SCIENCE CHINA
Information Sciences

• REVIEW •

December 2025, Vol. 68, Iss. 12, 221301:1–221301:40
https://doi.org/10.1007/s11432-025-4676-4

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[1], Yuqin CAO[1], 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*]

[1]*Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China*
[2]*Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[3]*College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China*
[4]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China*
[5]*Department of Psychiatry, Harvard University, Cambridge MA 02138, USA*
[6]*PengCheng Laboratory, Shenzhen 518055, China*
[7]*School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore*
[8]*Distributed & Interactive Systems Group, Delft University of Technology, Delft CN 2628, Netherlands*
[9]*Computer Science, University of British Columbia, Vancouver BC V6T 1Z4, Canada*
[10]*School of Computer Science and Informatics, Cardiff University, Cardiff CF10 3AT, UK*
[11]*College of Environmental Design, University of California, Berkeley, Berkeley CA 94720, USA*
[12]*Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China*
[13]*College of Engineering, Peking University, Beijing 100871, China*
[14]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[15]*Department of Computer Science and Engineering (CSE), The Chinese University of Hong Kong, Hong Kong 999077, China*
[16]*Ecole Polytechnique de l'Université de Nantes, Nantes Université, Nantes 44035, France*
[17]*School of Communication and Electronic Engineering, East China Normal University, Shanghai 200062, China*
[18]*Data Science and AI (DSAI), Monash University, Clayton VIC 3800, Australia*
[19]*Huawei Technologies Co., Ltd., Shanghai 200120, China*

Received 27 August 2025/Revised 25 September 2025/Accepted 8 November 2025/Published online 18 November 2025

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.

Keywords large multimodal models, multimodal understanding, multimodal generation

Citation Zhang Z C, Wang J Y, Wen F R, et al. Large multimodal models evaluation: a survey. Sci China Inf Sci, 2025, 68(12): 221301, https://doi.org/10.1007/s11432-025-4676-4

* Corresponding author (email: zhaiguangtao@sjtu.edu.cn)
† These authors contributed equally to this work.

© Science China Press 2025

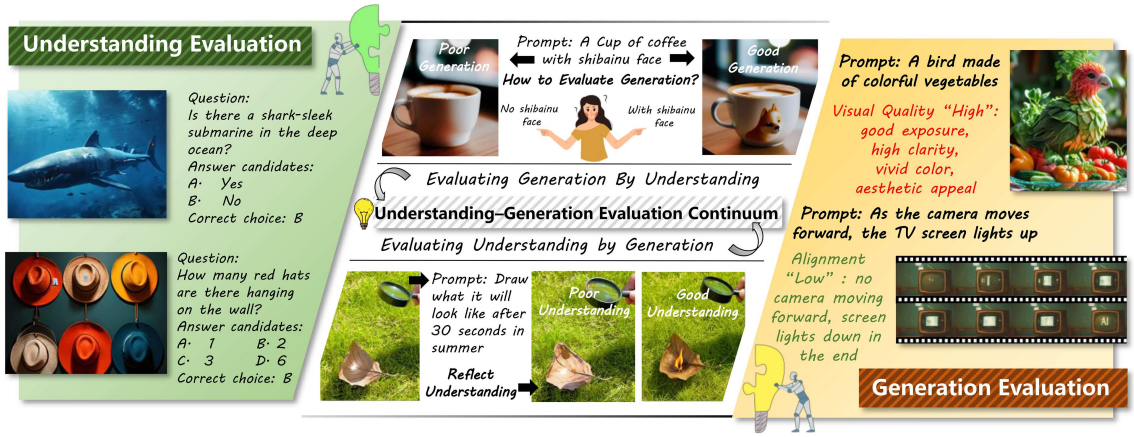

Figure 1 (Color online) Understanding-generation evaluation continuum. Understanding evaluation focuses on assessing LMM performance usually through question-answering accuracy, while generation evaluation emphasizes the quality of generated content. A growing trend reveals the convergence of these paradigms, where understanding can facilitate generation evaluation, and generation serves as a proxy for evaluating understanding.

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, marks a significant milestone in the development of artificial intelligence. These models (such as GPT [1, 2], Gemini [3], Grok [4]) unify vision, language, or other modalities under a shared framework, enabling a wide range of applications from multimodal understanding (image captioning [5], visual question answering [6]) to multimodal generation (text-to-image/video generation [7], text-to-video [8]). Therefore, LMM evaluation can be categorized into two pillars: understanding evaluation, which measures the ability of the 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 [9], emphasizing versatility across modalities, tasks, and domains. These capabilities include common tasks like instruction-following, dialog comprehension, and general visual grounding. However, as real-world demands increase, there is a growing emphasis on specialized capabilities [10], 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 [11]. Moreover, generation quality must be assessed with respect to both visual quality (e.g., technical 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 is bringing these paradigms closer. As shown in Figure 1, generation evaluation often relies on a 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.

In parallel with the development of benchmark and quality assessment, the broader ecosystem for LMM evaluation has been supported by the rapid emergence of community-driven leaderboards and open-source evaluation tools [10, 12–14], which have evolved into central platforms for aggregating evaluation results, tracking performance across models, and promoting reproducibility. Meanwhile, these evaluation tools

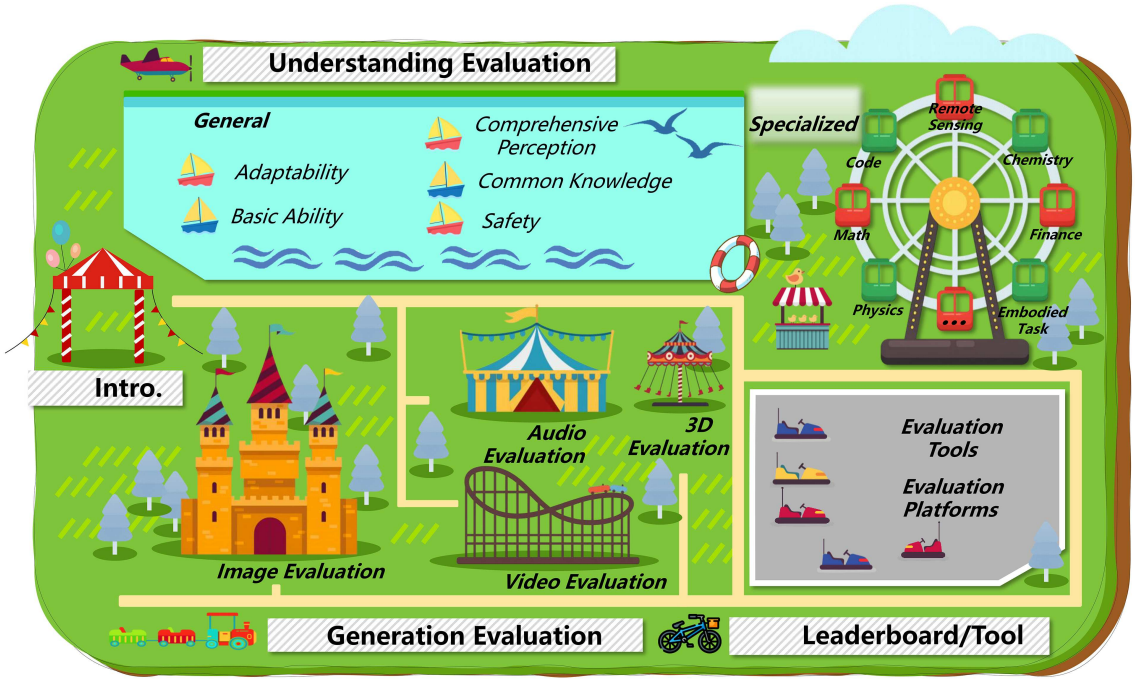

Figure 2 (Color online) Content overview of the survey.

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 [15–19]. 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 Figure 2.

2 Evaluation for understanding

Understanding evaluation (representative benchmarks are shown in Figure 3) 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)].

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

1) The 'VQA' in Sections 2 and 3 refers to visual question answering and video quality assessment, respectively.

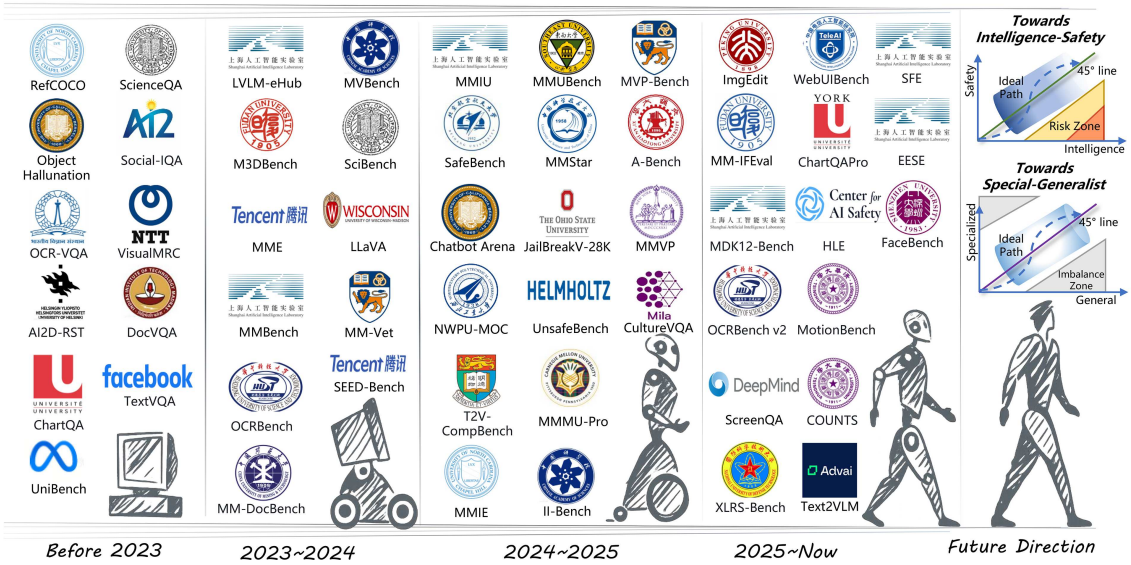

Figure 3 (Color online) Representative benchmarks for understanding evaluation. As the field evolves, benchmark development is increasingly emphasizing the synchronized advancement of intelligence and safety, as well as comprehensive assessment spanning specialized and generalist capabilities.

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 [20,21]. 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 [22]. LLaVA-Bench [23] 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 [24] 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 [25] 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 [26] 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 [27] 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 [28] 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 [29] 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 [30] targets sequential reasoning by introducing 4.7k dynamic image sequences, assessing whether models capture temporal changes and behavioral dynamics beyond static perception. MuirBench [31] expands coverage to 12 tasks and 10 relational types with over 11k images, pairing each instance with unanswerable variants to diagnose robustness. MMIU [32], 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 [33] pioneers relational reasoning benchmarks across four dimensions, requiring comparative analysis of up to 42 images in a task. Subsequent efforts such as MIBench [34] and II-Bench [35] emphasize higher-order perception and knowledge-seeking, while Mantis-Eval [36] bridges single- and multi-image reasoning in one unified benchmark. MileBench [37] highlights performance degradation under long input sequences, and ReMI [38] 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 [39] evaluates context-dependent comprehension by pairing images with contradictory textual cues, requiring models to dynamically reinterpret visual content. Sparkles [40] 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 a larger scale, MMIE [41] offers 20k queries across 12 fields, unifying open-ended and multiple-choice tasks, while InterleavedBench [42] supports arbitrary multimodal input-output orders and introduces InterleavedEval, a reference-free metric for text quality, perceptual fidelity, and cross-modal coherence. OpenING [43] 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 [44] and HERM-Bench [45] foreground fairness, inclusivity, and ethical reasoning in multimodal settings. UNIAA [46] and HumanBeauty [47] shift focus to aesthetics, drawing on large-scale human ratings to quantify visual appeal. Social-IQA [48] emphasizes social commonsense reasoning, requiring inference of motives, emotions, and event outcomes, while EmpathicStories [49] examines empathy through multimodal narratives with explicit emotional cues. In addition to task-specific datasets, community platforms provide direct human feedback. Chatbot Arena [50] implements double-blind pairwise comparison, where users vote on dialogue quality without model identity, thus capturing preference at scale. OpenAssistant Conversations [51] crowdsources human feedback from global volunteers in multi-turn dialogues, incorporating quality ratings and preference rankings. HCE [52] 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 [53] 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 [54] extends counting into the video domain, assessing compositional text-to-video generation abilities such as attribute binding and spatio-temporal consistency, while ConceptMIX [55] automates compositional evaluation for text-to-image models via generated prompts and VLM-based verification. PICD [56] 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 [57] frame recognition as question answering over images requiring the correct interpretation of embedded text, and OCR-VQA [58] scales this paradigm to 207.5k book-cover images with over one million QA pairs. Consolidated suites like OCRBench [59] integrate 29 datasets covering text recognition, scene-text VQA, document VQA, key information extraction, and handwritten mathematical expression recognition, while OCRBench v2 [60] quadruples task diversity (31 scenarios) and adds 10k human-verified QA pairs, introducing challenges such as text localization and logical reasoning. ASCIIEval [61] extends the text recognition to arts formulated in text strings. Recent benchmarks focus on reasoning traces and modality interaction: OCR-Reasoning [62] annotates both answers and reasoning processes across six abilities, and M4-ViteVQA [63] extends evaluation to video contexts, testing spatio-temporal grounding. SEED-Bench-2-Plus [64] 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 [66] 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 [67] emphasizes multi-page reasoning over 135 lengthy PDFs (1k expert questions), with 33.7% cross-page queries and 20.6% hallucination-detection items. UDA [68] 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 [69] (10k+ images and 30k+ QA) and DocVQA [70] (50k questions and 12k images). Scaling up, DocGenome [71] annotates 500k scientific documents with structured multimodal data, while GDI-Bench [72] 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 [73] offers 715k device-control episodes with 30k instructions, supporting multi-step gesture inference from demonstrations and language. ScreenSpot [74] evaluates GUI element grounding across mobile, desktop, and web platforms, while VisualWebBench [75] covers seven tasks with 1.5k curated instances from 139 websites (87 sub-domains). GUI-World [76] introduces dynamic and sequential content-conditions that often degrade LMM performance. Extending to programmatic competence, WebUIBench [77] 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 [78] 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 [79] establishes a large-scale foundation with human-written and generated questions targeting logical and arithmetic reasoning, while ChartQAPro [80] diversifies chart types (dashboards, infographics) and introduces unanswerable queries for robustness testing. Table-specific benchmarks include ComTQA [81] (about 9k QA) and TableVQA-Bench [82] (1.5k QA from generated tables), which reveal that visual processing remains more challenging than text-only inputs. CharXiv [83] (2.3k scientific charts) and SciFIBench [84] (2k figure questions) provide high-difficulty scientific contexts, while AI2D-RST [85] augments diagram QA with multi-layer discourse structure annotations. InfoChartQA [86] (5.6k infographic/plain chart pairs) and EvoChart-QA [87] (650 charts and 1.2k expert questions) expose robustness gaps under non-canonical designs. WikiMixQA [88] extends to cross-modal reasoning over tables and charts, combining 1k multiple-choice questions from 4k Wikipedia pages. ChartX [89] 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 [90] 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 [91] further pushes forward the low-level perception for LMMs on AIGC images. MVP-Bench [92] 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 [93] targets ultra-high-resolution remote sensing scenarios, defining 16 sub-tasks across 10 perceptual and six reasoning capabilities, with the largest average image size to date. HR-Bench [94] 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 [95] pushes realism further, offering 29k+ manually annotated QA pairs over diverse high-resolution, real-world scenes. V* Bench [96] 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 [97] addresses comprehensive face perception, cataloging over 210 attributes and nearly 50k QA pairs. MMAFFBen [98] 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 [99] jointly evaluates recognition and generation for fine-grained facial affective behaviors such as action unit recognition. Emotion-oriented datasets like MEMO-Bench [100], EmoBench [101], and EEmo-Bench [102] emphasize progressively fine-grained sentiment assessment. For aesthetic evaluation, AesBench [103] and UNIAA-Bench [46] offer structured frameworks across perception, empathy, assessment, and interpretation, while ImplicitAVE [104] and II-Bench [35] target implicit and abstract attribute extraction. CogBench [105] assesses comprehensive dimensions including time, location, character, event, and mental state, through images with rich reasoning chains. A4Bench [106] 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 [107] 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 [108] provides a vision-centric evaluation over 15 visual representations, alongside CV-Bench (2.6k VQA questions) for fine-grained capability diagnosis. MMUBench [109] pioneers machine unlearning evaluation for LMMs, measuring forgetting efficacy, generality, specificity, fluency, and diversity. MMVP [110] 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 [111] 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 the object or part level. The RefCOCO family [112, 113] 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 [114] updates these with higher annotation accuracy, a larger vocabulary, and longer expressions (average 24.2 words). MRES-32M [115] expands to multi-granularity segmentation with over 32.2M masks, while UrBench [116] broadens grounding to complex multi-view

