models. *arXiv preprint arXiv:2411.18145*, 2024.

[6] Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. Mm-phyqa: Multimodal physics question answering with multi-image cot prompting. *arXiv preprint arXiv:2404.08704*, 2024.

[7] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[8] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[9] Artificial-Analysis. Artificial-analysis. https://artificialanalysis.ai/, 2025.

[10] Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024.

[11] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

[12] Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*, 2024.

[13] Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*, 2023.

[14] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.

[15] Andrea Burns, Deniz Arsan, Sanjna Agrawal, Ranjitha Kumar, Kate Saenko, and Bryan A Plummer. A dataset for interactive vision-language navigation with unknown command feasibility. In *European Conference on Computer Vision*, pages 312–328. Springer, 2022.

[16] Roman Bushuiev, Anton Bushuiev, Niek de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

[17] He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

[18] Linhan Cao, Wei Sun, Kaiwei Zhang, Yicong Peng, Guangtao Zhai, and Xiongkuo Min. Breaking annotation barriers: Generalized video quality assessment via ranking-based self-supervision. *arXiv preprint arXiv:2505.03631*, 2025.

[19] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M. Rehg, and Chao Zheng. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21819–21830, June 2024.

[20] Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, and Guangtao Zhai. Agav-rater: adapting large multimodal model for ai-generated audio-visual quality assessment. *arXiv preprint arXiv:2501.18314*, 2025.

[21] Linzheng Chai, Jian Yang, Shukai Liu, Wei Zhang, Liran Wang, Ke Jin, Tao Sun, Congnan Liu, Chenchen Zhang, Hualei Zhu, et al. Multilingual multimodal software developer for code generation. *arXiv preprint arXiv:2507.08719*, 2025.

[22] Bhavik Chandna, Mariam Aboujenane, and Usman Naseem. Extremeaigc: Benchmarking lmm vulnerability to ai-generated extremist content. *arXiv preprint arXiv:2503.09964*, 2025.

[23] Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. Audio large language models can be descriptive speech quality evaluators. *arXiv preprint arXiv:2501.17202*, 2025.

[24] Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. Gui-world: A video benchmark and dataset for multimodal gui-oriented understanding. *arXiv preprint arXiv:2406.10819*, 2024.

[25] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv, abs/2402.18060*, 2024.

[26] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 513–524, 2025.

[27] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024.

[28] Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224*, 2023.

[29] Zeren Chen, Zhelun Shi, Xiaoya Lu, Lehan He, Sucheng Qian, Zhenfei Yin, Wanli Ouyang, Jing Shao, Yu Qiao, Cewu Lu, et al. Rh20t-p: A primitive-level robotic dataset towards composable generalization agents. *arXiv preprint arXiv:2403.19622*, 2024.

[30] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[31] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

[32] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Exploring the naturalness of ai-generated images, 2024.

[33] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

[34] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332, 2024.

[35] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

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

[37] Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23678–23686, 2025.

[38] Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, et al. Audio-aware large language models as judges for speaking styles. *arXiv preprint arXiv:2506.05984*, 2025.

[39] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[40] Hsu-kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *arXiv preprint arXiv:2502.09980*, 2025.

[41] Iya Chivileva, Philip Lynch, Tomas E. Ward, and Alan F. Smeaton. Measuring the quality of text-to-video model outputs: Metrics and dataset. 2023.

[42] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[43] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[44] Chuan Cui, Kejiang Chen, Zhihua Wei, Wen Shen, Weiming Zhang, and Nenghai Yu. M3-agiqa: Multimodal, multi-round, multi-aspect ai-generated image quality assessment, 2025.

[45] Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, et al. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. *arXiv preprint arXiv:2503.13517*, 2025.

[46] Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and Xuming Hu. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. *arXiv preprint arXiv:2505.15472*, 2025.

[47] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[48] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024.

[49] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

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

[51] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19119–19128, 2022.

[52] Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170, 2024.

[53] Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10707–10715, 2024.

[54] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. *arXiv preprint arXiv:2504.07957*, 2025.

[55] Gabriel Downer, Sean Craven, Damian Ruck, and Jake Thomas. Text2vlm: Adapting text-only datasets to evaluate alignment training in visual language models. *arXiv preprint arXiv:2507.20704*, 2025.

[56] Yuntao Du, Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. Mmke-bench: A multimodal editing benchmark for diverse visual knowledge. *arXiv preprint arXiv:2502.19870*, 2025.

[57] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.

[58] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3206–3217, 2025.

[59] Shivam Duggal, Yushi Hu, Oscar Michel, Aniruddha Kembhavi, William T Freeman, Noah A Smith, Ranjay Krishna, Antonio Torralba, Ali Farhadi, and Wei-Chiu Ma. Eval3d: Interpretable and fine-grained evaluation for 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13326–13336, 2025.

[60] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, 2021.

[61] Epoch-AI. Ai performance on a set of expert-level mathematics problems. https://epoch.ai/data/ai-benchmarking-dashboard, May 2025.

[62] Benjamin Estermann, Luca Lanzendörfer, Yannick Niedermayr, and Roger Wattenhofer. Puzzles: A benchmark for neural algorithmic reasoning. *Advances in Neural Information Processing Systems*, 37:127059–127098, 2024.

[63] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. GoT: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

[64] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding, 2024.

[65] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):220106, 2024.

[66] FlagEval. Nlp capability leaderboard. https://flageval.baai.ac.cn/#/leaderboard/, March 2025.

[67] Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. Wikimixqa: A multimodal benchmark for question answering over tables and charts. *arXiv preprint arXiv:2506.15594*, 2025.

[68] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

[69] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024.

[70] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025.

[71] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.

[72] Kang Fu, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, Xiongkuo Min, Jia Wang, and Zhai Guangtao. Si23dcqa: Perceptual quality assessment of single image-to-3d content. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2025.

[73] Kang Fu, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, Xiongkuo Min, Jia Wang, and Guangtao Zhai. Multi-dimensional quality assessment for text-to-3d assets: Dataset and model. *IEEE Transactions on Multimedia*, 2025.

[74] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

[75] Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, et al. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*, 2024.