urban scenarios with 11.6k questions across 14 task types. COUNTS [117] explicitly targets out-of-distribution generalization for object detectors and LMMs, introducing 14 natural distributional shifts with object-level annotations. MTVQA [118] 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 [119] pioneers large-scale geometric perception evaluation with 80k figures and 285k multiple-choice questions, highlighting foundational gaps in shape and structure comprehension. SpatialMQA [120] introduces 5.3k human-annotated samples over COCO2017 to test spatial relation understanding, while SpatialRGPT-Bench [121] extends the challenge to 3D cognition using indoor, outdoor, and simulated environments with precise ground-truth annotations. CoSpace [122] probes continuous space perception from sequences of spatially consistent images, complementing static-scene evaluations. LMM-CompBench [123] addresses comparative reasoning by assembling ∼40k image pairs for relative judgments across eight dimensions. SOK-Bench [124] integrates situated and open-world knowledge for video reasoning, while GSR-Bench [125] and What'sUp [126] target specific spatial relations such as object positioning and orientation. Q-Spatial Bench [127] 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 [128] 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 [129] exemplifies this trend by annotating 98k questions with explicit reasoning steps and bounding boxes over key visual regions, enabling interpretability analysis. LogicVista [130] evaluates five logical reasoning tasks spanning nine capabilities using 448 multiple-choice questions, while VisuLogic [131] mitigates language shortcuts by designing 1k human-verified problems focusing on genuine vision-centric inference. CoMT [132] expands the scope to visual creation, deletion, update, and selection-four categories that simulate complex visual operations. PUZZLES [133], 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 [134] equips models with the capacity to pose and evaluate questions in visual settings, with its EvalQABench containing 64k training and 5k testing samples. VLKEB [135] builds on multi-modal knowledge graphs to assess knowledge editing portability-whether changes apply consistently across relevant content. MMKE-Bench [136] 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 [137, 138], which decomposes multimodal knowledge into visual and textual components to isolate and fix misreadings or misrecognitions. NegVQA [139] 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 un-

derstanding through diverse image-based tasks. LVLM-eHub [140] offers a large-scale suite spanning six categories, from visual question answering to embodied AI, while its lightweight variant TinyLVLM-eHub [141] condenses evaluation into 2.1k image-text pairs for quick testing. Other large-scale frameworks expand coverage: LAMM [142] bridges 2D and 3D with 12 tasks, MME [143] incorporates 14 subtasks from object recognition to code understanding, and MMBench [144] stabilizes grading by converting free-form responses into multiple-choice format. The SEED-Bench series [64, 145, 146] scales this further across multimodal QA and text-rich scenarios, while MMT-Bench [147] and LMMs-Eval [13, 148] integrate dozens of tasks across modalities for unified comparison. To address potential biases in such broad suites, MMStar [149] curates 1.5k vision-indispensable samples and further introduces evaluation metrics like multi-modal gain and leakage, while NaturalBench [150] seeks to minimize language priors by constructing 10k adversarial human-verified samples. Finally, MM-Vet [151] and ChEF [152] 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 [153] provides one of the earliest large-scale benchmarks, with 900 videos spanning 254 h and 2.7k QA pairs, exposing deficiencies in temporal fusion. MMBench-Video [154] complements this with long-form YouTube videos and free-form QA, introducing ability-based categorization and GPT-4-based scoring. MVBench [155] emphasizes the evaluation of short-to-medium video reasoning, while LongVideo-Bench [156] pushes the scale to hour-long content with a novel referring reasoning task that requires models to locate and analyze specific temporal contexts. LVBench [157] further broadens this scope to extreme-length videos lasting several hours, thereby highlighting persistent weaknesses in long-term memory. MotionBench [158], 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 [159] 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 [160] 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 [161] 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 [162] 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 [163] covers eight tasks in 3D medical imaging, enabling systematic evaluation of multimodal medical reasoning. Space3D-Bench [164] 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 levels below high school.

ScienceQA [165] 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 [166] 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 [167] 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 [168] 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 [169] 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 [170] 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 [171] 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 [172] 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 [173] 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 [41] 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 [174] 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 [175] 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 [176–181]. 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. Representative safety benchmarks for LMMs

Table 1 Brief description of representative safety benchmarks for LMMs.

Safety Concern	Benchmark	Description
Jailbreak and adversarial robustness	Unicorn [182]	Systematic evaluation of jailbreak risks in vision-language models.
	JailbreakV-28K [183]	Large-scale dataset of 28K jailbreak samples covering diverse attack vectors.
	MM-SafetyBench [184]	Evaluates the impact of images on jailbreak attacks using multimodal systems.
	AVIBench [185]	Investigates adversarial visual instructions triggering unsafe responses.
	MMJ-Bench [186]	Unified platform for evaluating jailbreak attacks and defense mechanisms.
Comprehensive safety evaluations	USB [187]	Consolidates diverse safety evaluation tasks into a single benchmark suite.
	MLLMGuard [188]	Bilingual multimodal evaluation suite for safety in multilingual contexts.
	SafeBench [189]	Unified testing pipeline including both text and image modalities for safety.
	MemeSafetyBench [190]	Focuses on risks from internet culture and socially charged multimodal content.
	UnsafeBench [191]	Evaluates unsafe image classification on human-created and AI-generated images.
Hallucination and truthfulness	POPE [192]	Framework for detecting and mitigating hallucinations in LMMs.
	M-HalDetect [193]	Framework for detecting hallucinations in LMM outputs.
	Hal-Eval [194]	Fine-grained benchmark for hallucination detection.
	Hallu-pi [195]	Examines robustness under perturbed inputs.
	BEAF [196]	Introduces before-after comparison for measuring response stability.
	HallusionBench [197]	Large-scale hallucination test case generation.
	AutoHallusion [198]	Automatic generation of hallucination test cases.
	MultiTrust [199]	Benchmarks multiple aspects of model trustworthiness.
	MMDT [200]	Investigates safety at the decoding level.
Fairness and bias	CulturalVQA [201]	Curates image-question pairs from 11 countries, testing cultural understanding.
	ModScan [202]	Measures stereotype bias across vision and language modalities.
	FMBench [203]	Domain-specific benchmark for fairness in medical LMMs.
	FairMedFM [204]	Benchmarks fairness in medical LMMs.
	FairCLIP [205]	Integrates fairness considerations directly into model training.
Emerging safety concerns	DoxingBench [206]	Evaluates privacy leakage and how models might reveal user locations from images.
	PrivQA [207]	Investigates models' ability to follow privacy-preserving instructions.
	SHIELD [208]	Evaluates forgery and face-spoofing detection, focusing on deepfakes.
	ExtremeAIGC [209]	Benchmarks the vulnerability of LMMs to extremist AI-generated content.

are shown in Table 1.

(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 [182] provide the first systematic evaluation of jailbreak risks in vision-language models. Building on this, JailbreakV-28K [183] introduces a large-scale dataset of 28k jailbreak samples covering diverse attack vectors, establishing a more comprehensive baseline for measuring susceptibility. MM-SafetyBench [184] 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. [210] traced the landscape evolution from traditional text-only jailbreak strategies to multimodal ones, highlighting new risks introduced by visual prompts.

Recent studies emphasize increasingly sophisticated attack strategies. AVIBench [185] investigates adversarial visual instructions, where subtle changes in images can successfully trigger unsafe responses. MMJ-Bench [186] provides a unified platform for evaluating both jailbreak attacks and defense mechanisms. Guo et al. [211] introduced 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 [212]. 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) [187] exemplifies this trend by consolidating diverse safety evaluation tasks into a single benchmark suite. Similarly, MLLMGuard [188] offers a bilingual multimodal evaluation suite, addressing the need for safety evaluation in multi-lingual contexts. Benchmarks such as SafeBench [189] and MM-SafetyBench [213] 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 [190] focuses on the unique risks of internet culture, curating over 50k memes to evaluate how LMMs handle socially charged multimodal content. UnsafeBench [191] 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. [214] analyzed object hallucination in image captioning, laying the foundation for systematic evaluation. Subsequent studies, including POPE [192] and M-HalDetect [193], introduce frameworks for detecting and mitigating hallucinations. More recent contributions expand diagnostic coverage: Hal-Eval [194] develops a fine-grained benchmark, Hallu-pi [195] examines robustness under perturbed inputs, and BEAF [196] introduces before-after comparison to measure response stability. Large-scale benchmarks such as HallusionBench [197] and AutoHallusion [198] emphasize scalability by automatically generating hallucination test cases.

Beyond hallucination detection, researchers have also explored trustworthiness in a broader sense. MultiTrust [199] benchmarks multiple aspects of model trustworthiness, while MMDT [200] investigates safety at the decoding level. Dataset adaptation methods such as Text2VLM [215] extend text-only safety evaluation resources to multimodal models, and MOSSBench [216] 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 [217] proposed one of the first comprehensive multimodal bias evaluation frameworks, expanding beyond traditional axes of gender and race. To address cultural diversity, CulturalVQA [201] curates over 2.3k image-question pairs from 11 countries across 5 continents, testing the cultural understanding of LMMs. ModScan [202] provides a structured approach to measure stereotype bias jointly across vision and language modalities.

In healthcare, where fairness is especially critical, FMBench [203] and FairMedFM [204] provide domain-specific benchmarks for evaluating fairness in medical LMMs. Beyond evaluation, fairness-aware training approaches such as FairCLIP [205] 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 [206] demonstrates how multimodal agentic models may inadvertently reveal user locations from images, while PrivQA [207] investigates the ability of models to follow privacy-preserving instructions. In the context of misinformation and manipulation, SHIELD [208] evaluates forgery and face-spoofing detection, addressing the challenges of deepfakes. Meanwhile, ExtremeAIGC [209] 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 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 Figure 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 [218] integrates challenging problems from 31 multimodal datasets and further introduces curated subsets such as IQTest, FunctionQA, and PaperQA. PolyMATH (Gupta et al., 2024) [219] aggregates

Figure 4 (Color online) Quick references to the representative specialized benchmarks.

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) [220] collects problems from real math competitions, spanning 16 mathematical disciplines and five difficulty levels to emphasize authenticity and fine-grained control. Olympiad-Bench [221] extends to Olympiad-level bilingual problems in mathematics and physics, targeting scientific reasoning beyond pure math. PolyMath (Wang et al., 2025) [222] 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 [223] collects diagram-based problems and systematically transforms each into six versions, enabling modality ablation and robustness analysis. WE-MATH [224] 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 [225] 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 [226] is a Chinese multimodal benchmark spanning 12 grade levels, aligned with curriculum standards and common item types. MV-MATH [227] 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 [165] (physics subset), TQA [239], and AI2D [228], 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 [229] targets high school physics problems with multi-image reasoning and chain-of-thought prompting, while PhysUniBench [230] advances to undergraduate-level problems across eight sub-disciplines, providing both multiple-choice and open-ended formats. PhysicsArena [231] 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 [232] 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 [233], OlympiadBench [221], and SceMQA [234] 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 [235] evaluates physical commonsense reasoning through audiovisual videos, while GRASP [236] uses simulation-based videos to test intuitive physics reasoning about object permanence and dynamics. CausalVQA [237] focuses on video-based multiple-choice questions requiring causal and physical reasoning. LiveXiv [238] 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 targets 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) [240], which shares a sequential format with natural language but follows a unique chemical grammar. Benchmarks such as ChEBI-20 [241] and ChemBench [242] test sequence-sequence translation (e.g., SMILES $\leftrightarrow$ IUPAC names) and extend to reaction-centric prediction and cross-modal retrieval. Variants like SELFIES [243] and InChI [244] offer alternative encodings, while models such as MolX [254] employ specialized encoders to capture structural patterns.

2D representations include molecular graphs and rendered images, typically generated with RDKit[2)]. Graph-based inputs are aligned with text via projectors, as in GiT-Mol [245] and Instruct-Mol [246], improving property prediction and molecular captioning through explicit spatial-topological encoding. Image-based molecule recognition is evaluated by ChemOCR [247], which spans styles from hand-drawn to scanned and photographed depictions. Multimodal benchmarks such as MMChemBench [247] and ChEBI-20-MM [248] extend earlier datasets with captioning and property prediction tasks. Beyond molecules, chemical vision benchmarks like MMCR-Bench [247] and MACBench [249] target diagram reasoning, scientific figure interpretation, and practical laboratory knowledge.

3D representations capture stereochemistry and spatial configurations critical for chemical function. 3D-MoLM [250] integrates a 3D molecular encoder into an LLM for structure-grounded QA and retrieval, while M3-20M [255] 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 recognition and cross-modal reasoning for structure elucidation. MassSpecGym [251] standardizes MS/MS-based evaluation for de novo generation, retrieval, and simulation. Alberts et al. [252] extend to multi-spectral datasets (IR, MS, NMR) for 790k molecules, enabling cross-spectra modeling. MolPuzzle [253] 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 [256] provides a large-scale, high-quality dataset for chart- and table-based reasoning from financial reports, while FAMMA [257] introduces multilingual QA with textbook-derived and expert-authored questions in both basic and live professional settings. MME-Finance [258] 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 the scope and complexity. MultiFinBen [259] spans multilingual, multimodal (text, vision, audio), and difficulty-aware evaluation, covering tasks from entry-level QA to expert financial reasoning. CFBenchmark-MM [261] focuses on Chinese multimodal QA over diverse visual formats, and FinMMR [262] emphasizes bilingual numerical reasoning across 14 financial subdomains. Fin-Fact [263] pivots to multimodal fact-checking, pairing claims with textual and visual evidence, while FCMR [264] advances to cross-modal multi-hop reasoning that requires integrating textual reports, tables, and charts for multi-step inference.

2) RDKit: Open-source cheminformatics, http://www.rdkit.org.

In parallel, several studies introduce domain-specific models and agent-level evaluations. FinTral [265] presents a family of Mistral-based LMMs trained with financial data, outperforming GPT-4 in multiple benchmarks. Open-FinLLMs [266] offers open-source financial LMMs evaluated across 14 tasks, 30 datasets, and 4 multimodal settings, highlighting gains from targeted pre-training. FinGAIA [260] 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 [267–272] and visual question answering [274–279] tasks. Early VQA benchmarks, such as VQA-RAD [277], PathVQA [278], and SLAKE [279], primarily focus on interpreting individual radiological or pathological images. While these features carefully design question-answer pairs, they lack a complete clinical context and have limited task complexity. Benchmarks like AMOS-MM [280], RP3D-DiagDS [281], and PubMedQA [272], 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 [283] 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 [275] and OpenMM-Medical [284] 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 [285, 286]. Regulatory sequence benchmarks such as Genomics-Long-Range [273] 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 [282] 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 [287] 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 [283]. Although open-ended questions provide greater flexibility, they lack robust and reliable evaluation metrics to assess factual correctness and clinical appropriateness [288]. 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 knowl-

edge coherence [289]. 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 [290], which pairs 484 real webpage screenshots with HTML/CSS and evaluates output fidelity by rendering-based visual similarity and human judgment. Web2Code [291] 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 [292] 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 [293] 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 [298] 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 [294] 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 [295] extends repository-level issue solving to 619 multimodal tasks across 17 JavaScript repositories, where problem descriptions include screenshots or diagrams. Visual SWE-bench [296] similarly evaluates bug fixing with visual context, while MMCode [297] 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 [300] 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 [299], 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 [301] introduces an ego-centric dataset for importance ranking of traffic objects with natural language justifications, while DRAMA [302] collects over 17k interactive driving scenarios for joint risk localization and captioning. NuScenes-QA [303] 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 [304] 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 [305] builds a vehicle-to-vehicle QA benchmark for fusing distributed perception, while MAPLM-QA [316] focuses on traffic and HD map comprehension for domain-specific model fine-tuning. More specialized

benchmarks such as SURDS [306] target spatial reasoning categories like orientation and depth, demonstrating gains from reinforcement learning-based alignment, and AD^2-Bench [307] 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 [308] is the first action-driven benchmark for Vision-Language-Action models, linking perception to high-level driving actions via an action-rooted evaluation tree. DriveLMM-o1 [309] introduces step-by-step reasoning annotations for perception-prediction-planning QA, while DriveVLM [310] integrates VLMs into real-time driving pipelines. RoboTron-Sim [311] leverages simulated hard cases to boost real-world performance, and IDKB [312] evaluates models on over one million handbook, theory, and simulated test items to probe licensing-level knowledge. Complementary resources such as VLADBench [313] enable fine-grained capability breakdowns, and DriVQA [314] 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. [315] evaluated 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 [317] 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 [318], ClimateBERT [319], and WeatherQA [320] integrate textbook, scientific, and RS data to evaluate both comprehension and visual reasoning. For the hydrosphere, OceanBench [332] compiles 13k+ open-ended QA pairs from marine science literature, advancing free-text scientific reasoning.

Multi-sphere benchmarks further extend this coverage. OmniEarth-Bench [337] integrates heterogeneous databases and expert review to produce nearly 30k multiple-choice questions with rich metric annotations. MSEarth [322] emphasizes image-based VQA with chain-of-thought reasoning, featuring 11k+ expert-verified items across all spheres. EarthSE [328] 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 [323] provide hundreds to 10k labeled images, later expanded by NWPU-Captions [338] to 31k images for broader scene coverage. VQA-centric datasets include RSVQA-HRBEN/LRBEN [339] scales to one million automatically generated QA pairs, and DIOR-RSVG [324] enables object grounding evaluation. More recent datasets diversify task coverage, with VRSBench [325] supporting captioning, VQA, and localization, LRS-VQA [326] targeting large-scale RS VQA, and GeoChat-Bench [327] extending to multimodal instruction following.

Recent benchmarks further raise standards of scale, annotation fidelity, and resolution. XLRS-Bench [93] offers ultra-high-resolution imagery (average 8500×8500 pixels) across 16 perception and reasoning tasks. RSIEval [321] provides 2.5k manually annotated descriptions for a detail-oriented assessment. UrBench [116] integrates street-view and satellite imagery for urban perception tasks, while CHOICE [329] mitigates data leakage through independently collected imagery across 23 fine-grained tasks. SARChat-Bench-2M [330] standardizes Synthetic Aperture Radar evaluation with 2M samples covering six target categories.

Fine-grained RS understanding has also become a focus. LHRS-Bench [331] combines label filtering, balanced sampling, and GPT-4-generated instructions to evaluate recognition, spatial perception, and reasoning. FIT-RSFG [333] leverages high-resolution global RS imagery with scene graphs for relational

understanding. VLEO-Bench [334] supports scene understanding, localization, and change detection, while GEOBench-VLM [335] spans disaster monitoring, crop classification, and marine debris detection. NAIP-OSM [336] 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) [359] 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) [340] expands the task to natural language navigation within real building environments, enhancing the realism and complexity of navigation scenarios, while Reverie [341] contributes further challenging navigation tasks in this domain. To address more complex environmental interactions, Alfred [358] designs multi-step, compositional household tasks, and Calvin [356] 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 [355] and Ego4D [342] 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 [343] and SQA3D [344] introduce multi-hop reasoning tasks across space and time, strengthening the understanding of 3D spatial relationships. Open-EQA [345] 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 [346] 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 [357] and EgoTaskQA [348] 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 [349] and RH20T-P [347] 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 [351] 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 [350] aggregates 328 tasks and 125 3D scenes, encompassing navigation, interaction, social engagement, and multidimensional question answering. EmbodiedBench [352] 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 [353] 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 embodied intelligence. Additionally, EWMBench [354] assesses generative planning abilities of embodied world models through scene consistency, action correctness, and semantic alignment. Meanwhile, the NeurIPS 2025 Embodied Agent Interface Challenge [360] 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.

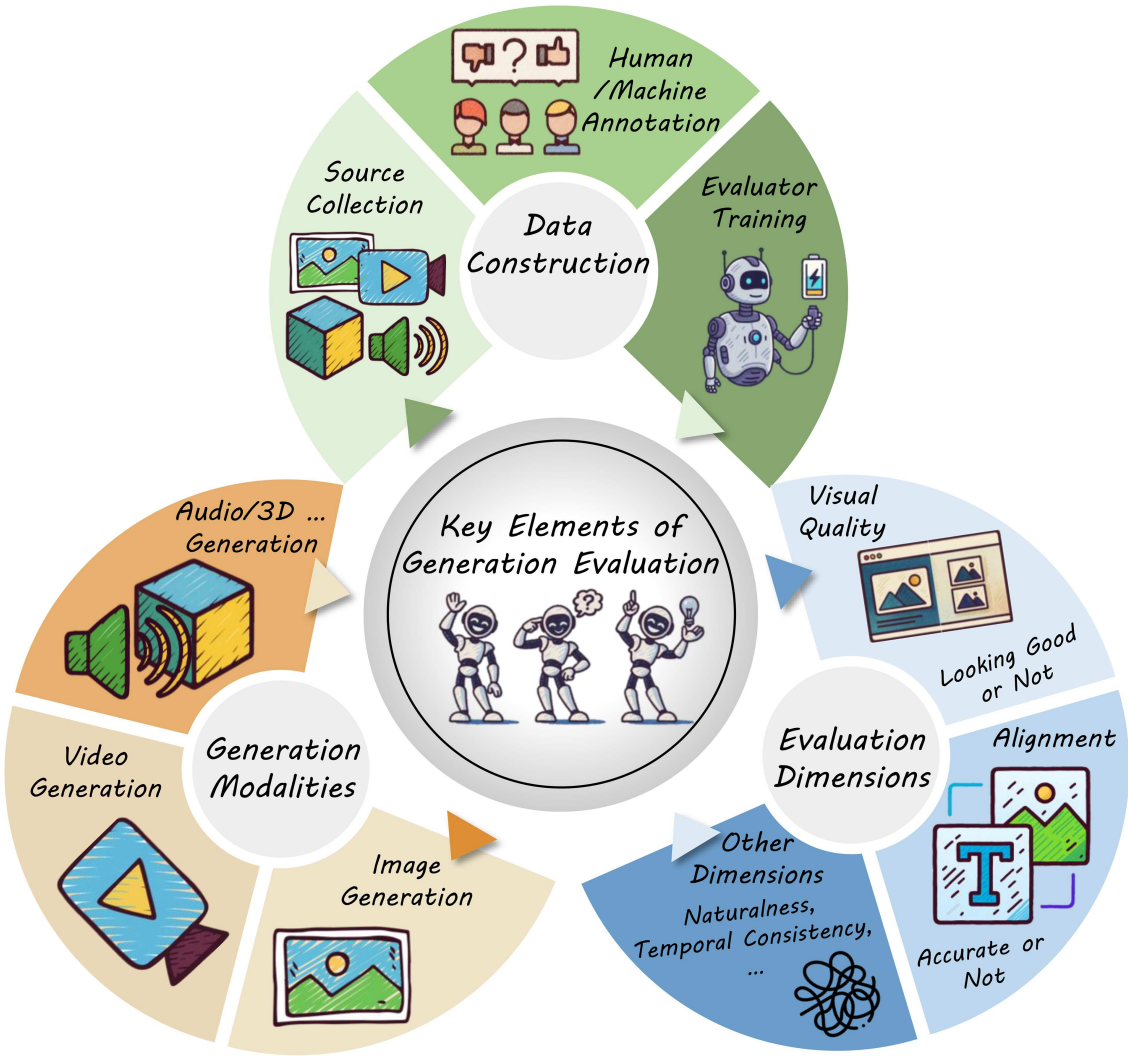

Figure 5 (Color online) Key elements of generation evaluation for large multimodal models. The process involves three complementary aspects: (1) data construction, which covers source collection, human/machine annotation, and evaluator training; (2) generation modalities, spanning image, video, audio, 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.

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.

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.

Table 2 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 [361]	2022	14.0M	–	1	None
HPD [362]	2023	98.8k	98.8k	1	Human preference
ImageReward [363]	2023	136.9k	136.9k	3	Human preference
Pick-A-Pic [364]	2023	500.0k	500.0k	6	Human preference
AGIQA-1K [365]	2023	1.1k	23.7k	2	Overall perceptual quality
AGIQA-3K [366]	2023	3.0k	125.0k	6	Perceptual quality, text alignment
AIGCIQA2023 [367]	2023	2.4k	48.0k	6	Perceptual quality, authenticity, correspondence
AGIN [368]	2023	6.1k	181.0k	18	Overall naturalness
AIGIQA-20K [369]	2024	20.0k	420.0k	15	Perceptual quality, text alignment
AIGCOIQA2024 [370]	2024	0.3k	6.0k	5	Perceptual quality, comfortability, text alignment
CMC-Bench [371]	2024	58.0k	160.0k	6	Ultra-low bitrate compression quality
PKU-I2IQA [372]	2023	3.2k	96.0k	6	Perceptual quality in NR and FR settings
SeeTRUE [373]	2023	31.9k	31.9k	–	Text-image semantic alignment verification
AIGCIQA2023+ [374]	2025	2.4k	48.0k	6	Perceptual quality, authenticity, correspondence
Q-Eval-100K [375]	2025	100.0k	960.0k	23	Visual quality, long-text alignment (images/videos)
HPD v2 [376]	2023	433.8k	798.1k	–	Human preference on diverse AIGIs
Quality assessment for AIGVs					
Chivileva et al. [377]	2023	1.0k	48.2k	5	Perceptual quality, text alignment
EvalCrafter [378]	2023	3.5k	8.6k	7	Perceptual quality, text alignment, temporal quality
FETV [379]	2023	2.5k	29.7k	3	Perceptual quality, text alignment, temporal quality
VBench [380]	2023	7.0k	–	4	Video quality, consistency
T2VQA-DB [381]	2024	10.0k	540.0k	9	Perceptual quality, text alignment
GAIA [382]	2024	9.2k	971.0k	18	Video action quality
AIGVQA-DB [383]	2025	36.6k	122.0k	15	Perceptual quality, temporal smoothness, dynamic degree, alignment
AIGVE-60K [384]	2025	58.5k	180.0k	30	Perceptual preference, text-video correspondence, task-specific accuracy
Human-AGVQA-DB [385]	2025	6.0k	630.0k	22	Human activity quality
TDVE-DB [386]	2025	3.9k	173.6k	12	Edited quality, editing alignment, structural consistency
AGAVQA-3K [387]	2025	3.1k	9.3k	8	Audio-visual quality, content consistency, overall quality
Quality assessment for AIGAs					
Qwen-ALLD [388]	2025	20.0k	60.0k	–	MOS, SIM, A/B preference (speech quality)
BASE-TTS [389]	2024	–	–	–	Fine-grained semantic capture, emotion, prosody
ATT [390]	2025	–	–	–	Human-likeness, multi-dimensional TTS quality
TTSDS2 [391]	2025	0.3k	–	20	Prosody, intelligibility, multi-lingual TTS quality
Quality assessment for AIG3Ds					
MATE-3D [392]	2024	1.3k	107.5k	8	Alignment, geometry, texture, overall quality
3DGCQA [393]	2025	0.3k	12.5k	7	Alignment, overall quality
AIGC-T23DAQA [394]	2025	1.0k	1.0k	6	Quality, authenticity, text-content correspondence
SI23DCQA [395]	2025	1.5k	–	5	Overall, color, shape quality
3DGS-IEval-15K [396]	2025	15.2k	228.0k	6	Image quality for compressed 3D Gaussian splatting

3.1.1 *Datasets for AI-generated image quality assessment*

This section surveys datasets for generation-oriented evaluation of AI-generated images (AIGIs) from T2I and related pipelines, emphasizing perceptual quality, semantic alignment, authenticity, and aesthetics. A brief comparison of quality assessment datasets for visual AIGC is shown in Table 2. 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 [361] 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) [362] and HPD v2 [376] collect 98.8k and 798k human choices, respectively, serving as foundational resources for training HPS-style reward models. Pick-a-Pic [364] 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 [363] 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 [365] contains 1k images annotated for technical quality, aesthetics, and text alignment. AGIQA-3K [366] expands this to 2.9k images, while AIGIQA-20K [369] scales up to 20k images from 15 text-to-image models with 420k human ratings. PKU-I2IQA [372] focuses on image-to-image generation under both no-reference and full-reference settings. For assessing naturalness, AGIN [368] 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 [373] 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 [397] 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 [367] 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+ [374] enriches these assessments with human preference scores and explanatory annotations, adding interpretability. Building on this foundation, Q-Eval-100K [375] 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 [371] 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) [398] 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 [399] 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 [362] and PickScore [364] learn to mimic human choices on AIGIs. To improve semantic robustness, BLIP-based ImageReward/ReFL [363] 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 [373]. A complementary formulation, VQAScore [397], 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 [400]) provide open testbeds and leaderboards that calibrate these alignment/quality assessors and track progress across methods.

(3) LMM-driven evaluation with instruction-tuned assessors. LMMs can directly emit quantitative quality judgments as well. Q-Bench [90] 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 [111, 401–403] introduce instruction/training procedures that align LMMs to IQA objectives and improve rating consistency. DepictQA [404] elicits fine-grained, language-based rationales for human-like judgments, while M3-AGIQA [405] conducts multi-round, multi-aspect prompting for holistic, human-aligned assessment across visual and textual facets. A unified evaluator, Q-Eval-Score [375], scores both perceptual quality and long-text alignment, improving robustness on complex prompts. Crucially, evaluation signals can close the loop: Q-Refine [406] 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 [407] incorporates semantic guidance by injecting prompt cues and combining them through a mixture-of-experts framework, while SF-IQA [408] focuses on fusing quality and similarity features using a multi-layer extractor built on a fine-tuned vision-language backbone. SC-AGIQA [409] explicitly enforces text-visual semantic constraints to jointly evaluate alignment and perceptual distortion, and MINT-IQA [374] extends this direction by learning multi-perspective human preferences and leveraging instruction tuning to enhance explainability. TSP-MGS [410] further separates perceptual quality and alignment using task-specific prompts and combines information across multiple granularities, whereas MoE-AGIQA [411] 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 [412] fuses global and local features with alignment information, PSCR [413] uses patch-sampling contrastive regression to improve robustness and reduce geometric bias, and NR/FR-AIGCIQA [372] compares no-reference (NR) regression with full-reference (FR) feature fusion. Other approaches combine multimodal information more explicitly: TIER [414] regresses quality jointly from text and image encoders, JOINT [368] models technical and rationality cues to capture naturalness, and IPCE [415] transforms calibrated classification probabilities into continuous regression targets to improve CLIP-based quality assessment.