[76] Junyu Gao, Liangliang Zhao, and Xuelong Li. Nwpu-moc: A benchmark for fine-grained multicategory object counting in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[77] Lancheng Gao, Ziheng Jia, Yunhao Zeng, Wei Sun, Yiming Zhang, Wei Zhou, Guangtao Zhai, and Xiongkuo Min. Eemo-bench: A benchmark for multi-modal large language models on image evoked emotion assessment. *arXiv preprint arXiv:2504.16405*, 2025.

[78] Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. Lmm-vqa: Advancing video quality assessment with large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[79] Gemini Team, Google DeepMind. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next-Generation Agentic Capabilities. Technical Report v2.5, Google DeepMind, June 2025.

[80] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[81] Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Yujiu Yang, Yan Teng, Yu Qiao, et al. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:7256–7295, 2024.

[82] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[83] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

[84] Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024.

[85] Siyuan Guo, Lexuan Wang, Chang Jin, Jinxian Wang, Han Peng, Huayang Shi, Wengen Li, Jihong Guan, and Shuigeng Zhou. M3-20m: A large-scale multi-modal molecule dataset for ai-driven drug design and discovery. *Journal of bioinformatics and computational biology*, 23(2):2550006, 2025.

[86] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. Surds: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. *arXiv preprint arXiv:2411.13112*, 2024.

[87] Yangyang Guo, Fangkai Jiao, Liqiang Nie, and Mohan Kankanhalli. The vllm safety paradox: Dual ease in jailbreak attack and defense. *arXiv preprint arXiv:2411.08410*, 2024.

[88] Yijin Guo, Kaiyuan Ji, Xiaorong Zhu, Junying Wang, Farong Wen, Chunyi Li, Zicheng Zhang, and Guangtao Zhai. Human-centric evaluation for foundation models. *arXiv preprint arXiv:2506.01793*, 2025.

[89] Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. Polymath: A challenging multi-modal mathematical reasoning benchmark. *arXiv preprint arXiv:2410.14702*, 2024.

[90] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024.

[91] Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, Peng Jia, et al. Driveaction: A benchmark for exploring human-like driving decisions in vla models. *arXiv preprint arXiv:2506.05667*, 2025.

[92] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

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

[94] Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning. *arXiv preprint arXiv:2401.14011*, 2024.

[95] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23, 2015.

[96] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55(3):661–688, 2021.

[97] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models, 2025.

[98] Wenpin Hou and Zhicheng Ji. Geneturing tests gpt models in genomics. *BioRxiv*, pages 2023–03, 2023.

[99] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891, 2021.

[100] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Carbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots, 2025.

[101] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

[102] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 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, 2024.

[103] Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Vlkeb: A large vision-language model knowledge editing benchmark. *Advances in Neural Information Processing Systems*, 37:9257–9280, 2024.

[104] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024.

[105] Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, et al. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*, 2024.

[106] Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. InstructTTSEval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*, 2025.

[107] Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*, 2025.

[108] Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3680–3688, 2025.

[109] Wen-Chin Huang, Erica Cooper, and Tomoki Toda. Mos-bench: Benchmarking generalization abilities of subjective speech quality assessment models. *arXiv preprint arXiv:2411.03715*, 2024.

[110] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. LDNet: Unified listener dependent modeling in mos prediction for synthetic speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 896–900, 2022.

[111] Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024.

[112] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023.

[113] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8362–8371, June 2024.

[114] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. 2023.

[115] Yulong Hui, Yao Lu, and Huanchen Zhang. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *Advances in Neural Information Processing Systems*, 37:67200–67217, 2024.

[116] Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *arXiv preprint arXiv:2503.10621*, 2025.

[117] Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, 2023.

[118] Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. GRASP: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023.

[119] Tao Qin Jes Frellsen, Kun Zhang. Neurips 2025 embodied agent interface challenge. `https://blog.neurips.cc/2025/06/27/neurips-2025-competitions-announced/`.

[120] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[121] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022.

[122] Qi Jia, Xiang Yue, Shanshan Huang, Ziheng Qin, Yizhu Liu, Bill Yuchen Lin, and Yang You. Visual perception in text strings. *arXiv preprint arXiv:2410.01733*, 2024.

[123] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Vqa2: Visual question answering for video quality assessment. *arXiv preprint arXiv:2411.03795*, 2024.

[124] Ziheng Jia, Zicheng Zhang, Zeyu Zhang, Yingji Liang, Xiaorong Zhu, Chunyi Li, Jinliang Han, Haoning Wu, Bin Wang, Haoran Zhang, et al. Scaling-up perceptual video quality assessment. *arXiv preprint arXiv:2505.22543*, 2025.

[125] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534, 2024.

[126] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

[127] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

[128] Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. *arXiv preprint arXiv:2503.11117*, 2025.

[129] Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. Modscan: Measuring stereotypical bias in large vision-language models from vision and language modalities. *CoRR*, 2024.

[130] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv, abs/1909.06146*, 2019.

[131] Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: Fairness benchmarking for medical imaging foundation models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[132] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

[133] Mengxue Kang, Xinyu Zhang, Fei Wei, and Shuang Xu. Enhancing image editing with chain-of-thought reasoning and multimodal large language models. In *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, April 2025. IEEE.

[134] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2024.

[135] Aniruddha Kembhavi, Matt Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. Diagram understanding in geometry questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[136] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4999–5007, 2017.

[137] Jihyung Kil, Zheda Mai, Justin Lee, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, Arpita Chowdhury, and Wei-Lun Chao. Compbench: A comparative reasoning benchmark for multimodal llms. *arXiv e-prints*, pages arXiv–2407, 2024.

[138] Seunghee Kim, Changhyeon Kim, and Taeuk Kim. Fcmr: Robust evaluation of financial cross-modal multi-hop reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[139] Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024.

[140] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[141] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

[142] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*, 2021.

[143] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. 2024.

[144] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[145] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.

[146] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.

[147] Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent Van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*, 2024.

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

[149] Khiem Le, Zhichun Guo, Kaiwen Dong, Xiaobao Huang, Bozhao Nan, Roshni Iyer, Xiangliang Zhang, Olaf Wiest, Wei Wang, and Nitesh V Chawla. Molx: Enhancing large language models for molecular learning with a multi-modal extension. *arXiv preprint arXiv:2406.06777*, 2024.

[150] DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. Are vision-language models safe in the wild? a meme-based benchmark study. *arXiv preprint arXiv:2505.15389*, 2025.

[151] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. MBNet: Mos prediction for synthesized speech with mean-bias network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 391–395, 2021.