(5) Reasoning-driven generation and editing evaluation. A complementary sub-direction 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 [416] 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 [417] 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 [418] explores complex 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 [419] probes world-knowledge-informed semantic correctness for T2I, complementing editing benchmarks along the knowledge axis, while KRIS-Bench [420] organizes editing tasks by factual, conceptual, and procedural knowledge with multi-dimensional reasoning diagnostics. Methodologically, CoT-editing [421] demonstrates that injecting chain-of-thought into the editing loop (plan-act-verify) yields measurable gains on multi-constraint, multi-step edits. Together, these studies 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) [422], particularly targeting user-generated content (UGC) [423, 424]. LMM-VQA [425] 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 [426] introduces fine-grained MOS annotations across six dimensions and applies efficient LoRA-based fine-tuning for multi-dimensional UGC video quality prediction. VQA^2 [427] 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 [428] 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 [429] draws from the Compare2Score paradigm [430], 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 studies such as Chivileva et al. [377] and EvalCrafter [378] focus on perceptual and alignment quality with 1k–2.5k videos, while FETV [379] and VBench [380] introduce fine-grained attribute annotations and preference-based evaluation at moderate scales. Recent general-purpose benchmarks have scaled substantially, including AIGVQA-DB [383] (36.6k videos, 122k MOS/pairwise annotations), AIGVE-60K [384] (60k videos with $\sim$12k MOS and 60k instruction pairs), and T2VQA-DB [381] (10k videos with MOS from 27 subjects). Meanwhile, specialized datasets address more focused aspects of video and audio-visual evaluation. GAIA [382] emphasizes motion realism, providing 9k video-action pairs with 971k human ratings. Human-AGVQA-DB [385] targets the quality of human activity in videos, while TDVE-DB [386] concentrates on assessing text-driven video editing. Extending beyond vision, AGAVQA-3K [387] incorporates audio-visual content 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 [383] 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 [392] leverages foreground-background prompting to probe spatial-temporal disentanglement. GHVQ [385] addresses human activity quality assessment using a dual-branch architecture (spatial quality analyzer and action quality analyzer), integrated with CLIP-based regression. LOVE [384] systematically evaluates both AIGV quality and LMM understanding across three dimensions and twenty subtasks. TDVE-Assessor [386] benchmarks TDVE-specific tasks, focusing on editing-relevant quality dimensions. VQ-Insight [431] 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 [387] 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.

3.2.4 *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" [432] and "THQA" [433] 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" [434] addresses animated human imitation, combining subjective studies and objective metrics to assess realism and fidelity. Methodologically, "MI3S" [435] 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" [436] explores practical pipelines for digital human generation via multimodal fusion, bridging algorithmic evaluation with engineering deployment. Together, these studies 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 [437] pioneered deep learning-based MOS prediction for converted speech. MOSA-Net+ [438] extended this approach by leveraging acoustic features extracted from Whisper. MOSLight [439] pursued a lightweight design using 1D convolutions for faster inference. MBNet [440], DeePMOS [441], and LDNet [442] incorporated listener-specific perceived quality scores in addition to the mean opinion scores. ADTMOS [443] further enhanced robustness with a frame-wise MOS generator and an audio distortion token extractor. UAMOS [444] proposed an uncertainty-aware MOS framework to improve reliability in open-world applications. Audiobox Aesthetics [445] decomposed human listening perspectives into four perceptual axes for aesthetic evaluation. While most methods resample audio to a fixed rate, HighRateMOS [446] 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. [447] 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 [448], Qwen-Audio [449], Qwen2-Audio [450]) for multiple assessment tasks, including MOS and speaker similarity (SIM) prediction, A/B preference testing, and natural language quality descriptions [451]. Qwen-ALLD [388] introduced the first natural language-based speech quality dataset and employed LLM distillation to align ALMs for MOS, SIM, and A/B testing. QualiSpeech [452] further advanced natural language-based evaluation by integrating reasoning and contextual cues to improve accuracy and interpretability.

(3) Benchmark-driven evaluation. To systematically evaluate (text-to-speech) TTS and speech synthesis systems, several multi-dimensional benchmark frameworks have been proposed. BASE-TTS [389] presented an English emergent abilities suite covering seven categories of text (e.g., emotions, paralinguistics, syntactic complexities) to test fine-grained semantic capture. DiscreteEval [453] assessed five dimensions: speaking style, intelligibility, speaker consistency, prosodic variation, and spontaneity. EmergentTTS-Eval [454] addressed six challenging scenarios (e.g., emotions, foreign words, complex pronunciation) with ALM-generated test cases evaluated by an ALM. ATT [390] 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 [391] benchmarked 20 open-source TTS systems across 14 languages along four axes, including prosody and intelligibility. InstructTTSEval [455] evaluated instruction-driven TTS models on acoustic-parameter control, descriptive-style directives, and role-play scenarios. Mos-Bench [456] 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 [457,458] are developed based on limited-scale datasets. Later, several benchmarks target the perceptual evaluation of 3D assets generated from textual descriptions. MATE-3D [392] contains 1280 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 [393] 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 [394] 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 [459] provides ∼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 [395] 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 studies include NeRF-NQA [460], 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 [461] focuses on compression, providing subjective scores for multiple parameter levels of 22 source objects. Martin et al. [462,463] conducted extensive subjective studies across scene types and benchmark FR/NR metrics against human scores. For Gaussian splatting, GS-QA [464] evaluates static GS methods under 360° and forward-facing trajectories with subjective ratings and 18 objective metrics. 3DGS-IEval-15K [396] 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 studies have developed automatic evaluation pipelines aligned with human judgments. GPT-4V Evaluator [465] 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 [466] integrates pretrained models and LMMs to score Geometric Consistency, Semantic Consistency, Structural Consistency, Text-3D Alignment, and Aesthetics. 3DGen-Bench [467] combines CLIP-based 3DGen-Score and MLLM-based 3DGen-Eval to better correlate with human preferences through multimodal reasoning. LMM-PCQA [468] 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 [469] 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 [469] 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.

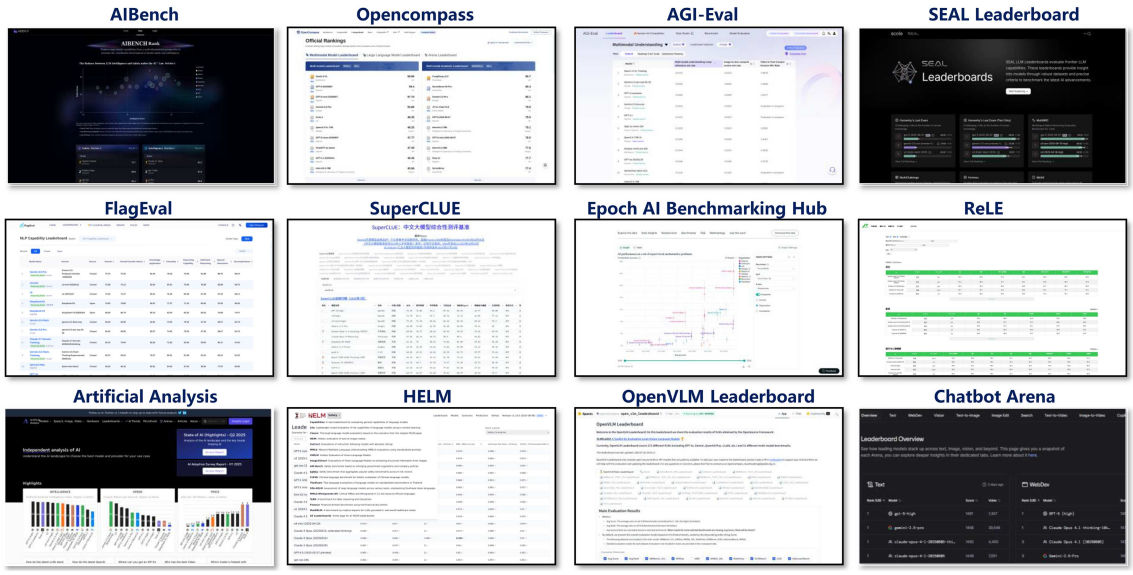

Figure 6 (Color online) Screenshots of the representative evaluation platforms.

Beyond toolkit-style designs, LMMS-Eval [13] 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 [470] 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 Figure 6). OpenCompass [12] exemplifies this trend by consolidating evaluations across reasoning, LLMs, vision-language, and spatial domains, providing multi-capability insights under standardized workflows. Similarly, HELM [471] 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 [472] 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 [473] visualizes historical benchmark performance trends, and Artificial Analysis[3)] provides comparative intelligence across providers with insights into latency, price, and safety. Scale's SEAL Leaderboards[4)] adopt expert-curated, high-complexity evaluations, ensuring robustness by testing models only on unseen prompts. FlagEval[5)] introduces customizable tasks and debate-style comparisons, while AGI-Eval [474] 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 [475] provides a live-updating leaderboard covering hundreds of fine-grained dimensions in education, finance, healthcare, and law. SuperCLUE [476] 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 [10] 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.

3) Artificial-Analysis, https://artificialanalysis.ai/.
4) Scale, SEAL Leaderboards, https://scale.com/leaderboard/.
5) FlagEval, NLP Capability Leaderboard, https://flageval.baai.ac.cn/#/leaderboard/.

5 Outlook & conclusion

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.

Acknowledgements This work was supported by Shanghai Artificial Intelligence Laboratory and National Natural Science Foundation of China (Grant Nos. 623B2073, 62101326, 62225112, 62301316).

References

1 Hurst A, Lerer A, Goucher A P, et al. GPT-4o system card. 2024. ArXiv:2410.21276
2 OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. 2024. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
3 Comanici G, Bieber E, Schaekermann M, et al. Gemini 2.5: pushing the Frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025. arXiv:2507.06261
4 xAI Team. Grok-3: The Age of Reasoning Agents. Technical Report, 2025. https://x.ai/news/grok-3
5 Bai S, An S. A survey on automatic image caption generation. Neurocomputing, 2018, 311: 291–304
6 Wu Q, Teney D, Wang P, et al. Visual question answering: a survey of methods and datasets. Comput Vision Image Understanding, 2017, 163: 21–40
7 Zhang C, Zhang C, Zhang M, et al. Text-to-image diffusion models in generative AI: a survey. 2023. ArXiv:2303.07909
8 Singh A. A survey of AI text-to-image and AI text-to-video generators. In: Proceedings of the 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC), 2023. 32–36
9 Yang C, Lu C, Wang Y, et al. Towards ai-45° law: a roadmap to trustworthy agi. 2024. ArXiv:2412.14186
10 Zhang Z, Wang J, Guo Y, et al. AIBench: towards trustworthy evaluation under the 45° law. Displays, 2026, 91: 103255
11 Zhang Z, Zhou Y, Li C, et al. Quality assessment in the era of large models: a survey. ACM Trans Multimedia Comput Commun Appl, 2025, 21: 1–31
12 Opencompass. Compassbench large language model leaderboard. 2025. https://rank.opencompass.org.cn/leaderboard-llm/
13 Zhang K, Li B, Zhang P, et al. Lmms-eval: reality check on the evaluation of large multimodal models. 2024. ArXiv:2407.12772
14 Zhong W, Cui R, Guo Y, et al. Agieval: a human-centric benchmark for evaluating foundation models. 2023. ArXiv:2304.06364
15 Fu C, Zhang Y F, Yin S, et al. Mme-survey: a comprehensive survey on evaluation of multimodal llms. 2024. ArXiv:2411.15296
16 Li L, Chen G, Shi H, et al. A survey on multimodal benchmarks: In the era of large AI models. 2024. ArXiv:2409.18142
17 Huang J, Zhang J. A survey on evaluation of multimodal large language models. 2024. ArXiv:2408.15769
18 Min X K, Duan H Y, Sun W, et al. Perceptual video quality assessment: a survey. Sci China Inf Sci, 2024, 67: 211301
19 Zhai G T, Min X K. Perceptual image quality assessment: a survey. Sci China Inf Sci, 2020, 63: 211301
20 Chen Z, Wang W Y, Tian H, et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. Sci China Inf Sci, 2024, 67: 220101
21 Zhang Y, Ji Z, Pang Y W, et al. Modality-experts coordinated adaptation for large multimodal models. Sci China Inf Sci, 2024, 67: 220107
22 Liu Y Z, Cao Y, Gao Z W, et al. MMInstruct: a high-quality multi-modal instruction tuning dataset with extensive diversity. Sci China Inf Sci, 2024, 67: 220103
23 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 36: 34892–34916

24 Qian Y, Ye H, Fauconnier J P, et al. Mia-bench: towards better instruction following evaluation of multimodal LLMs. 2024. ArXiv:2407.01509
25 Ding S, Wu S, Zhao X, et al. Mm-ifengine: towards multimodal instruction following. 2025. ArXiv:2504.07957
26 Bitton Y, Bansal H, Hessel J, et al. Visit-bench: a benchmark for vision-language instruction following inspired by real-world use. 2023. ArXiv:2308.06595
27 Liu Z, Chu T, Zang Y, et al. Mmdu: a multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMS. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 8698–8733
28 Liu S, Ying K, Zhang H, et al. Convbench: a multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. 2024. ArXiv:2403.20194
29 Kottur S, Moon S, Geramifard A, et al. Simmc 2.0: a task-oriented dialog dataset for immersive multimodal conversations. 2021. ArXiv:2104.08667
30 Wang X, Zhou Y, Liu X, et al. Mementos: a comprehensive benchmark for multimodal large language model reasoning over image sequences. 2024. ArXiv:2401.10529
31 Wang F, Fu X, Huang J Y, et al. Muirbench: a comprehensive benchmark for robust multi-image understanding. 2024. ArXiv:2406.09411
32 Meng F, Wang J, Li C, et al. Mmiu: multimodal multi-image understanding for evaluating large vision-language models. 2024. ArXiv:2408.02718
33 Zhao B, Zong Y, Zhang L, et al. Benchmarking multi-image understanding in vision and language models: perception, knowledge, reasoning, and multi-hop reasoning. 2024. ArXiv:2406.12742
34 Liu H, Zhang X, Xu H, et al. Mibench: evaluating multimodal large language models over multiple images. 2024. ArXiv:2407.15272
35 Liu Z, Fang F, Feng X, et al. Ii-bench: an image implication understanding benchmark for multimodal large language models. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 46378–46480
36 Jiang D, He X, Zeng H, et al. Mantis: interleaved multi-image instruction tuning. 2024. ArXiv:2405.01483
37 Song D, Chen S, Chen G H, et al. Milebench: benchmarking MLLMs in long context. 2024. ArXiv:2404.18532
38 Kazemi M, Dikkala N, Anand A, et al. Remi: a dataset for reasoning with multiple images. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 60088–60109
39 Luo F, Chen C, Wan Z, et al. Codis: benchmarking context-dependent visual comprehension for multimodal large language models. 2024. ArXiv:2402.13607
40 Huang Y, Meng Z, Liu F, et al. Sparkles: unlocking chats across multiple images for multimodal instruction-following models. 2023. ArXiv:2308.16463
41 Xia P, Han S, Qiu S, et al. Mmie: massive multimodal interleaved comprehension benchmark for large vision-language models. 2024. ArXiv:2410.10139
42 Liu M, Xu Z, Lin Z, et al. Holistic evaluation for interleaved text-and-image generation. 2024. ArXiv:2406.14643
43 Zhou P, Peng X, Song J, et al. Opening: a comprehensive benchmark for judging open-ended interleaved image-text generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 56–66
44 Raza S, Narayanan A, Khazaie V R, et al. Humanibench: a human-centric framework for large multimodal models evaluation. 2025. ArXiv:2505.11454
45 Li K, Yang Z, Zhao J, et al. Herm: benchmarking and enhancing multimodal LLMs for human-centric understanding. 2024. ArXiv:2410.06777
46 Zhou Z, Wang Q, Lin B, et al. Uniaa: a unified multi-modal image aesthetic assessment baseline and benchmark. 2024. ArXiv:2404.09619
47 Liao Z, Liu X, Qin W, et al. Humanaesexpert: advancing a multi-modality foundation model for human image aesthetic assessment. 2025. ArXiv:2503.23907
48 Sap M, Rashkin H, Chen D, et al. Socialiqa: commonsense reasoning about social interactions. 2019. ArXiv:1904.09728
49 Shen J, Kim Y, Hulse M, et al. Empathicstories++: a multimodal dataset for empathy towards personal experiences. 2024. ArXiv:2405.15708
50 Chiang W L, Zheng L, Sheng Y, et al. Chatbot Arena: an open platform for evaluating LLMs by human preference. In: Proceedings of the 41st International Conference on Machine Learning, 2024
51 Köpf A, Kilcher Y, Von Rütte D, et al. OpenAssistant conversations-democratizing large language model alignment. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 36: 47669–47681
52 Guo Y, Ji K, Zhu X, et al. Human-centric evaluation for foundation models. 2025. ArXiv:2506.01793
53 Gao J, Zhao L, Li X. Nwpu-moc: a benchmark for fine-grained multicategory object counting in aerial images. IEEE Trans Geosci Remote Sensing, 2024, 62: 1–14
54 Sun K, Huang K, Liu X, et al. T2v-compbench: a comprehensive benchmark for compositional text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 8406–8416
55 Wu X, Yu D, Huang Y, et al. Conceptmix: a compositional image generation benchmark with controllable difficulty. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 86004–86047
56 Zhao Z, Lu P, Zhang A, et al. Can machines understand composition? Dataset and benchmark for photographic image composition embedding and understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 14411–14421
57 Singh A, Natarajan V, Shah M, et al. Towards VQA models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 8317–8326
58 Mishra A, Shekhar S, Singh A K, et al. Ocr-vqa: visual question answering by reading text in images. In: Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019. 947–952
59 Liu Y L, Li Z, Huang M X, et al. OCRBench: on the hidden mystery of OCR in large multimodal models. Sci China Inf Sci, 2024, 67: 220102
60 Fu L, Kuang Z, Song J, et al. Ocrbench v2: an improved benchmark for evaluating large multimodal models on visual text localization and reasoning. 2024. ArXiv:2501.00321
61 Jia Q, Yue X, Huang S, et al. Visual perception in text strings. 2024. ArXiv:2410.01733
62 Huang M, Shi Y, Peng D, et al. OCR-reasoning benchmark: unveiling the true capabilities of MLLMs in complex text-rich image reasoning. 2025. ArXiv:2505.17163
63 Zhao M, Li B, Wang J, et al. Towards video text visual question answering: benchmark and baseline. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 35549–35562