[152] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples, 2025.

[153] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models. `https://github.com/EvolvingLMMs-Lab/lmms-eval`, March 2024. Version v0.1.0.

[154] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[155] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, 2024.

[156] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[157] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Haoning Wu, Weixia Zhang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Aigiqa-20k: A large database for ai-generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[158] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Q-refine: A perceptual quality refiner for ai-generated image. *arXiv preprint arXiv:2401.01117*, 2024.

[159] Chunyi Li, Xiele Wu, Haoning Wu, Donghui Feng, Zicheng Zhang, Guo Lu, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Cmc-bench: Towards a new paradigm of visual signal compression. *arXiv preprint arXiv:2406.09356*, 2024.

[160] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[161] Guanzhen Li, Yuxi Xie, and Min-Yen Kan. Mvp-bench: Can large vision–language models conduct multi-level visual perception like humans? *arXiv preprint arXiv:2410.04345*, 2024.

[162] Jiangtong Li, Yiyun Zhu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. Cfbenchmark-mm: Chinese financial assistant benchmark for multimodal large language model. *arXiv preprint arXiv:2506.13055*, 2025.

[163] Jiansheng Li, Xingxuan Zhang, Hao Zou, Yige Guo, Renzhe Xu, Yilong Liu, Chuzhao Zhu, Yue He, and Peng Cui. Counts: Benchmarking object detectors and multimodal large language models under distribution shifts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9186–9198, 2025.

[164] Jiaqi Li, Miaozeng Du, Chuanyi Zhang, Yongrui Chen, Nan Hu, Guilin Qi, Haiyun Jiang, Siyuan Cheng, and Bozhong Tian. Mike: A new benchmark for fine-grained multimodal entity knowledge editing. *arXiv preprint arXiv:2402.14835*, 2024.

[165] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453, 2024.

[166] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423, 2025.

[167] Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, Zhiyong Huang, and Jing Ma. Mmcode: Benchmarking multimodal large language models for code generation with visually rich programming problems. *arXiv preprint arXiv:2404.09486*, 2024.

[168] Keliang Li, Zaifei Yang, Jiahe Zhao, Hongze Shen, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Herm: Benchmarking and enhancing multimodal llms for human-centric understanding. *arXiv preprint arXiv:2410.06777*, 2024.

[169] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024.

[170] Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024.

[171] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. M3dbench: Let's instruct large models with multi-modal 3d prompts, 2023.

[172] Qiang Li, Qingsen Yan, Haojian Huang, Peng Wu, Haokui Zhang, and Yanning Zhang. Text-visual semantic constrained ai-generated image quality assessment, 2025.

[173] Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*, 2024.

[174] Siqi Li, Yufan Shen, Xiangnan Chen, Jiayi Chen, Hengwei Ju, Haodong Duan, Song Mao, Hongbin Zhou, Bo Zhang, Bin Fu, et al. Gdi-bench: A benchmark for general document intelligence with vision and reasoning decoupling. *arXiv preprint arXiv:2505.00063*, 2025.

[175] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024.

[176] Xirui Li, Hengguang Zhou, Ruochen Wang, Tianyi Zhou, Minhao Cheng, and Cho-Jui Hsieh. Mossbench: Is your multimodal language model oversensitive to safe queries? *arXiv preprint arXiv:2406.17806*, 2024.

[177] Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-Omni-1.5 Technical Report, January 2025.

[178] Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. Facial affective behavior analysis with instruction tuning. In *European Conference on Computer Vision*, pages 165–186. Springer, 2024.

[179] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[180] Yue Li, Meng Tian, Zhenyu Lin, Jiangtong Zhu, Dechang Zhu, Haiqiang Liu, Zining Wang, Yueyi Zhang, Zhiwei Xiong, and Xinhai Zhao. Fine-grained evaluation of large vision-language models in autonomous driving. *arXiv preprint arXiv:2503.21505*, 2025.

[181] Zitong Li and Wei Li. MOSLight: A lightweight data-efficient system for non-intrusive speech quality assessment. In *Proc. Interspeech 2023*, pages 5386–5390, 2023.

[182] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification.

[183] Qiao Liang, Ying Shen, Tiantian Chen, Lin Zhang, and Shengjie Zhao. ADTMOS–synthesized speech quality assessment based on audio distortion tokens. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[184] Xinyu Liang, Fredrik Cumlin, Christian Schüldt, and Saikat Chatterjee. DeePMOS: deep posterior mean-opinion-score of speech. In *Proceedings of INTERSPEECH*, pages 526–530, 2023.

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

[186] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models, 2024.

[187] Zhichao Liao, Xiaokun Liu, Wenyu Qin, Qingyu Li, Qiulin Wang, Pengfei Wan, Di Zhang, Long Zeng, and Pingfa Feng. Humanaesexpert: Advancing a multimodality foundation model for human image aesthetic assessment. *arXiv preprint arXiv:2503.23907*, 2025.

[188] Minzhi Lin, Tianchi Xie, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. Infochartqa: A benchmark for multimodal question answering on infographic charts. *arXiv preprint arXiv:2505.19028*, 2025.

[189] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[190] Zhiyu Lin, Zhengda Zhou, Zhiyuan Zhao, Tianrui Wan, Yilun Ma, Junyu Gao, and Xuelong Li. Webuibench: A comprehensive benchmark for evaluating multimodal large language models in webui-to-code. *arXiv preprint arXiv:2506.07818*, 2025.

[191] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[192] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[193] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*, 2024.

[194] Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? *arXiv preprint arXiv:2505.19015*, 2025.

[195] Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*, 2024.

[196] Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *arXiv, abs/2407.10990*, 2024.

[197] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024.

[198] Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073, 2024.

[199] Pengfei Liu, Jun Tao, and Zhixiang Ren. A quantitative analysis of knowledge-learning preferences in large language models in molecular science. *Nature Machine Intelligence*, 7(2):315–327, 2025.

[200] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024.

[201] Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*, 2024.