64 Li B, Ge Y, Chen Y, et al. Seed-bench-2-plus: benchmarking multimodal large language models with text-rich visual comprehension. 2024. ArXiv:2404.16790
65 Feng H, Liu Q, Liu H, et al. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. Sci China Inf Sci, 2024, 67: 220106
66 Zhu F, Liu Z, Ng X Y, et al. Mmdocbench: benchmarking large vision-language models for fine-grained visual document understanding. 2024. ArXiv:2410.21311
67 Ma Y, Zang Y, Chen L, et al. Mmlongbench-doc: benchmarking long-context document understanding with visualizations. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 95963–96010
68 Hui Y, Lu Y, Zhang H. Uda: a benchmark suite for retrieval augmented generation in real-world document analysis. In: Proceedings of Advances in Neural Information Processing Systems, 2024, 37: 67200–67217
69 Tanaka R, Nishida K, Yoshida S. Visualmrc: machine reading comprehension on document images. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 13878–13888
70 Mathew M, Karatzas D, Jawahar C. Docvqa: a dataset for vqa on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021. 2200–2209
71 Xia R, Mao S, Yan X, et al. Docgenome: an open large-scale scientific document benchmark for training and testing multi-modal large language models. 2024. ArXiv:2406.11633
72 Li S, Shen Y, Chen X, et al. Gdi-bench: a benchmark for general document intelligence with vision and reasoning decoupling. 2025. ArXiv:2505.00063
73 Rawles C, Li A, Rodriguez D, et al. Androidinthewild: a large-scale dataset for Android device control. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 36: 59708–59728
74 Cheng K, Sun Q, Chu Y, et al. Seeclick: harnessing GUI grounding for advanced visual GUI agents. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 9313–9332
75 Liu J, Song Y, Lin B Y, et al. Visualwebbench: how far have multimodal LLMs evolved in web page understanding and grounding? 2024. ArXiv:2404.05955
76 Chen D, Huang Y, Wu S, et al. Gui-world: a video benchmark and dataset for multimodal GUI-oriented understanding. 2024. ArXiv:2406.10819
77 Lin Z, Zhou Z, Zhao Z, et al. Webuibench: a comprehensive benchmark for evaluating multimodal large language models in webui-to-code. 2025. ArXiv:2506.07818
78 Hsiao Y C, Zubach F, Baechler G, et al. Screenqa: large-scale question-answer pairs over mobile app screenshots. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025
79 Masry A, Do X L, Tan J Q, et al. ChartQA: a benchmark for question answering about charts with visual and logical reasoning. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 2263–2279
80 Masry A, Islam M S, Ahmed M, et al. Chartqapro: a more diverse and challenging benchmark for chart question answering. 2025. ArXiv:2504.05506
81 Zhao W, Feng H, Liu Q, et al. Tabpedia: towards comprehensive visual table understanding with concept synergy. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 7185–7212
82 Kim Y, Yim M, Song K Y. Tablevqa-bench: a visual question answering benchmark on multiple table domains. 2024. ArXiv:2404.19205
83 Wang Z, Xia M, He L, et al. Charting gaps in realistic chart understanding in multimodal LLMs. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 113569–113697
84 Roberts J, Han K, Houlsby N, et al. Scifibench: benchmarking large multimodal models for scientific figure interpretation. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 18695–18728
85 Hiippala T, Alikhani M, Haverinen J, et al. AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. Lang Resour Evaluation, 2021, 55: 661–688
86 Lin M, Xie T, Liu M, et al. Infochartqa: a benchmark for multimodal question answering on infographic charts. 2025. ArXiv:2505.19028
87 Huang M, Lai H, Zhang X, et al. Evochart: a benchmark and a self-training approach towards real-world chart understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 3680–3688
88 Foroutan N, Romanou A, Ansaripour M, et al. Wikimixqa: a multimodal benchmark for question answering over tables and charts. 2025. ArXiv:2506.15594
89 Xia R, Zhang B, Ye H, et al. Chartx & chartvlm: a versatile benchmark and foundation model for complicated chart reasoning. 2024. ArXiv:2402.12185
90 Wu H, Zhang Z, Zhang E, et al. Q-bench: a benchmark for general-purpose foundation models on low-level vision. 2023. ArXiv:2309.14181
91 Zhang Z, Wu H, Li C, et al. A-bench: are lmms masters at evaluating AI-generated images? 2024. ArXiv:2406.03070
92 Li G, Xie Y, Kan M Y. MVP-bench: can large vision–language models conduct multi-level visual perception like humans? 2024. ArXiv:2410.04345
93 Wang F, Wang H, Guo Z, 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, 2025. 14325–14336
94 Wang W, Ding L, Zeng M, et al. 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, 2025. 7907–7915
95 Zhang Y F, Zhang H, Tian H, et al. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? 2024. ArXiv:2408.13257
96 Wu P, Xie S. V*: guided visual search as a core mechanism in multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 13084–13094
97 Wang X, Ma X, Hou X, et al. 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, 2025. 9154–9164
98 Liu Z, Qian L, Xie Q, et al. Mmaffben: a multilingual and multimodal affective analysis benchmark for evaluating LLMs and VLMS. 2025. ArXiv:2505.24423
99 Li Y, Dao A, Bao W, et al. Facial affective behavior analysis with instruction tuning. In: Proceedings of European Conference on Computer Vision, 2024. 165–186
100 Zhou Y, Zhang Z, Cao J, et al. Memo-bench: a multiple benchmark for text-to-image and multimodal large language models on human emotion analysis. 2024. ArXiv:2411.11235

101 Sabour S, Liu S, Zhang Z, et al. Emobench: evaluating the emotional intelligence of large language models. 2024. ArXiv:2402.12071
102 Gao L, Jia Z, Zeng Y, et al. Eemo-bench: a benchmark for multi-modal large language models on image evoked emotion assessment. 2025. ArXiv:2504.16405
103 Huang Y, Yuan Q, Sheng X, et al. Aesbench: an expert benchmark for multimodal large language models on image aesthetics perception. 2024. ArXiv:2401.08276
104 Zou H P, Samuel V, Zhou Y, et al. Implicitave: an open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction. 2024. ArXiv:2404.15592
105 Song X, Wu M, Zhu K Q, et al. 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, 2025. 6392–6409
106 Wang J, Li W, Wu Y, et al. Affordance benchmark for MLLMs. 2025. ArXiv:2506.00893
107 Wang Y, Liao Y, Liu H, et al. Mm-sap: a comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. 2024. ArXiv:2401.07529
108 Tong P, Brown E, Wu P, et al. Cambrian-1: a fully open, vision-centric exploration of multimodal LLMs. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 87310–87356
109 Li J, Wei Q, Zhang C, et al. Single image unlearning: efficient machine unlearning in multimodal large language models. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 35414–35453
110 Tong S, Liu Z, Zhai Y, et al. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 9568–9578
111 Wu H, Zhu H, Zhang Z, et al. Towards open-ended visual quality comparison. 2024. ArXiv:2402.16641
112 Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions. In: Proceedings of European Conference on Computer Vision, 2016. 69–85
113 Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 11–20
114 Chen J, Wei F, Zhao J, et al. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 513–524
115 Wang W, Yue T, Zhang Y, et al. Unveiling parts beyond objects: towards finer-granularity referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 12998–13008
116 Zhou B, Yang H, Chen D, et al. Urbench: a comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 10707–10715
117 Li J, Zhang X, Zou H, et al. Counts: benchmarking object detectors and multimodal large language models under distribution shifts. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 9186–9198
118 Tang J, Liu Q, Ye Y, et al. Mtvqa: benchmarking multilingual text-centric visual question answering, 2024. ArXiv:2405.11985
119 Xing S, Xiang C, Han Y, et al. Gepbench: evaluating fundamental geometric perception for multimodal large language models. 2024. ArXiv:2412.21036
120 Liu J, Liu Z, Cen Z, et al. Can multimodal large language models understand spatial relations? 2025. ArXiv:2505.19015
121 Cheng A C, Yin H, Fu Y, et al. Spatialrgpt: grounded spatial reasoning in vision-language models. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 135062–135093
122 Zhu Y, Wang Z, Zhang C, et al. Cospace: benchmarking continuous space perception ability for vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 29569–29579
123 Kil J, Mai Z, Lee J, et al. MLLM-Compbench: a comparative reasoning benchmark for multimodal LLMs. 2024. ArXiv:2407.16837
124 Wang A, Wu B, Chen S, et al. 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, 2024. 13384–13394
125 Rajabi N, Kosecka J. Gsr-bench: a benchmark for grounded spatial reasoning evaluation via multimodal LLMs. 2024. ArXiv:2406.13246
126 Kamath A, Hessel J, Chang K W. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. 2023. ArXiv:2310.19785
127 Liao Y H, Mahmood R, Fidler S, et al. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models, 2024. ArXiv:2409.09788
128 Wang W, Ren Y, Luo H, et al. The all-seeing project v2: towards general relation comprehension of the open world, 2024. ArXiv:2402.19474
129 Shao H, Qian S, Xiao H, et al. Visual cot: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 8612–8642
130 Xiao Y, Sun E, Liu T, et al. Logicvista: multimodal LLM logical reasoning benchmark in visual contexts. 2024. ArXiv:2407.04973
131 Xu W, Wang J, Wang W, et al. Visulogic: a benchmark for evaluating visual reasoning in multi-modal large language models. 2025. ArXiv:2504.15279
132 Cheng Z, Chen Q, Zhang J, et al. Comt: a novel benchmark for chain of multi-modal thought on large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 23678–23686
133 Estermann B, Lanzendörfer L, Niedermayr Y, et al. Puzzles: a benchmark for neural algorithmic reasoning. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 127059–127098
134 Zhao H H, Zhou P, Gao D, et al. Lova3: learning to visual question answering, asking and assessment. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 115146–115175
135 Huang H, Zhong H, Yu T, et al. Vlkeb: a large vision-language model knowledge editing benchmark. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 9257–9280
136 Du Y, Jiang K, Gao Z, et al. Mmke-bench: a multimodal editing benchmark for diverse visual knowledge. 2025. ArXiv:2502.19870
137 Zhang J, Zhang H, Yin X, et al. Mc-mke: a fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. 2024. ArXiv:2406.13219
138 Li J, Du M, Zhang C, et al. Mike: a new benchmark for fine-grained multimodal entity knowledge editing. 2024. ArXiv:2402.14835