[202] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, Fei Peng, Huiyuan Fu, Anlong Ming, Chuanming Wang, Huadong Ma, Shuai He, Zifei Dou, Shu Chen, Huacong Zhang, Haiyi Xie, Chengwei Wang, Baoying Chen, Jishen Zeng, Jianquan Yang, Weigang Wang, Xi Fang, Xiaoxin Lv, Jun Yan, Tianwu Zhi, Yabin Zhang, Yaohui Li, Yang Li, Jingwen Xu, Jianzhao Liu, Yiting Liao, Junlin Li, Zihao Yu, Yiting Lu, Xin Li, Hossein Motamednia, S. Farhad Hosseini-Benvidi, Fengbin Guan, Ahmad Mahmoudi-Aznaveh, Azadeh Mansouri, Ganzorig Gankhuyag, Kihwan Yoon, Yifang Xu, Haotian Fan, Fangyuan Kong, Shiling Zhao, Weifeng Dong, Haibing Yin, Li Zhu, Zhiling Wang, Bingchen Huang, Avinab Saha, Sandeep Mishra, Shashank Gupta, Rajesh Sureddi, Oindrila Saha, Luigi Celona, Simone Bianco, Paolo Napoletano, Raimondo Schettini, Junfeng Yang, Jing Fu, Wei Zhang, Wenzhi Cao, Limei Liu, Han Peng, Weijun Yuan, Zhan Li, Yihang Cheng, Yifan Deng, Haohui Li, Bowen Qu, Yao Li, Shuqing Luo, Shunzhou Wang, Wei Gao, Zihao Lu, Marcos V. Conde, Xinrui Wang, Zhibo Chen, Ruling Liao, Yan Ye, Qiulin Wang, Bing Li, Zhaokun Zhou, Miao Geng, Rui Chen, Xin Tao, Xiaoyu Liang, Shangkun Sun, Xingyuan Ma, Jiaze Li, Mengduo Yang, Haoran Xu, Jie Zhou, Shiding Zhu, Bohan Yu, Pengfei Chen, Xinrui Xu, Jiabin Shen, Zhichao Duan, Erfan Asadi, Jiahe Liu, Qi Yan, Youran Qu, Xiaohui Zeng, Lele Wang, and Renjie Liao. Ntire 2024 quality assessment of ai-generated content challenge, 2024.

[203] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024.

[204] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*, 2023.

[205] Xuannan Liu, Xing Cui, Peipei Li, Zekun Li, Huaibo Huang, Shuhan Xia, Miaoxuan Zhang, Yueying Zou, and Ran He. Jailbreak attacks and defenses against multimodal generative models: A survey. *arXiv preprint arXiv:2411.09259*, 2024.

[206] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multimodal instruction tuning dataset with extensive diversity. *Science China Information Sciences*, 67(12):220103, 2024.

[207] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. 2023.

[208] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[209] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. 2023.

[210] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[211] Zhiwei Liu, Lingfei Qian, Qianqian Xie, Jimin Huang, Kailai Yang, and Sophia Ananiadou. Mmaffben: A multilingual and multimodal affective analysis benchmark for evaluating llms and vlms. *arXiv preprint arXiv:2505.24423*, 2025.

[212] Ziqiang Liu, Feiteng Fang, Xi Feng, Xeron Du, Chenhao Zhang, Noah Wang, Qixuan Zhao, Liyang Fan, CHENGGUANG GAN, Hongquan Lin, et al. Ii-bench: An image implication understanding benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:46378–46480, 2024.

[213] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *Advances in Neural Information Processing Systems*, 37:8698–8733, 2024.

[214] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.

[215] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

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

[217] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[218] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[219] Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5838–5846, 2025.

[220] Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, et al. Codis: Benchmarking context-dependent visual comprehension for multimodal large language models. *arXiv preprint arXiv:2402.13607*, 2024.

[221] J Luo, Z Pang, Y Zhang, T Wang, L Wang, B Dang, J Lao, J Wang, J Chen, Y Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding, arxiv. *arXiv preprint arXiv:2406.10100*, 2024.

[222] Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, et al. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[223] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

[224] Weidi Luo, Qiming Zhang, Tianyu Lu, Xiaogeng Liu, Yue Zhao, Zhen Xiang, and Chaowei Xiao. Doxing via the lens: Revealing privacy leakage in image geolocation for agentic multi-modal large reasoning model. *arXiv e-prints*, pages arXiv–2504, 2025.

[225] Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315, 2025.

[226] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024.

[227] Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

[228] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[229] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

[230] Zhiming Ma, Xiayang Xiao, Sihao Dong, Peidong Wang, HaiPeng Wang, and Qingyun Pan. Sarchat-bench-2m: A multi-task vision-language benchmark for sar image interpretation. *arXiv preprint arXiv:2502.08168*, 2025.

[231] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024.

[232] U Mall, CP Phoo, MK Liu, C Vondrick, B Hariharan, and K Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. arxiv 2023. *arXiv preprint arXiv:2312.06960*.

[233] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023.

[234] Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. Climaqa: An automated evaluation framework for climate foundation models. *arXiv preprint arXiv:2410.16701*, 2024.

[235] Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. EmergentTTS-Eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *arXiv preprint arXiv:2505.23009*, 2025.

[236] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[237] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024.

[238] Pedro Martin, António Rodrigues, João Ascenso, and Maria Paula Queluz. Nerf-qa: Neural radiance fields quality assessment database. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 107–110. IEEE, 2023.

[239] Pedro Martin, António Rodrigues, João Ascenso, and Maria Paula Queluz. Nerf view synthesis: Subjective quality assessment and objective metrics evaluation. *IEEE Access*, 2024.

[240] Pedro Martin, António Rodrigues, João Ascenso, and Maria Paula Queluz. Gs-qa: Comprehensive quality assessment benchmark for gaussian splatting view synthesis. *arXiv preprint arXiv:2502.13196*, 2025.

[241] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022.

[242] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

[243] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[244] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

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

[246] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11):211301, 2024.

[247] Christoph Minixhofer, Ondrej Klejch, and Peter Bell. TTSDS2: Resources and benchmark for evaluating human-quality text to speech systems. *arXiv preprint arXiv:2506.19441*, 2025.

[248] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[249] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024.

[250] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, 2024.

[251] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[252] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, July 2024.

[253] OpenAI. Hello gpt-4o. System card / technical report, OpenAI, May 2024.

[254] Opencompass. Compassbench large language model leaderboard. https://rank.opencompass.org.cn/leaderboard-llm/, May 2025.

[255] Fei Peng, Huiyuan Fu, Anlong Ming, Chuanming Wang, Huadong Ma, Shuai He, Zifei Dou, and Shu Chen. Aigc image quality assessment via image-prompt correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6432–6441, 2024.