139 Zhang Y, Su Y, Liu Y, et al. Negvqa: can vision language models understand negation? 2025. ArXiv:2505.22946
140 Xu P, Shao W, Zhang K, et al. Lvlm-ehub: a comprehensive evaluation benchmark for large vision-language models. 2023. ArXiv:2306.09265
141 Shao W, Lei M, Hu Y, et al. Tinylvlm-ehub: towards comprehensive and efficient evaluation for large vision-language models. 2024. ArXiv:2308.03729
142 Yin Z, Wang J, Cao J, et al. Lamm: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. 2024. ArXiv:2306.06687
143 Fu C, Chen P, Shen Y, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. 2024. ArXiv:2306.13394
144 Liu Y, Duan H, Zhang Y, et al. Mmbench: is your multi-modal model an all-around player? 2023. ArXiv:2307.06281
145 Li B, Wang R, Wang G, et al. Seed-bench: benchmarking multimodal LLMs with generative comprehension. 2023. ArXiv:2307.16125
146 Li B, Ge Y, Ge Y, et al. Seed-bench-2: benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 13299–13308
147 Ying K, Meng F, Wang J, et al. Mmt-bench: a comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. 2024. ArXiv:2404.16006
148 Li B, Zhang P, Zhang K, et al. Lmms-eval: accelerating the development of large multimodal models. Version v0.1.0. 2024. https://github.com/EvolvingLMMs-Lab/lmms-eval
149 Chen L, Li J, Dong X, et al. Are we on the right way for evaluating large vision-language models? 2024. ArXiv:2403.20330
150 Li B, Lin Z, Peng W, et al. Naturalbench: evaluating vision-language models on natural adversarial samples, 2025. ArXiv:2410.14669
151 Yu W, Yang Z, Li L, et al. Mm-vet: evaluating large multimodal models for integrated capabilities. 2024. ArXiv:2308.02490
152 Shi Z, Wang Z, Fan H, et al. Chef: a comprehensive evaluation framework for standardized assessment of multimodal large language models. 2023. ArXiv:2311.02692
153 Fu C, Dai Y, Luo Y, et al. Video-mme: the first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. 2025. ArXiv:2405.21075
154 Fang X, Mao K, Duan H, et al. Mmbench-video: a long-form multi-shot benchmark for holistic video understanding. 2024. ArXiv:2406.14515
155 Li K, Wang Y, He Y, et al. Mvbench: a comprehensive multi-modal video understanding benchmark. 2024. ArXiv:2311.17005
156 Wu H, Li D, Chen B, et al. Longvideobench: a benchmark for long-context interleaved video-language understanding. 2024. ArXiv:2407.15754
157 Wang W, He Z, Hong W, et al. Lvbench: an extreme long video understanding benchmark. 2025. ArXiv:2406.08035
158 Hong W, Cheng Y, Yang Z, et al. Motionbench: benchmarking and improving fine-grained video motion understanding for vision language models. 2025. ArXiv:2501.02955
159 Wang B, Zou X, Lin G, et al. Audiobench: a universal benchmark for audio large language models. 2025. ArXiv:2406.16020
160 Yang Q, Xu J, Liu W, et al. Air-bench: benchmarking large audio-language models via generative comprehension. 2024. ArXiv:2402.07729
161 yu Huang C, Lu K H, Wang S H, et al. Dynamic-superb: towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. 2024. ArXiv:2309.09510
162 Li M, Chen X, Zhang C, et al. M3dbench: let's instruct large models with multi-modal 3d prompts. 2023. ArXiv:2312.10763
163 Bai F, Du Y, Huang T, et al. M3d: advancing 3d medical image analysis with multi-modal large language models. 2024. ArXiv:2404.00578
164 Szymanska E, Dusmanu M, Buurlage J W, et al. Space3d-bench: spatial 3d question answering benchmark. 2024. ArXiv:2408.16662
165 Lu P, Mishra S, Xia T, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 2507–2521
166 He Z, Wu X, Zhou P, et al. Cmmu: a benchmark for Chinese multi-modal multi-type question understanding and reasoning. 2024. ArXiv:2401.14011
167 Wang X, Hu Z, Lu P, et al. Scibench: evaluating college-level scientific problem-solving abilities of large language models. 2023. ArXiv:2307.10635
168 Das R J, Hristov S E, Li H, et al. Exams-v: a multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. 2024. ArXiv:2403.10378
169 Yue X, Ni Y, Zhang K, 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, 2024. 9556–9567
170 Yue X, Zheng T, Ni Y, et al. Mmmu-pro: a more robust multi-discipline multimodal understanding benchmark. 2024. ArXiv:2409.02813
171 Phan L, Gatti A, Han Z, et al. Humanity's last exam. 2025. ArXiv:2501.14249
172 Cui H, Shamsi Z, Cheon G, et al. Curie: evaluating LLMs on multitask scientific long context understanding and reasoning. 2025. ArXiv:2503.13517
173 Zhou Y, Wang Y, He X, et al. Scientists' first exam: probing cognitive abilities of MLLM via perception, understanding, and reasoning. 2025. ArXiv:2506.10521
174 Zhou P, Zhang F, Peng X, et al. Mdk12-bench: a multi-discipline benchmark for evaluating reasoning in multimodal large language models. 2025. ArXiv:2504.05782
175 Wang J, Zhang Z, Guo Y, et al. The ever-evolving science exam. 2025. ArXiv:2507.16514
176 Sun R, Chang J, Pearce H, et al. Sok: unifying cybersecurity and cybersafety of multimodal foundation models with an information theory approach. 2024. ArXiv:2411.11195
177 Zhang C, Zhou L, Xu X, et al. Adversarial attacks of vision tasks in the past 10 years: a survey. ACM Comput Surv, 2026, 58: 1–42
178 Liu X, Cui X, Li P, et al. Jailbreak attacks and defenses against multimodal generative models: a survey. 2024. ArXiv:2411.09259
179 Ye M, Rong X, Huang W, et al. A survey of safety on large vision-language models: attacks, defenses and evaluations. 2025. ArXiv:2502.14881
180 Wang J, Zhang H, Yuan Y. 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), 2025. 21001–21010

181 Yin S K, Fu C Y, Zhao S R, et al. Woodpecker: hallucination correction for multimodal large language models. Sci China Inf Sci, 2024, 67: 220105
182 Tu H, Cui C, Wang Z, et al. How many unicorns are in this image? A safety evaluation benchmark for vision LLMs. 2023. ArXiv:2311.16101
183 Luo W, Ma S, Liu X, et al. Jailbreakv-28k: a benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. 2024. ArXiv:2404.03027
184 Liu X, Zhu Y, Lan Y, et al. Query-relevant images jailbreak large multi-modal models. 2023. ArXiv:2311.17600
185 Zhang H, Shao W, Liu H, et al. B-AVIBench: toward evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. IEEE Trans Inform Forensic Secur, 2025, 20: 1434–1446
186 Weng F, Xu Y, Fu C, et al. MMJ-Bench: a comprehensive study on jailbreak attacks and defenses for vision language models. 2024. ArXiv:2408.08464
187 Zheng B, Chen G, Zhong H, et al. Usb: a comprehensive and unified safety evaluation benchmark for multimodal large language models. 2025. ArXiv:2505.23793
188 Gu T, Zhou Z, Huang K, et al. Mllmguard: a multi-dimensional safety evaluation suite for multimodal large language models. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 7256–7295
189 Ying Z, Liu A, Liang S, et al. Safebench: a safety evaluation framework for multimodal large language models. 2024. ArXiv:2410.18927
190 Lee D, Jang J, Jeong J, et al. Are vision-language models safe in the wild? A meme-based benchmark study. 2025. ArXiv:2505.15389
191 Qu Y, Shen X, Wu Y, et al. Unsafebench: benchmarking image safety classifiers on real-world and AI-generated images. 2024. ArXiv:2405.03486
192 Li Y, Du Y, Zhou K, et al. Evaluating object hallucination in large vision-language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023
193 Gunjal A, Yin J, Bas E. Detecting and preventing hallucinations in large vision language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 18135–18143
194 Jiang C, Jia H, Dong M, et al. 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, 2024. 525–534
195 Ding P, Wu J, Kuang J, et al. Hallu-pi: evaluating hallucination in multi-modal large language models within perturbed inputs. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 10707–10715
196 Ye-Bin M, Hyeon-Woo N, Choi W, et al. Beaf: observing before-after changes to evaluate hallucination in vision-language models. In: Proceedings of European Conference on Computer Vision, 2025. 232–248
197 Guan T, Liu F, Wu X, 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, 2024. 14375–14385
198 Wu X, Guan T, Li D, et al. Autohallusion: automatic generation of hallucination benchmarks for vision-language models. 2024. ArXiv:2406.10900
199 Zhang Y, Huang Y, Sun Y, et al. Benchmarking trustworthiness of multimodal large language models: a comprehensive study. 2024. ArXiv:2406.07057
200 Xu C, Zhang J, Chen Z, et al. Mmdt: decoding the trustworthiness and safety of multimodal foundation models. 2025. ArXiv:2503.14827
201 Nayak S, Jain K, Awal R, et al. Benchmarking vision language models for cultural understanding. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 5769–5790
202 Jiang Y, Li Z, Shen X, et al. Modscan: measuring stereotypical bias in large vision-language models from vision and language modalities. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024
203 Wu P, Liu C, Chen C, et al. Fmbench: benchmarking fairness in multimodal large language models on medical tasks. 2024. ArXiv:2410.01089
204 Jin R, Xu Z, Zhong Y, et al. Fairmedfm: fairness benchmarking for medical imaging foundation models. In: Proceedings of the 38th Conference on Neural Information Processing Systems, 2024
205 Luo Y, Shi M, Khan M O, et al. Fairclip: harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 12289–12301
206 Luo W, Zhang Q, Lu T, et al. Doxing via the lens: revealing privacy leakage in image geolocation for agentic multi-modal large reasoning model. 2025. ArXiv:2504.19373
207 Chen Y, Mendes E, Das S, et al. Can language models be instructed to protect personal information? 2023. ArXiv:2310.02224
208 Shi Y, Gao Y, Lai Y, et al. Shield: an evaluation benchmark for face spoofing and forgery detection with multimodal large language models. 2024. ArXiv:2402.04178
209 Chandna B, Aboujenane M, Naseem U. Extremeaigc: benchmarking LMM vulnerability to AI-generated extremist content. 2025. ArXiv:2503.09964
210 Wang S, Long Z, Fan Z, et al. From LLMs to MLLMs: exploring the landscape of multimodal jailbreaking. 2024. ArXiv:2406.14859
211 Guo Y, Jiao F, Nie L, et al. The VLLM safety paradox: dual ease in jailbreak attack and defense. 2024. ArXiv:2411.08410
212 Ying Z, Liu A, Liu X, et al. Unveiling the safety of GPT-4o: an empirical study using jailbreak attacks. 2024. ArXiv:2406.06302
213 Liu X, Zhu Y, Gu J, et al. Mm-safetybench: a benchmark for safety evaluation of multimodal large language models. In: Proceedings of European Conference on Computer Vision, 2024. 386–403
214 Rohrbach A, Hendricks L A, Burns K, et al. Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. 4035–4045
215 Downer G, Craven S, Ruck D, et al. Text2vlm: adapting text-only datasets to evaluate alignment training in visual language models. 2025. ArXiv:2507.20704
216 Li X, Zhou H, Wang R, et al. Mossbench: is your multimodal language model oversensitive to safe queries? 2024. ArXiv:2406.17806
217 Janghorbani S, De Melo G. 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, 2023. 1725–1735
218 Lu P, Bansal H, Xia T, et al. Mathvista: evaluating mathematical reasoning of foundation models in visual contexts. 2023.

ArXiv:2310.02255
219 Gupta H, Verma S, Anantheswaran U, et al. Polymath: a challenging multi-modal mathematical reasoning benchmark. 2024. ArXiv:2410.14702
220 Wang K, Pan J, Shi W, et al. Measuring multimodal mathematical reasoning with math-vision dataset. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 95095–95169
221 He C, Luo R, Bai Y, et al. Olympiadbench: a challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. 2024. ArXiv:2402.14008
222 Wang Y, Zhang P, Tang J, et al. Polymath: evaluating mathematical reasoning in multilingual contexts. 2025. ArXiv:2504.18428
223 Zhang R, Jiang D, Zhang Y, et al. Mathverse: does your multi-modal LLM truly see the diagrams in visual math problems? In: Proceedings of European Conference on Computer Vision, 2024. 169–186
224 Qiao R, Tan Q, Dong G, et al. We-math: does your large multimodal model achieve human-like mathematical reasoning? 2024. ArXiv:2407.01284
225 Zhou M, Liang H, Li T, et al. Mathscape: evaluating MLLMs in multimodal math scenarios through a hierarchical benchmark. 2024. ArXiv:2408.07543
226 Liu W, Pan Q, Zhang Y, et al. Cmm-math: a Chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. 2024. ArXiv:2409.02834
227 Wang P, Li Z Z, Yin F, et al. Mv-math: evaluating multimodal math reasoning in multi-visual contexts. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 19541–19551
228 Kembhavi A, Salvato M, Kolve E, et al. Diagram understanding in geometry questions. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2016
229 Anand A, Kapuriya J, Singh A, et al. Mm-phyqa: multimodal physics question answering with multi-image cot prompting. 2024. ArXiv:2404.08704
230 Wang L, Su E, Liu J, et al. Physunibench: an undergraduate-level physics reasoning benchmark for multimodal models. 2025. ArXiv:2506.17667
231 Dai S, Yan Y, Su J, et al. Physicsarena: the first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. 2025. ArXiv:2505.15472
232 Xiang K, Li H, Zhang T J, et al. Seephys: does seeing help thinking? Benchmarking vision-based physics reasoning. 2025. ArXiv:2505.19099
233 Zhang X, Dong Y, Wu Y, et al. Physreason: a comprehensive benchmark towards physics-based reasoning. 2025. ArXiv:2502.12054
234 Liang Z, Guo K, Liu G, et al. Scemqa: a scientific college entrance level multimodal question answering benchmark. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, 2024. 109–119
235 Yu S, Wu P, Liang P P, et al. PACS: a dataset for physical audiovisual commonsense reasoning. 2022. ArXiv:2203.11130
236 Jassim S, Holubar M, Richter A, et al. GRASP: a novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. 2023. ArXiv:2311.09048
237 Foss A, Evans C, Mitts S, et al. Causalvqa: a physically grounded causal reasoning benchmark for video models. 2025. ArXiv:2506.09943
238 Shabtay N, Polo F M, Doveh S, et al. Livexiv—a multi-modal live benchmark based on arxiv papers content. 2024. ArXiv:2410.10783
239 Kembhavi A, Seo M, Schwenk D, et al. 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), 2017. 4999–5007
240 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci, 1988, 28: 31–36
241 Edwards C, Zhai C, Ji H. Text2mol: cross-modal molecule retrieval with natural language queries. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021. 595–607
242 Zhang D, Liu W, Tan Q, et al. Chemllm: a chemical large language model. 2024. ArXiv:2402.06852
243 Krenn M, Häse F, Nigam A, et al. Self-referencing embedded strings (selfies): a 100% robust molecular string representation. Mach Learn Sci Tech, 2020, 1: 045024
244 Heller S R, McNaught A, Pletnev I, et al. InChI, the IUPAC international chemical identifier. J Cheminform, 2015, 7: 23
245 Liu P, Ren Y, Tao J, et al. GIT-Mol: a multi-modal large language model for molecular science with graph, image, and text. Comput Biol Med, 2024, 171: 108073
246 Cao H, Liu Z, Lu X, et al. Instructmol: multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. 2023. ArXiv:2311.16208
247 Li J, Zhang D, Wang X, et al. Chemvlm: exploring the power of multimodal large language models in chemistry area. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 415–423
248 Liu P, Tao J, Ren Z. A quantitative analysis of knowledge-learning preferences in large language models in molecular science. Nat Mach Intell, 2025, 7: 315–327
249 Alampara N, Schilling-Wilhelmi M, Ríos-García M, et al. Probing the limitations of multimodal language models for chemistry and materials research. 2024. ArXiv:2411.16955
250 Li S, Liu Z, Luo Y, et al. Towards 3d molecule-text interpretation in language models. 2024. ArXiv:2401.13923
251 Bushuiev R, Bushuiev A, de Jonge N, et al. MassSpecGym: a benchmark for the discovery and identification of molecules. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 110010–110027
252 Alberts M, Schilter O, Zipoli F, et al. Unraveling molecular structure: a multimodal spectroscopic dataset for chemistry. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 125780–125808
253 Guo K, Nan B, Zhou Y, et al. Can LLMs solve molecule puzzles? A multimodal benchmark for molecular structure elucidation. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 134721–134746
254 Le K, Guo Z, Dong K, et al. Molx: enhancing large language models for molecular learning with a multi-modal extension. 2024. ArXiv:2406.06777
255 Guo S, Wang L, Jin C, et al. M^3-20M: a large-scale multi-modal molecule dataset for AI-driven drug design and discovery. J Bioinform Comput Biol, 2025, 23: 2550006
256 Luo J, Kou Z, Yang L, 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

257 Xue S, Li X, Zhou F, et al. Famma: a benchmark for financial domain multilingual multimodal question answering. 2024. ArXiv:2410.04526
258 Gan Z, Lu Y, Zhang D, et al. Mme-finance: a multimodal finance benchmark for expert-level understanding and reasoning. 2024. ArXiv:2411.03314
259 Peng X, Qian L, Wang Y, et al. Multifinben: a multilingual, multimodal, and difficulty-aware benchmark for financial LLM evaluation. 2025. ArXiv:2506.14028
260 Zeng L, Lou F, Wang Z, et al. Fingaia: an end-to-end benchmark for evaluating AI agents in finance. 2025. ArXiv:2507.17186
261 Li J, Zhu Y, Cheng D, et al. Cfbenchmark-mm: Chinese financial assistant benchmark for multimodal large language model. 2025. ArXiv:2506.13055
262 Tang Z, Liu J, Yang Z, et al. Finmmr: make financial numerical reasoning more multimodal, comprehensive, and challenging. 2025. ArXiv:2508.04625
263 Rangapur A, Wang H, Jian L, et al. Fin-fact: a benchmark dataset for multimodal financial fact-checking and explanation generation. In: Proceedings of Companion Proceedings of the ACM Web Conference, 2025. 785–788
264 Kim S, Kim C, Kim T. 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
265 Bhatia G, Nagoudi E M B, Cavusoglu H, et al. Fintral: a family of GPT-4 level multimodal financial large language models. 2024. ArXiv:2402.10986
266 Huang J, Xiao M, Li D, et al. Open-finllms: open multimodal large language models for financial applications. 2024. ArXiv:2408.11878
267 Wu S, Koo M, Blum L, et al. A comparative study of open-source large language models, GPT-4 and Claude 2: multiple-choice test taking in nephrology. 2023. ArXiv:2308.04709
268 Liu M, Ding J, Xu J, et al. Medbench: a comprehensive, standardized, and reliable benchmarking system for evaluating Chinese medical large language models. 2024. ArXiv:2407.10990
269 Chen H, Fang Z, Singla Y, et al. Benchmarking large language models on answering and explaining challenging medical questions. 2024. ArXiv:2402.18060
270 Krithara A, Nentidis A, Bougiatiotis K, et al. BioASQ-QA: a manually curated corpus for biomedical question answering. Sci Data, 2023, 10: 170
271 Zhang N, Chen M, Bi Z, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022
272 Jin Q, Dhingra B, Liu Z, et al. Pubmedqa: a dataset for biomedical research question answering. 2019. ArXiv:1909.06146
273 Trop E, Schiff Y, Marroquin E M, et al. The genomics long-range benchmark: advancing DNA language models. In: Proceedings of ICLR, 2024
274 Hu Y, Li T, Lu Q, et al. Omnimedvqa: a new large-scale comprehensive evaluation benchmark for medical lvlm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 22170–22183
275 Ye J, Wang G, Li Y, et al. Gmai-mmbench: a comprehensive multimodal evaluation benchmark towards general medical AI. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 94327–94427
276 Zuo Y, Qu S, Li Y, et al. Medxpertqa: benchmarking expert-level medical reasoning and understanding. 2025. ArXiv:2501.18362
277 Lau J J, Gayen S, Ben Abacha A, et al. A dataset of clinically generated visual questions and answers about radiology images. Sci Data, 2018, 5: 180251
278 He X, Zhang Y, Mou L, et al. Pathvqa: 30000+ questions for medical visual question answering. 2020. ArXiv:2003.10286
279 Liu B, Zhan L M, Xu L, et al. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: Proceedings of 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021. 1650–1654
280 Ji Y, Bai H, Ge C, et al. Amos: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 36722–36732
281 Zheng Q, Zhao W, Wu C, et al. Large-scale long-tailed disease diagnosis on radiology images. Nat Commun, 2024, 15: 10147
282 Hou W, Ji Z. Geneturing tests GPT models in genomics. BioRxiv, 2023, doi: 10.1101/2023.03.11.532238
283 Arora R K, Wei J, Hicks R S, et al. Healthbench: evaluating large language models towards improved human health. 2025. ArXiv:2505.08775
284 Li Y, Liu J, Zhang T, et al. Baichuan-Omni-1.5 Technical Report. 2025. ArXiv:2501.15368
285 Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods, 2015, 12: 931–934
286 Howe K L, Achuthan P, Allen J, et al. Ensembl 2021. Nucleic Acids Res, 2021, 49: D884–D891
287 Yin M, Qu Y, Liu D, et al. Genome-bench: a scientific reasoning benchmark from real-world expert discussions. 2025. ArXiv:2505.19501
288 Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med, 2024, 30: 2613–2622
289 Sellergren A, Kazemzadeh S, Jaroensri T, et al. Medgemma technical report. 2025. ArXiv:2507.05201
290 Si C, Zhang Y, Li R, et al. Design2code: benchmarking multimodal code generation for automated front-end engineering. 2024. ArXiv:2403.03163
291 Yun S, Thushara R, Bhat M, et al. Web2code: a large-scale webpage-to-code dataset and evaluation framework for multimodal LLMs. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 112134–112157
292 Wu C, Ge Y, Guo Q, et al. Plot2code: a comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. 2024. ArXiv:2405.07990
293 Yang C, Shi C, Liu Y, et al. Chartmimic: evaluating LMM's cross-modal reasoning capability via chart-to-code generation. 2024. ArXiv:2406.09961
294 Wang H, Zhou X, Xu Z, et al. Code-vision: evaluating multimodal LLMs logic understanding and code generation capabilities. 2025. ArXiv:2502.11829
295 Yang J, Jimenez C E, Zhang A L, et al. Swe-bench multimodal: do AI systems generalize to visual software domains? 2024. ArXiv:2410.03859
296 Zhang L, Zan D, Yang Q, et al. Codev: issue resolving with visual data. 2024. ArXiv:2412.17315
297 Li K, Tian Y, Hu Q, et al. Mmcode: benchmarking multimodal large language models for code generation with visually rich programming problems. 2024. ArXiv:2404.09486
298 Zhang F, Wu L, Bai H, et al. Humaneval-v: benchmarking high-level visual reasoning with complex diagrams in coding

tasks. 2024. ArXiv:2410.12381
299 Rodriguez J, Jian X, Panigrahi S S, et al. Bigdocs: an open dataset for training multimodal models on document and code tasks. 2024. ArXiv:2412.04626
300 Chai L, Yang J, Liu S, et al. Multilingual multimodal software developer for code generation. 2025. ArXiv:2507.08719
301 Sachdeva E, Agarwal N, Chundi S, et al. Rank2tell: a multimodal driving dataset for joint importance ranking and reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024. 7513–7522
302 Malla S, Choi C, Dwivedi I, et al. Drama: joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023. 1043–1052
303 Qian T, Chen J, Zhuo L, et al. Nuscenes-qa: a multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 4542–4550
304 Marcu A M, Chen L, Hünermann J, et al. Lingoqa: visual question answering for autonomous driving. In: Proceedings of European Conference on Computer Vision, 2024. 252–269
305 Chiu H k, Hachiuma R, Wang C Y, et al. V2v-llm: vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. 2025. ArXiv:2502.09980
306 Guo X, Zhang R, Duan Y, et al. Surds: benchmarking spatial understanding and reasoning in driving scenarios with vision language models. 2024. ArXiv:2411.13112
307 Wei Z, Qiang C, Jiang B, et al. Ad^2-bench: a hierarchical cot benchmark for MLLM in autonomous driving under adverse conditions. 2025. ArXiv:2506.09557
308 Hao Y, Li Z, Sun L, et al. Driveaction: a benchmark for exploring human-like driving decisions in VLA models. 2025. ArXiv:2506.05667
309 Ishaq A, Lahoud J, More K, et al. Drivelmm-o1: a step-by-step reasoning dataset and large multimodal model for driving scenario understanding. 2025. ArXiv:2503.10621
310 Tian X, Gu J, Li B, et al. Drivevlm: the convergence of autonomous driving and large vision-language models. 2024. ArXiv:2402.12289
311 Xiao B, Feng C, Huang Z, et al. Robotron-sim: improving real-world driving via simulated hard-case. 2025. ArXiv:2508.04642
312 Lu Y, Yao Y, Tu J, et al. Can lvlms obtain a driver's license? A benchmark towards reliable AGI for autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 5838–5846
313 Li Y, Tian M, Lin Z, et al. Fine-grained evaluation of large vision-language models in autonomous driving. 2025. ArXiv:2503.21505
314 Rekanar K, Joyce J M, Hayes M, et al. DriVQA: a gaze-based dataset for visual question answering in driving scenarios. Data Brief, 2025, 59: 111367
315 Wen L, Yang X, Fu D, et al. On the road with GPT-4v (ision): explorations of utilizing visual-language model as autonomous driving agent. In: Proceedings of ICLR, 2024
316 Cao X, Zhou T, Ma Y, et al. 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), 2024. 21819–21830
317 Deng C, Zhang T, He Z, 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, 2024. 161–170
318 Manivannan V V, Jafari Y, Eranky S, et al. Climaqa: an automated evaluation framework for climate foundation models. 2024. ArXiv:2410.16701
319 Webersinke N, Kraus M, Bingler J A, et al. Climatebert: a pretrained language model for climate-related text. 2021. ArXiv:2110.12010
320 Ma C, Hua Z, Anderson-Frey A, et al. Weatherqa: can multimodal language models reason about severe weather? 2024. ArXiv:2406.11217
321 Hu Y, Yuan J, Wen C, et al. RSGPT: a remote sensing vision language model and benchmark. ISPRS J Photogrammetry Remote Sens, 2025, 224: 272–286
322 Zhao X, Xu W, Liu B, et al. MSEarth: a benchmark for multimodal scientific comprehension of earth science. 2025. ArXiv:2505.20740
323 Lu X, Wang B, Zheng X, et al. Exploring models and data for remote sensing image caption generation. IEEE Trans Geosci Remote Sens, 2017, 56: 2183–2195
324 Zhan Y, Xiong Z, Yuan Y. Rsvg: exploring data and models for visual grounding on remote sensing data. IEEE Trans Geosci Remote Sensing, 2023, 61: 1–13
325 Li X, Ding J, Elhoseiny M. Vrsbench: a versatile vision-language benchmark dataset for remote sensing image understanding. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 3229–3242
326 Luo J, Zhang Y, Yang X, et al. When large vision-language model meets large remote sensing imagery: coarse-to-fine text-guided token pruning. 2025. ArXiv:2503.07588
327 Kuckreja K, Danish M S, Naseer M, et al. Geochat: grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 27831–27840
328 Xu W, Zhao X, Zhou Y, et al. Earthse: a benchmark evaluating earth scientific exploration capability for large language models. 2025. ArXiv:2505.17139
329 An X, Sun J, Gui Z, et al. Choice: benchmarking the remote sensing capabilities of large vision-language models. 2024. ArXiv:2411.18145
330 Ma Z, Xiao X, Dong S, et al. Sarchat-bench-2m: a multi-task vision-language benchmark for SAR image interpretation. 2025. ArXiv:2502.08168
331 Muhtar D, Li Z, Gu F, et al. Lhrs-bot: empowering remote sensing with VGI-enhanced large multimodal language model. In: Proceedings of European Conference on Computer Vision. Springer, 2024. 440–457
332 Bi Z, Zhang N, Xue Y, et al. Oceangpt: a large language model for ocean science tasks. 2023. ArXiv:2310.02031
333 Luo J, Pang Z, Zhang Y, et al. Skysensegpt: a fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. 2024. ArXiv:2406.10100
334 Zhang C, Wang S. Good at captioning, bad at counting: benchmarking GPT-4v on earth observation data. 2024. ArXiv:2401.17600
335 Danish M S, Munir M A, Shah S R A, et al. Geobench-vlm: benchmarking vision-language models for geospatial tasks. 2024. ArXiv:2411.19325

336 Mall U, Phoo C, Liu M, et al. Remote sensing vision-language foundation models without annotations via ground remote alignment. 2023. ArXiv:2312.06960
337 Wang F, Chen M, He X, et al. Omniearth-bench: towards holistic evaluation of earth's six spheres and cross-spheres interactions with multimodal observational earth data. 2025. ArXiv:2505.23522
338 Cheng Q, Huang H, Xu Y, et al. Nwpu-captions dataset and mlca-net for remote sensing image captioning. IEEE Trans Geosci Remote Sensing, 2022, 60: 1–19
339 Lobry S, Marcos D, Murray J, et al. RSVQA: visual question answering for remote sensing data. IEEE Trans Geosci Remote Sens, 2020, 58: 8555–8566
340 Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3674–3683
341 Qi Y, Wu Q, Anderson P, et al. Reverie: remote embodied visual referring expression in real indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 9982–9991
342 Grauman K, Westbury A, Byrne E, 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, 2022. 18995–19012
343 Datta S, Dharur S, Cartillier V, et al. Episodic memory question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 19119–19128
344 Ma X, Yong S, Zheng Z, et al. Sqa3d: Situated question answering in 3d scenes. 2022. ArXiv:2210.07474
345 Majumdar A, Ajay A, Zhang X, 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, 2024. 16488–16498
346 Ren A Z, Clark J, Dixit A, et al. Explore until confident: efficient exploration for embodied question answering. 2024. ArXiv:2403.15941
347 Chen Z, Shi Z, Lu X, et al. Rh20t-p: a primitive-level robotic dataset towards composable generalization agents. 2024. ArXiv:2403.19622
348 Jia B, Lei T, Zhu S C, et al. Egotaskqa: understanding human tasks in egocentric videos. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 3343–3360
349 Wang T, Mao X, Zhu C, 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, 2024. 19757–19767
350 Cheng Z, Tu Y, Li R, et al. Embodiedeval: evaluate multimodal LLMs as embodied agents. 2025. ArXiv:2501.11858
351 Jiang K, Liu Y, Chen W, et al. Beyond the destination: a novel benchmark for exploration-aware embodied question answering. 2025. ArXiv:2503.11117
352 Yang R, Chen H, Zhang J, et al. Embodiedbench: comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. 2025. ArXiv:2502.09560
353 Zhang S, Xu Z, Liu P, et al. Vlabench: a large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. 2024. ArXiv:2412.18194
354 Yue H, Huang S, Liao Y, et al. Ewmbench: evaluating scene, motion, and semantic quality in embodied world models. 2025. ArXiv:2505.09694
355 Damen D, Doughty H, Farinella G M, et al. Scaling egocentric vision: the epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 720–736
356 Mees O, Hermann L, Rosete-Beas E, et al. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robot Autom Lett, 2022, 7: 7327–7334
357 Burns A, Arsan D, Agrawal S, et al. A dataset for interactive vision-language navigation with unknown command feasibility. In: Proceedings of European Conference on Computer Vision, 2022. 312–328
358 Shridhar M, Thomason J, Gordon D, et al. Alfred: a benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10740–10749
359 Das A, Datta S, Gkioxari G, et al. Embodied question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1–10
360 Frellsen T Q J, Zhang K. Neurips 2025 embodied agent interface challenge. 2025. https://blog.neurips.cc/2025/06/27/neurips-2025-competitions-announced/
361 Wang Z J, Montoya E, Munechika D, et al. Diffusiondb: a large-scale prompt gallery dataset for text-to-image generative models. 2022. ArXiv:2210.14896
362 Wu X, Sun K, Zhu F, et al. Human preference score: better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. 2096–2105
363 Xu J, Liu X, Wu Y, et al. Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of Advances in Neural Information Processing Systems, 2024
364 Kirstain Y, Polyak A, Singer U, et al. Pick-a-pic: an open dataset of user preferences for text-to-image generation. In: Proceedings of Advances in Neural Information Processing Systems, 2024
365 Zhang Z, Li C, Sun W, et al. A perceptual quality assessment exploration for AIGC images. In: Proceedings of 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2023. 440–445
366 Li C, Zhang Z, Wu H, et al. AGIQA-3K: an open database for AI-generated image quality assessment. IEEE Trans Circuits Syst Video Technol, 2024, 34: 6833–6846
367 Wang J, Duan H, Liu J, et al. Aigciqa2023: a large-scale image quality assessment database for AI generated images: from the perspectives of quality, authenticity and correspondence. In: Proceedings of CAAI International Conference on Artificial Intelligence, 2023. 46–57
368 Chen Z, Sun W, Wu H, et al. Exploring the naturalness of AI-generated images. 2024. ArXiv:2312.05476
369 Li C, Kou T, Gao Y, et al. 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
370 Yang L, Duan H, Teng L, et al. Aigcoiqa2024: perceptual quality assessment of AI generated omnidirectional images. 2024. ArXiv:2404.01024
371 Li C, Wu X, Wu H, et al. Cmc-bench: towards a new paradigm of visual signal compression. 2024. ArXiv:2406.09356
372 Yuan J, Cao X, Li C, et al. Pku-i2iqa: an image-to-image quality assessment database for AI generated images. 2023. ArXiv:2311.15556
373 Yarom M, Bitton Y, Changpinyo S, et al. What you see is what you read? Improving text-image alignment evaluation. 2023. ArXiv:2305.10400
374 Wang J, Duan H, Zhai G, et al. Quality assessment for AI generated images with instruction tuning. 2025. ArXiv:2405.07346