[256] Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Jeff Zhao, Huan He, Yi Han, et al. Multifinben: A multilingual, multimodal, and difficulty-aware benchmark for financial llm evaluation. *arXiv preprint arXiv:2506.14028*, 2025.

[257] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, et al. Humanity's last exam. https://arxiv.org/abs/2501.14249, 2025.

[258] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[259] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550, 2024.

[260] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.

[261] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[262] Qiang Qu, Hanxue Liang, Xiaoming Chen, Yuk Ying Chung, and Yiran Shen. Nerf-nqa: No-reference quality assessment for scenes generated by nerf and neural view synthesis methods. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2129–2139, 2024.

[263] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486*, 2024.

[264] Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *arXiv preprint arXiv:2406.13246*, 2024.

[265] Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation. In *Companion Proceedings of the ACM Web Conference 2025 (WWW '25 Companion)*, pages 785–788, 2025.

[266] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Androidinthewild: A large-scale dataset for android device control. *Advances in Neural Information Processing Systems*, 36:59708–59728, 2023.

[267] Shaina Raza, Aravind Narayanan, Vahid Reza Khazaie, Ashmal Vayani, Mukund S Chettiar, Amandeep Singh, Mubarak Shah, and Deval Pandya. Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*, 2025.

[268] Kaavya Rekanar, John M. Joyce, Martin Hayes, and Ciarán Eising. Drivqa: A gaze-based dataset for visual question answering in driving scenarios. *Data in Brief*, 59:111367, 2025.

[269] RELM. Really reliable live evaluation for llm. https://nonelinear.com/static/benchmarking.html, 2025.

[270] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024.

[271] Wenze Ren, Yi-Cheng Lin, Wen-Chin Huang, Ryandhimas E Zezario, Szu-Wei Fu, Sung-Feng Huang, Erica Cooper, Haibin Wu, Hung-Yu Wei, Hsin-Min Wang, et al. HighRateMOS: Sampling-rate aware modeling for speech quality assessment. *arXiv preprint arXiv:2506.21951*, 2025.

[272] Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728, 2024.

[273] Juan Rodriguez, Xiangru Jian, Siba Smarak Panigrahi, Tianyu Zhang, Aarash Feizi, Abhay Puri, Akshay Kalkunte, François Savard, Ahmed Masry, Shravan Nayak, et al. Bigdocs: An open dataset for training multimodal models on document and code tasks. *arXiv preprint arXiv:2412.04626*, 2024.

[274] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[275] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

[276] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7513–7522, 2024.

[277] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[278] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[279] Scale. Seal leaderboards. https://scale.com/leaderboard/, April 2025.

[280] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[281] Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, M. Jehanzeb Mirza, Leshem Chosen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. Livexiv – a multi-modal live benchmark based on arxiv papers content. *arXiv preprint arXiv:2410.10783*, 2024.

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

[283] Wenqi Shao, Meng Lei, Yutao Hu, Peng Gao, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, and Ping Luo. Tinylvlm-ehub: Towards comprehensive and efficient evaluation for large vision-language models. *arXiv preprint arXiv:2308.03729*, 2024.

[284] Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Won Park. Empathicstories++: A multimodal dataset for empathy towards personal experiences. *arXiv preprint arXiv:2405.15708*, 2024.

[285] Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024.

[286] Zhelun Shi, Zhipin Wang, Hongxing Fan, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. Chef: A comprehensive evaluation framework for standardized assessment of multimodal large language models, 2023.

[287] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[288] Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. *arXiv preprint arXiv:2403.03163*, 2024.

[289] Aditi Singh. A survey of ai text-to-image and ai text-to-video generators. In *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, pages 32–36. IEEE, 2023.

[290] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

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

[292] Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, and Yanyi Chen. A cognitive evaluation benchmark of image reasoning and description for large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6392–6409, 2025.

[293] Sitong Su, Xiao Cai, Lianli Gao, Pengpeng Zeng, Qinhong Du, Mengqi Li, Heng Tao Shen, and Jingkuan Song. Gt23d-bench: A comprehensive general text-to-3d generation benchmark. *arXiv preprint arXiv:2412.09997*, 2024. A 400 k-sample multimodal benchmark for Text-to-3D evaluation.

[294] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025.

[295] Ruoxi Sun, Jiamin Chang, Hammond Pearce, Chaowei Xiao, Bo Li, Qi Wu, Surya Nepal, and Minhui Xue. Sok: Unifying cybersecurity and cybersafety of multimodal foundation models with an information theory approach. *arXiv preprint arXiv:2411.11195*, 2024.

[296] Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark, 2024.

[297] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021.

[298] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[299] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024.

[300] Zichen Tang, Jiacheng Liu, Zhongjun Yang, Rongjin Li, Zihua Rong, Haoyang He, Zhuodi Hao, Xinyang Hu, Kun Ji, Ziyan Ma, et al. Finmmr: Make financial numerical reasoning more multimodal, comprehensive, and challenging. *arXiv preprint arXiv:2508.04625*, 2025.

[301] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[302] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

[303] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[304] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[305] Evan Trop, Yair Schiff, Edgar Mariano Marroquin, Chia Hsiang Kao, Aaron Gokaslan, McKinley Polen, Mingyi Shao, Bernardo P de Almeida, Thomas Pierrot, Yang I Li, et al. The genomics long-range benchmark: Advancing dna language models. 2024.

[306] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

[307] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[308] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13394, 2024.

[309] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models, 2025.

[310] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.

[311] Fengxiang Wang, Mingshuo Chen, Xuming He, YiFan Zhang, Feng Liu, Zijie Guo, Zhenghao Hu, Jiong Wang, Jingyi Xu, Zhangrui Li, et al. Omniearth-bench: Towards holistic evaluation of earth's six spheres and cross-spheres interactions with multimodal observational earth data. *arXiv preprint arXiv:2505.23522*, 2025.

[312] Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14325–14336, 2025.

[313] Hanbin Wang, Xiaoxuan Zhou, Zhipeng Xu, Keyuan Cheng, Yuxin Zuo, Kai Tian, Jingwei Song, Junting Lu, Wenhui Hu, and Xueyang Liu. Code-vision: Evaluating multimodal llms logic understanding and code generation capabilities. *arXiv preprint arXiv:2502.11829*, 2025.

[314] Hui Wang, Shiwan Zhao, Jiaming Zhou, Xiguang Zheng, Haoqin Sun, Xuechen Wang, and Yong Qin. Uncertainty-aware mean opinion score prediction. *arXiv preprint arXiv:2408.12829*, 2024.

[315] Jiarui Wang, Huiyu Duan, Ziheng Jia, Yu Zhao, Woo Yi Yang, Zicheng Zhang, Zijian Chen, Juntong Wang, Yuke Xing, Guangtao Zhai, et al. Love: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. *arXiv preprint arXiv:2505.12098*, 2025.

[316] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence*, pages 46–57. Springer, 2023.

[317] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Quality assessment for ai generated images with instruction tuning, 2025.

[318] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18869–18880, 2025.

[319] Juntong Wang, Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Tdve-assessor: Benchmarking and evaluating the quality of text-driven video editing with lmms. *arXiv preprint arXiv:2505.19535*, 2025.

[320] Junying Wang, Wenzhe Li, Yalun Wu, Yingji Liang, Yijin Guo, Chunyi Li, Haodong Duan, Zicheng Zhang, and Guangtao Zhai. Affordance benchmark for mllms, 2025.

[321] Junying Wang, Hongyuan Zhang, and Yuan Yuan. Adv-cpg: A customized portrait generation framework with facial adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21001–21010, June 2025.

[322] Junying Wang, Zicheng Zhang, Yijin Guo, Farong Wen, Ye Shen, Yingji Liang, Yalun Wu, Wenzhe Li, Chunyi Li, Zijian Chen, et al. The ever-evolving science exam. *arXiv preprint arXiv:2507.16514*, 2025.

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

[324] Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, et al. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*, 2025.

[325] Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19541–19551, 2025.

[326] Puyi Wang, Wei Sun, Zicheng Zhang, Jun Jia, Yanwei Jiang, Zhichao Zhang, Xiongkuo Min, and Guangtao Zhai. Large multi-modality model assisted ai-generated image quality assessment. *arXiv preprint arXiv:2404.17762*, 2024.

[327] Siyang Wang and Éva Székely. Evaluating text-to-speech synthesis from a large discrete token-based speech language model. *arXiv preprint arXiv:2405.09768*, 2024.

[328] Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. Qualispeech: A speech quality assessment dataset with natural language reasoning and descriptions. *arXiv preprint arXiv:2503.20290*, 2025.

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

[330] Siyuan Wang, Zhuohan Long, Zhihao Fan, and Zhongyu Wei. From llms to mllms: Exploring the landscape of multimodal jailbreaking. *arXiv preprint arXiv:2406.14859*, 2024.

[331] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024.

[332] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2025.

[333] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The all-seeing project v2: Towards general relation comprehension of the open world, 2024.

[334] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025.

[335] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12998–13008, 2024.

[336] Xiaoqin Wang, Xusen Ma, Xianxu Hou, Meidan Ding, Yudong Li, Junliang Chen, Wenting Chen, Xiaoyang Peng, and Linlin Shen. Facebench: A multi-view multi-level facial attribute vqa dataset for benchmarking face perception mllms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9154–9164, 2025.

[337] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

[338] Xihuai Wang, Ziyi Zhao, Siyu Ren, Shao Zhang, Song Li, Xiaoyu Li, Ziwen Wang, Lin Qiu, Guanglu Wan, Xuezhi Cao, et al. Audio Turing Test: Benchmarking the human-likeness of large language model-based text-to-speech systems in chinese. *arXiv preprint arXiv:2505.11200*, 2025.

[339] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

[340] Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*, 2025.

[341] Yuhao Wang, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yu Wang, and Yanfeng Wang. Mm-sap: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. *arXiv preprint arXiv:2401.07529*, 2024.

[342] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[343] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

[344] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

[345] Zhaoyang Wei, Chenhui Qiang, Bowen Jiang, Xumeng Han, Xuehui Yu, and Zhenjun Han. Adˆ2-bench: A hierarchical cot benchmark for mllm in autonomous driving under adverse conditions. *arXiv preprint arXiv:2506.09557*, 2025.

[346] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[347] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Explorations of utilizing visual-language model as autonomous driving agent. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[348] Fenghua Weng, Yue Xu, Chengyan Fu, and Wenjie Wang. \textit {MMJ-Bench}: A comprehensive study on jailbreak attacks and defenses for vision language models. *arXiv e-prints*, pages arXiv–2408, 2024.

[349] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark, 2025.

[350] Chengyue Wu, Yixiao Ge, Qiushan Guo, Jiahao Wang, Zhixuan Liang, Zeyu Lu, Ying Shan, and Ping Luo. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. *arXiv preprint arXiv:2405.07990*, 2024.

[351] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024.

[352] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

[353] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. 2023.

[354] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. *arXiv preprint arXiv:2311.06783*, 2023.

[355] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[356] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. *arXiv preprint arXiv:2402.16641*, 2024.

[357] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. *arXiv preprint arXiv:2402.16641*, 2024.

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

[359] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[360] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[361] Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*, 2023.

[362] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22227–22238, 2024.

[363] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[364] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105.

[365] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024.

[366] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models, 2024.

[367] Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. KRIS-Bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025.

[368] xAI Team. Grok-3: The Age of Reasoning Agents. Technical report, xAI, February 2025.

[369] Jili Xia, Lihuo He, Fei Gao, Kaifan Zhang, Leida Li, and Xinbo Gao. Ai-generated image quality assessment based on task-specific prompt and multi-granularity similarity, 2024.

[370] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024.

[371] Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024.

[372] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning, 2025.

[373] Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. Seephys: Does seeing help thinking? benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*, 2025.

[374] Baihui Xiao, Chengjian Feng, Zhijian Huang, Yujie Zhong, Lin Ma, et al. Robotron-sim: Improving real-world driving via simulated hard-case. *arXiv preprint arXiv:2508.04642*, 2025.

[375] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[376] Shangyu Xing, Changhao Xiang, Yuteng Han, Yifan Yue, Zhen Wu, Xinyu Liu, Zhangtai Wu, Fei Zhao, and Xinyu Dai. Gepbench: Evaluating fundamental geometric perception for multimodal large language models. *arXiv preprint arXiv:2412.21036*, 2024.