375 Zhang Z, Kou T, Wang S, et al. Q-eval-100k: evaluating visual quality and alignment level for text-to-vision content. 2025. ArXiv:2503.02357
376 Wu X, Hao Y, Sun K, et al. Human preference score v2: a solid benchmark for evaluating human preferences of text-to-image synthesis. 2023. ArXiv:2306.09341
377 Chivileva I, Lynch P, Ward T E, et al. Measuring the quality of text-to-video model outputs: metrics and dataset. 2023. ArXiv:2309.08009
378 Liu Y, Cun X, Liu X, et al. Evalcrafter: benchmarking and evaluating large video generation models. 2023. ArXiv:2310.11440
379 Liu Y, Li L, Ren S, et al. Fetv: a benchmark for fine-grained evaluation of open-domain text-to-video generation. 2023. ArXiv:2311.01813
380 Huang Z, He Y, Yu J, et al. Vbench: comprehensive benchmark suite for video generative models. 2023. ArXiv:2311.17982
381 Kou T, Liu X, Zhang Z, et al. Subjective-aligned dataset and metric for text-to-video quality assessment. 2024. ArXiv:2403.11956
382 Chen Z, Sun W, Tian Y, et al. Gaia: rethinking action quality assessment for ai-generated videos. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2024
383 Wang J, Duan H, Zhai G, et al. 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, 2025. 18869–18880
384 Wang J, Duan H, Jia Z, et al. Love: benchmarking and evaluating text-to-video generation and video-to-text interpretation. 2025. ArXiv:2505.12098
385 Zhang Z, Sun W, Li X, et al. Human-activity AGV quality assessment: a benchmark dataset and an objective evaluation metric. 2024. ArXiv:2411.16619
386 Wang J, Wang J, Duan H, et al. Tdve-assessor: benchmarking and evaluating the quality of text-driven video editing with LMMS. 2025. ArXiv:2505.19535
387 Cao Y, Min X, Gao Y, et al. Agav-rater: adapting large multimodal model for AI-generated audio-visual quality assessment. 2025. ArXiv:2501.18314
388 Chen C, Hu Y, Wang S, et al. Audio large language models can be descriptive speech quality evaluators. 2025. ArXiv:2501.17202
389 Łajszczak M, Cámbara G, Li Y, et al. Base TTS: lessons from building a billion-parameter text-to-speech model on 100k hours of data. 2024. ArXiv:2402.08093
390 Wang X, Zhao Z, Ren S, et al. Audio Turing test: benchmarking the human-likeness of large language model-based text-to-speech systems in Chinese. 2025. ArXiv:2505.11200
391 Minixhofer C, Klejch O, Bell P. TTSDS2: resources and benchmark for evaluating human-quality text to speech systems. 2025. ArXiv:2506.19441
392 Zhang Y, Cui B, Yang Q, et al. Benchmarking and learning multi-dimensional quality evaluator for text-to-3D generation. 2024. ArXiv:2412.11170
393 Zhou Y, Zhang Z, Wen F, et al. 3dgcqa: a quality assessment database for 3d ai-generated contents. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. 1–5
394 Fu K, Duan H, Zhang Z, et al. Multi-dimensional quality assessment for text-to-3D assets: dataset and model. IEEE Trans Multimedia, 2025, doi: 10.1109/tmm.2025.3604905
395 Fu K, Duan H, Zhang Z, et al. Si23dcqa: perceptual quality assessment of single image-to-3d content. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2025
396 Xing Y, Wang J, Niu P, et al. 3dgs-ieval-15k: a large-scale image quality evaluation database for 3d Gaussian-splatting. 2025. ArXiv:2506.14642
397 Lin Z, Pathak D, Li B, et al. Evaluating text-to-visual generation with image-to-text generation. In: Proceedings of European Conference on Computer Vision, 2024. 366–384
398 Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 29
399 Unterthiner T, van Steenkiste S, Kurach K, et al. Towards accurate generative models of video: a new metric & challenges. 2018. ArXiv:1812.01717
400 Liu X, Min X, Zhai G, et al. Ntire 2024 quality assessment of AI-generated content challenge. 2024. ArXiv:2404.16687
401 Wu H, Zhang Z, Zhang E, et al. Q-instruct: improving low-level visual abilities for multi-modality foundation models. 2023. ArXiv:2311.06783
402 Wu H, Zhang Z, Zhang W, et al. Q-align: teaching LMMS for visual scoring via discrete text-defined levels. 2023. ArXiv:2312.17090
403 Zhang Z, Wu H, Ji Z, et al. Q-boost: on visual quality assessment ability of low-level multi-modality foundation models. 2023. ArXiv:2312.15300
404 You Z, Li Z, Gu J, et al. Depicting beyond scores: advancing image quality assessment through multi-modal language models. In: Proceedings of European Conference on Computer Vision, 2024. 259–276
405 Cui C, Chen K, Wei Z, et al. M3-agiqa: multimodal, multi-round, multi-aspect AI-generated image quality assessment. 2025. ArXiv:2502.15167
406 Li C, Wu H, Zhang Z, et al. Q-refine: a perceptual quality refiner for AI-generated image. 2024. ArXiv:2401.01117
407 Wang P, Sun W, Zhang Z, et al. Large multi-modality model assisted AI-generated image quality assessment. 2024. ArXiv:2404.17762
408 Yu Z, Guan F, Lu Y, et al. 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, 2024. 6692–6701
409 Li Q, Yan Q, Huang H, et al. Text-visual semantic constrained AI-generated image quality assessment. 2025. ArXiv:2507.10432
410 Xia J, He L, Gao F, et al. AI-generated image quality assessment based on task-specific prompt and multi-granularity similarity. 2024. ArXiv:2411.16087
411 Yang J, Fu J, Zhang W, et al. Moe-agiqa: mixture-of-experts boosted visual perception-driven and semantic-aware quality assessment for AI-generated images. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024. 6395–6404
412 Zhou T, Tan S, Zhou W, et al. Adaptive mixed-scale feature fusion network for blind AI-generated image quality assessment. 2024. ArXiv:2404.15163
413 Yuan J, Cao X, Cao L, et al. Pscr: patches sampling-based contrastive regression for AIGC image quality assessment. 2023.

ArXiv:2312.05897
414 Yuan J, Cao X, Che J, et al. Tier: text-image encoder-based regression for AIGC image quality assessment. 2024. ArXiv:2401.03854
415 Peng F, Fu H, Ming A, et al. AIGC image quality assessment via image-prompt correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 6432–6441
416 Zhao X, Zhang P, Tang K, et al. Envisioning beyond the pixels: benchmarking reasoning-informed visual editing. 2025. ArXiv:2504.02826
417 Fang R, Duan C, Wang K, et al. GoT: unleashing reasoning capability of multimodal large language model for visual generation and editing. 2025. ArXiv:2503.10639
418 Huang Y, Xie L, Wang X, et al. 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), 2024. 8362–8371
419 Niu Y, Ning M, Zheng M, et al. WISE: a world knowledge-informed semantic evaluation for text-to-image generation. 2025. ArXiv:2503.07265
420 Wu Y, Li Z, Hu X, et al. KRIS-Bench: benchmarking next-level intelligent image editing models. 2025. ArXiv:2505.16707
421 Kang M, Zhang X, Wei F, et al. Enhancing image editing with chain-of-thought reasoning and multimodal large language models. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, 2025
422 Zhu H, Sui X, Chen B, et al. 2AFC prompting of large multimodal models for image quality assessment. IEEE Trans Circuits Syst Video Technol, 2024, 34: 12873–12878
423 Zhu H, Chen B, Zhu L, et al. Video quality assessment for spatio-temporal resolution adaptive coding. IEEE Trans Circuits Syst Video Technol, 2024, 34: 6403–6415
424 Zhu H, Chen B, Zhu L, et al. Deepdc: deep distance correlation as a perceptual image quality evaluator. 2022. ArXiv:2211.04927
425 Ge Q, Sun W, Zhang Y, et al. LMM-VQA: advancing video quality assessment with large multimodal models. IEEE Trans Circuits Syst Video Technol, 2025, 35: 11083–11096
426 Duan H, Hu Q, Wang J, et al. Finevq: fine-grained user generated content video quality assessment. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 3206–3217
427 Jia Z, Zhang Z, Qian J, et al. Vqa2: visual question answering for video quality assessment. 2024. ArXiv:2411.03795
428 Jia Z, Zhang Z, Zhang Z, et al. Scaling-up perceptual video quality assessment. 2025. ArXiv:2505.22543
429 Cao L, Sun W, Zhang K, et al. Breaking annotation barriers: Generalized video quality assessment via ranking-based self-supervision. 2025. ArXiv:2505.03631
430 Zhu H, Wu H, Li Y, et al. Adaptive image quality assessment via teaching large multimodal model to compare. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 32611–32629
431 Zhang X, Li W, Zhao S, et al. Vq-insight: teaching VLMs for AI-generated video quality understanding via progressive visual reinforcement learning. 2025. ArXiv:2506.18564
432 Zhou Y, Cao J, Zhang Z, et al. Who is a better talker: subjective and objective quality assessment for AI-generated talking heads. 2025. ArXiv:2507.23343
433 Zhou Y, Zhang Z, Sun W, et al. Thqa: a perceptual quality assessment database for talking heads. 2024. ArXiv:2404.09003
434 Zhou Y, Zhang Z, Jia J, et al. Who is a better imitator: subjective and objective quality assessment of animated humans. IEEE Trans Circuits Syst Video Technol, 2025, 35: 10047–10058
435 Zhou Y, Zhang Z, Wu S, et al. MI3S: a multimodal large language model assisted quality assessment framework for AI-generated talking heads. Inf Processing Manage, 2026, 63: 104321
436 Zhou Y, Chen Y, Bi K, et al. An implementation of multimodal fusion system for intelligent digital human generation. 2023. ArXiv:2310.20251
437 Lo C C, Fu S W, Huang W C, et al. Mosnet: deep learning based objective assessment for voice conversion. 2019. ArXiv:1904.08352
438 Zezario R E, Chen Y W, Fu S W, et al. A study on incorporating whisper for robust speech assessment. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2024. 1–6
439 Li Z, Li W. MOSLight: a lightweight data-efficient system for non-intrusive speech quality assessment. In: Proceedings of Interspeech, 2023. 5386–5390
440 Leng Y, Tan X, Zhao S, et al. MBNet: MOS prediction for synthesized speech with mean-bias network. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021. 391–395
441 Liang X, Cumlin F, Schüldt C, et al. DeePMOS: deep posterior mean-opinion-score of speech. In: Proceedings of Interspeech, 2023. 526–530
442 Huang W C, Cooper E, Yamagishi J, et al. LDNet: unified listener dependent modeling in MOS prediction for synthetic speech. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2022. 896–900
443 Liang Q, Shen Y, Chen T, et al. ADTMOS—synthesized speech quality assessment based on audio distortion tokens. IEEE Trans Audio Speech Lang Process, 2025, 33: 1493–1507
444 Wang H, Zhao S, Zhou J, et al. Uncertainty-aware mean opinion score prediction. 2024. ArXiv:2408.12829
445 Tjandra A, Wu Y C, Guo B, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. 2025. ArXiv:2502.05139
446 Ren W, Lin Y C, Huang W C, et al. HighRateMOS: sampling-rate aware modeling for speech quality assessment. 2025. ArXiv:2506.21951
447 Chiang C H, Wang X, Lin C C, et al. Audio-aware large language models as judges for speaking styles. 2025. ArXiv:2506.05984
448 Tang C, Yu W, Sun G, et al. Salmonn: towards generic hearing abilities for large language models. In: Proceedings of International Conference on Learning Representations (ICLR), 2024
449 Chu Y, Xu J, Zhou X, et al. Qwen-audio: advancing universal audio understanding via unified large-scale audio-language models. 2023. ArXiv:2311.07919
450 Chu Y, Xu J, Yang Q, et al. Qwen2-audio technical report. 2024. ArXiv:2407.10759
451 Wang S, Yu W, Yang Y, et al. Enabling auditory large language models for automatic speech quality evaluation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2025. 1–5
452 Wang S, Yu W, Chen X, et al. Qualispeech: a speech quality assessment dataset with natural language reasoning and

descriptions. 2025. ArXiv:2503.20290
453 Wang S, Székely É. Evaluating text-to-speech synthesis from a large discrete token-based speech language model. 2024. ArXiv:2405.09768
454 Manku R R, Tang Y, Shi X, et al. EmergentTTS-Eval: evaluating TTS models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. 2025. ArXiv:2505.23009
455 Huang K, Tu Q, Fan L, et al. InstructTTSEval: benchmarking complex natural-language instruction following in text-to-speech systems. 2025. ArXiv:2506.16381
456 Huang W C, Cooper E, Toda T. Mos-bench: benchmarking generalization abilities of subjective speech quality assessment models. 2024. ArXiv:2411.03715
457 Zhang Z, Sun W, Min X, et al. No-reference quality assessment for 3D colored point cloud and mesh models. IEEE Trans Circuits Syst Video Technol, 2022, 32: 7618–7631
458 Zhang Z, Sun W, Min X, et al. Mm-pcqa: multi-modal learning for no-reference point cloud quality assessment. 2022. ArXiv:2209.00244
459 Su S, Cai X, Gao L, et al. Gt23d-bench: a comprehensive general text-to-3d generation benchmark. 2024. ArXiv:2412.09997
460 Qu Q, Liang H, Chen X, et al. NeRF-NQA: no-reference quality assessment for scenes generated by NeRF and neural view synthesis methods. IEEE Trans Visual Comput Graphics, 2024, 30: 2129–2139
461 Xing Y, Yang Q, Yang K, et al. Explicit-nerf-qa: a quality assessment database for explicit nerf model compression. In: Proceedings of 2024 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2024. 1–5
462 Martin P, Rodrigues A, Ascenso J, et al. Nerf-qa: neural radiance fields quality assessment database. In: Proceedings of 2023 15th International Conference on Quality of Multimedia Experience (QoMEX), 2023. 107–110
463 Martin P, Rodrigues A, Ascenso J, et al. Nerf view synthesis: subjective quality assessment and objective metrics evaluation. 2024. ArXiv:2405.20078
464 Martin P, Rodrigues A, Ascenso J, et al. Gs-qa: comprehensive quality assessment benchmark for Gaussian splatting view synthesis. 2025. ArXiv:2502.13196
465 Wu T, Yang G, Li Z, et al. 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, 2024. 22227–22238
466 Duggal S, Hu Y, Michel O, et al. Eval3d: Interpretable and fine-grained evaluation for 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025. 13326–13336
467 Zhang Y, Zhang M, Wu T, et al. 3dgen-bench: comprehensive benchmark suite for 3d generative models. 2025. ArXiv:2503.21745
468 Zhang Z, Wu H, Zhou Y, et al. LMM-PCQA: assisting point cloud quality assessment with large multimodal models. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2024. 1234–1243
469 Duan H, Yang J, Qiao Y, et al. Vlmevalkit: an open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM international conference on multimedia, 2024. 11198–11201
470 Jiang D, Ku M, Li T, et al. Genai Arena: an open evaluation platform for generative models. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 79889–79908
471 Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann New York Acad Sci, 2023, 1525: 140–146
472 White C, Dooley S, Roberts M, et al. Livebench: a challenging, contamination-limited LLM benchmark. 2025. ArXiv:2406.19314
473 Epoch-AI. AI performance on a set of expert-level mathematics problems. https://epoch.ai/data/ai-benchmarking-dashboard, 2025
474 AGI-Eval. Large language model leaderboard. 2025. https://agi-eval.cn/mvp/listSummaryIndex
475 RELM. Really reliable live evaluation for LLM. 2025. https://nonelinear.com/static/benchmarking.html
476 Xu L, Li A, Zhu L, et al. Superclue: a comprehensive Chinese large language model benchmark. 2023. ArXiv:2307.15020