[377] Yuke Xing, Jiarui Wang, Peizhi Niu, Wenjie Huang, Guangtao Zhai, and Yiling Xu. 3dgs-ieval-15k: A large-scale image quality evaluation database for 3d gaussian-splatting. *arXiv preprint arXiv:2506.14642*, 2025.

[378] Yuke Xing, Qi Yang, Kaifa Yang, Yiling Xu, and Zhu Li. Explicit-nerf-qa: A quality assessment database for explicit nerf model compression. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2024.

[379] Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, et al. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*, 2025.

[380] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[381] Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*, 2023.

[382] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

[383] Wanghan Xu, Xiangyu Zhao, Yuhao Zhou, Xiaoyu Yue, Ben Fei, Fenghua Ling, Wenlong Zhang, and Lei Bai. Earthse: A benchmark evaluating earth scientific exploration capability for large language models. *arXiv preprint arXiv:2505.17139*, 2025.

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

[385] Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. Famma: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint arXiv:2410.04526*, 2025.

[386] Chao Yang, Chaochao Lu, Yingchun Wang, and Bowen Zhou. Towards ai-45° law: A roadmap to trustworthy agi. *arXiv preprint arXiv:2412.14186*, 2024.

[387] Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024.

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

[389] Junfeng Yang, Jing Fu, Wei Zhang, Wenzhi Cao, Limei Liu, and Han Peng. Moe-agiqa: Mixture-of-experts boosted visual perception-driven and semantic-aware quality assessment for ai-generated images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6395–6404, 2024.

[390] Liu Yang, Huiyu Duan, Long Teng, Yucheng Zhu, Xiaohong Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Aigcoiqa2024: Perceptual quality assessment of ai generated omnidirectional images. *arXiv preprint arXiv:2404.01024*, 2024.

[391] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension, 2024.

[392] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents, 2025.

[393] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation, 2023.

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

[395] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025.

[396] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pages 232–248. Springer, 2025.

[397] Ming Yin, Yuanhao Qu, Dyllan Liu, Ling Yang, Le Cong, and Mengdi Wang. Genome-bench: A scientific reasoning benchmark from real-world expert discussions. *bioRxiv*, pages 2025–06, 2025.

[398] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[399] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. 2024.

[400] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

[401] Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*, 2024.

[402] Zonghao Ying, Aishan Liu, Xianglong Liu, and Dacheng Tao. Unveiling the safety of gpt-4o: An empirical study using jailbreak attacks. *arXiv preprint arXiv:2406.06302*, 2024.

[403] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pages 259–276. Springer, 2024.

[404] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016.

[405] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. PACS: A dataset for physical audiovisual commonsense reasoning. *arXiv preprint arXiv:2203.11130*, 2022.

[406] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024.

[407] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen. Sf-iqa: Quality and similarity integration for ai generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6692–6701, 2024.

[408] 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, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech, 2024.

[409] Jiquan Yuan, Xinyan Cao, Linjing Cao, Jinlong Lin, and Xixin Cao. Pscr: Patches sampling-based contrastive regression for aigc image quality assessment, 2023.

[410] Jiquan Yuan, Xinyan Cao, Jinming Che, Qinyuan Wang, Sen Liang, Wei Ren, Jinlong Lin, and Xixin Cao. Tier: Text-image encoder-based regression for aigc image quality assessment, 2024.

[411] Jiquan Yuan, Xinyan Cao, Changjin Li, Fanyi Yang, Jinlong Lin, and Xixin Cao. Pku-i2iqa: An image-to-image quality assessment database for ai generated images, 2023.

[412] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

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

[414] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[415] Sukmin Yun, Rusiru Thushara, Mohammad Bhat, Yongxin Wang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, Haonan Li, Preslav Nakov, et al. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. *Advances in neural information processing systems*, 37:112134–112157, 2024.

[416] Lingfeng Zeng, Fangqi Lou, Zixuan Wang, Jiajie Xu, Jinyi Niu, Mengping Li, Yifan Dong, Qi Qi, Wei Zhang, Ziwei Yang, et al. Fingaia: An end-to-end benchmark for evaluating ai agents in finance. *arXiv preprint arXiv:2507.17186*, 2025.

[417] Ryandhimas E Zezario, Yu-Wen Chen, Szu-Wei Fu, Yu Tsao, Hsin-Min Wang, and Chiou-Shann Fuh. A study on incorporating whisper for robust speech assessment. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2024.

[418] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63(11):211301, 2020.

[419] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[420] Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. arxiv. 2023.

[421] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[422] Chiyu Zhang, Lu Zhou, Xiaogang Xu, Jiafei Wu, and Zhe Liu. Adversarial attacks of vision tasks in the past 10 years: A survey. *ACM Computing Surveys*, 2024.

[423] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

[424] Fengji Zhang, Linquan Wu, Huiyu Bai, Guancheng Lin, Xiao Li, Xiao Yu, Yue Wang, Bei Chen, and Jacky Keung. Humaneval-v: Benchmarking high-level visual reasoning with complex diagrams in coding tasks. *arXiv preprint arXiv:2410.12381*, 2024.

[425] Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions. *CoRR*, 2024.

[426] Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, and Xiaojun Wan. Mc-mke: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. *arXiv preprint arXiv:2406.13219*, 2024.

[427] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.

[428] Linhao Zhang, Daoguang Zan, Quanshun Yang, Zhirong Huang, Dong Chen, Bo Shen, Tianyu Liu, Yongshun Gong, Pengjie Huang, Xudong Lu, et al. Codev: Issue resolving with visual data. *arXiv preprint arXiv:2412.17315*, 2024.

[429] Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. CBLUE: A chinese biomedical language understanding evaluation benchmark.
urlhttps://github.com/CBLUEbenchmark/CBLUE, 2021.

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

[431] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024.

[432] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.

[433] Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. Vq-insight: Teaching vlms for ai-generated video quality understanding via progressive visual reinforcement learning. *arXiv preprint arXiv:2506.18564*, 2025.

[434] Yan Zhang, Zhong Ji, Yanwei Pang, Jungong Han, and Xuelong Li. Modality-experts coordinated adaptation for large multimodal models. *Science China Information Sciences*, 67(12):220107, 2024.

[435] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[436] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv e-prints*, pages arXiv–2406, 2024.

[437] Yuhan Zhang, Mengchen Zhang, Tong Wu, Tengfei Wang, Gordon Wetzstein, Dahua Lin, and Ziwei Liu. 3dgen-bench: Comprehensive benchmark suite for 3d generative models. *arXiv preprint arXiv:2503.21745*, 2025.

[438] Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. Negvqa: Can vision language models understand negation? *arXiv preprint arXiv:2505.22946*, 2025.

[439] Yujie Zhang, Bingyang Cui, Qi Yang, Zhu Li, and Yiling Xu. Benchmarking and learning multi-dimensional quality evaluator for text-to-3d generation. *arXiv preprint arXiv:2412.11170*, 2024.

[440] Zhichao Zhang, Wei Sun, Xinyue Li, Yunhao Li, Qihang Ge, Jun Jia, Zicheng Zhang, Zhongpeng Ji, Fengyu Sun, Shangling Jui, et al. Human-activity agv quality assessment: A benchmark dataset and an objective evaluation metric. *arXiv preprint arXiv:2411.16619*, 2024.

[441] Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content, 2025.

[442] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 440–445. IEEE, 2023.

[443] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai. No-reference quality assessment for 3d colored point cloud and mesh models. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7618–7631, 2022.

[444] Zicheng Zhang, Wei Sun, Xiongkuo Min, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai. Mm-pcqa: Multi-modal learning for no-reference point cloud quality assessment. *arXiv preprint arXiv:2209.00244*, 2022.

[445] Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, et al. Aibench: Towards trustworthy evalua-tion under the 45 law. 2025.

[446] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al. Q-boost: On visual quality assessment ability of low-level multi-modality foundation models. *arXiv preprint arXiv:2312.15300*, 2023.

[447] Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmms masters at evaluating ai-generated images?, 2024.

[448] Zicheng Zhang, Haoning Wu, Yingjie Zhou, Chunyi Li, Wei Sun, Chaofeng Chen, Xiongkuo Min, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. LMM-PCQA: Assisting point cloud quality assessment with large multimodal models. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1234–1243. ACM, 2024.

[449] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[450] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

[451] Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175, 2024.

[452] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, et al. Towards video text visual question answering: Benchmark and baseline. *Advances in Neural Information Processing Systems*, 35:35549–35562, 2022.

[453] Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37:7185–7212, 2024.

[454] Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. Msearth: A benchmark for multimodal scientific comprehension of earth science. *arXiv preprint arXiv:2505.20740*, 2025.

[455] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

[456] Zhaoran Zhao, Peng Lu, Anran Zhang, Peipei Li, Xia Li, Xuannan Liu, Yang Hu, Shiyi Chen, Liwei Wang, and Wenhao Guo. Can machines understand composition? dataset and benchmark for photographic image composition embedding and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14411–14421, 2025.

[457] Baolin Zheng, Guanlin Chen, Hongqiong Zhong, Qingyang Teng, Yingshui Tan, Zhendong Liu, Weixun Wang, Jiaheng Liu, Jian Yang, Huiyun Jing, et al. Usb: A comprehensive and unified safety evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2505.23793*, 2025.

[458] Qiaoyu Zheng, Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Lisong Dai, Hengyu Guan, Yuehua Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-scale long-tailed disease diagnosis on radiology images. *Nature Communications*, 15(1):10147, 2024.

[459] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

[460] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10707–10715, 2025.

[461] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[462] Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, et al. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*, 2024.

[463] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 56–66, 2025.

[464] Pengfei Zhou, Fanrui Zhang, Xiaopeng Peng, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, et al. Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. *arXiv preprint arXiv:2504.05782*, 2025.

[465] Tianwei Zhou, Songbai Tan, Wei Zhou, Yu Luo, Yuan-Gen Wang, and Guanghui Yue. Adaptive mixed-scale feature fusion network for blind ai-generated image quality assessment, 2024.

[466] Yingjie Zhou, Jiezhang Cao, Zicheng Zhang, Farong Wen, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Who is a better talker: Subjective and objective quality assessment for ai-generated talking heads. arXiv preprint arXiv:2507.23343, 2025.

[467] Yingjie Zhou, Yaodong Chen, Kaiyue Bi, Lian Xiong, and Hui Liu. An implementation of multimodal fusion system for intelligent digital human generation. arXiv preprint arXiv:2310.20251, 2023.

[468] Yingjie Zhou, Xiaohong Liu, et al. Mi3s: A multimodal large language model assisted quality assessment framework for ai-generated talking heads. *Information Processing & Management*, 2025.

[469] Yingjie Zhou, Zicheng Zhang, Jiezhang Cao, Jun Jia, Yanwei Jiang, Farong Wen, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Memo-bench: A multiple benchmark for text-to-image and multimodal large language models on human emotion analysis. *arXiv preprint arXiv:2411.11235*, 2024.

[470] Yingjie Zhou, Zicheng Zhang, Jun Jia, Guangtao Zhai, et al. Who is a better imitator: Subjective and objective quality assessment of animated humans. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[471] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Zhihua Wang, Xiao-Ping Zhang, and Guangtao Zhai. Thqa: A perceptual quality assessment database for talking heads. *arXiv preprint arXiv:2404.09003*, 2024. 800 talking head videos, MOS annotations, dataset publicly released.

[472] Yingjie Zhou, Zicheng Zhang, Farong Wen, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 3dgcqa: A quality assessment database for 3d ai-generated contents. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[473] Yuhao Zhou, Yiheng Wang, Xuming He, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, Wenxuan Huang, Jiaqi Wei, Dan Si, Xiuqi Yao, Jia Bu, Haiwen Huang, Tianfan Fu, Shixiang Tang, Ben Fei, Dongzhan Zhou, Fenghua Ling, Yan Lu, Siqi Sun, Chenhui Li, Guanjie Zheng, Jiancheng Lv, Wenlong Zhang, and Lei Bai. Scientists' first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning, 2025.

[474] Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.

[475] Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.

[476] Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*, 2024.

[477] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37:32611–32629, 2024.

[478] Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. Cospace: Benchmarking continuous space perception ability for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29569–29579, 2025.

[479] Henry Peng Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip S Yu, and Cornelia Caragea. Implicitave: An open-source dataset and multimodal llms benchmark for implicit attribute value extraction. *arXiv preprint arXiv:2404.15592*, 2024.

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