Review

Not peer-reviewed version

Vision–Language Foundation Models and Multimodal Large Language Models: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges

Gurpreet Singh *, Lamia Qamar, Nicholas Valentino Volta, Amruta Velamuri, Aya Khanyile

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0467.v2

Keywords: vision-language models (VLMs); multimodal large language models (MLLMs); cross-modal alignment; visual question answering (VQA); self supervised multimodal learning; contrastive visionlanguage pretraining; vision transformers; multimodal fusion; foundation models; multimodal reasoning

 doi:10.20944/preprints202602.0467.v2

Review

Vision–Language Foundation Models and Multimodal Large Language Models: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges

Gurpreet Singh [1,*], Lamia Qamare [2], Nicholas Valentino Volta [3], Amruta Velamuri [4] and Aya Khanyile [5]

[1] Graduated - Endicott College of International Studies, Woosong University, Republic of Korea
[2] Graduated - Heriot-Watt University, Edinburgh, UK
[3] Student, Flex Department, Florida Virtual School, USA
[4] Student, CTE Department, Rouse High School, United States of America
[5] Student, College of Law, University of South Africa, South Africa
* Correspondence: gurpreetsinghmcse@gmail.com

Abstract

Vision-based multimodal learning has experienced rapid advancement through the integration of large-scale vision-language models (VLMs) and multimodal large language models (MLLMs). In this review, we adopt a historical and task-oriented perspective to systematically examine the evolution of multimodal vision models from early visual-semantic embedding frameworks to modern instruction-tuned MLLMs. We categorize model developments across major architectural paradigms, including dual-encoder contrastive frameworks, transformer-based fusion architectures, and unified generative models. Further, we analyze their practical implementations across key vision-centric tasks such as image captioning, visual question answering (VQA), visual grounding, and cross-modal generation. Comparative insights are drawn between traditional multimodal fusion strategies and the emerging trend of large-scale multimodal pretraining. We also provide a detailed overview of benchmark datasets, evaluating their representativeness, scalability, and limitations in real-world multimodal scenarios. Building upon this analysis, we identify open challenges in the field, including fine-grained cross-modal alignment, computational efficiency, generalization across modalities, and multimodal reasoning under limited supervision. Finally, we discuss potential research directions such as self-supervised multimodal pretraining, dynamic fusion via adaptive attention mechanisms, and the integration of multimodal reasoning with ethical and human-centered AI principles. Through this comprehensive synthesis of past and present multimodal vision research, we aim to establish a unified reference framework for advancing future developments in visual-language understanding and cross-modal intelligence.

Keywords: vision-language models (VLMs); multimodal large language models (MLLMs); cross-modal alignment; visual question answering (VQA); self-supervised multimodal learning; contrastive vision-language pretraining; vision transformers; multimodal fusion; foundation models; multimodal reasoning

1. Introduction

With the rapid advancement of artificial intelligence and machine learning, multimodal fusion techniques and vision-language models (VLMs) have emerged as critical components driving innovation across diverse sectors. By integrating visual and linguistic modalities, these models enable richer semantic understanding, enhanced interaction, and improved automation in complex tasks. In the healthcare domain, VLMs and Multimodal Large Language Models (MLLMs) have been utilized to interpret medical imagery such as X-rays, CT, and MRI scans in conjunction with radiology

 doi:10.20944/preprints202602.0467.v2

reports, supporting automated diagnosis and clinical decision-making [1]. In agriculture, multimodal fusion approaches are applied to precision farming tasks including crop health monitoring, disease detection, and yield estimation by combining visual crop data with textual or sensor-based contextual information [2]. In the retail and manufacturing industries, VLMs facilitate visual product search, automated tagging, and defect detection by aligning product images with descriptive textual data, improving quality control and recommendation systems [3]. In education, these models enhance multimodal learning environments by generating visual explanations and captions that aid accessibility and support learners with visual impairments [4]. Furthermore, in robotics, vision-language and vision-language-action models empower autonomous systems to jointly process visual scenes and natural language instructions, thereby advancing navigation, manipulation, and human-robot interaction [5]. Collectively, these developments demonstrate that VLMs and MLLMs are not only transforming traditional visual understanding tasks but are also establishing the foundation for cross-domain, context-aware intelligent systems that bridge perception, reasoning, and interaction in real-world environments.

Building on this foundation, the exponential rise of vision-language models (VLMs) and multimodal large language models (MLLMs) has significantly transformed the paradigm of multimodal fusion in recent years. Large-scale pretrained VLMs such as CLIP [6], ALIGN [7], and BLIP-2 [8] possess remarkable cross-modal alignment and generalization capabilities, enabling robust performance in zero-shot image classification and retrieval tasks [6,7]. These models further demonstrate strong potential in instruction-following scenarios, where natural language prompts are seamlessly mapped to visual tasks through multimodal reasoning [8,9]. Similarly, in visual question answering (VQA), transformer-based architectures such as LXMERT [10] and ViLBERT [11] have achieved state-of-the-art results in understanding fine-grained relationships between vision and language [10,11]. This evolution signifies a paradigm shift in robotic vision systems from passive perception toward proactive, semantically aware, and linguistically interactive agents capable of understanding and reasoning about their environment [12,13].

Despite these advancements, several practical challenges persist in deploying multimodal fusion for robotic applications. First, effectively integrating heterogeneous data across visual, textual, and sensory modalities remains a fundamental obstacle, particularly regarding modality alignment, unified feature representation, and spatiotemporal synchronization [14,15]. Second, robotic systems impose stringent constraints on real-time processing and computational efficiency, demanding lightweight yet accurate fusion architectures that balance inference speed and performance [15,16]. Third, although pretrained VLMs exhibit strong generalization capabilities, their adaptability to task-specific robotic environments such as dynamic scene understanding, manipulation, and embodied reasoning remains limited [17]. Addressing these challenges necessitates future research focusing on self-supervised multimodal pretraining, adaptive attention mechanisms for efficient fusion, and domain-specific fine-tuning strategies to enhance robotic perception and interaction in real-world contexts [18,19].

2. Background / Theoretical Foundation

2.1. What Is Multimodality?

The concept of multimodality fundamentally refers to the integration and processing of multiple types of data or semiotic resources, known as modalities, to communicate or process information [20,21]. In the domain of communication and semiotics, Multimodal Discourse is defined as "the combination of different semiotic modes for example, language and music in a communicative artifact or event" [22]. Human communication is inherently multimodal, involving not only language but also other modes such as gesture, gaze, and facial expression. A mode itself is characterized as "a socially and culturally given semiotic resource for making meaning" [20]. Historically, fields like academia favored monomodality (text-only documents), but this has reversed with the rise of modern media and digital tools that incorporate color illustrations, sophisticated layout, and typography. In the realm of Artificial Intelligence (AI), multimodality specifically refers to systems that integrate and process

 doi:10.20944/preprints202602.0467.v2

diverse data streams, such as text, audio, images, or video, to achieve a more holistic understanding of complex inputs [21,23]. Multimodal models, especially large multimodal models (MLLMs), enhance AI capabilities by integrating visual and textual data, mimicking human learning processes. These systems gain richer context and better reasoning skills by combining different forms of information, allowing them to perform complex tasks like Visual Question Answering (VQA), image captioning, and visual dialogue [21,23]. For instance, a multimodal AI system must accurately and efficiently manage different types of information, such as finding relevant images based on a text query or explaining an image's content in natural language. This approach is crucial in applications like robotics, where machines combine inputs from cameras (vision), microphones (sound), and force sensors (touch) to interact effectively with the environment [24]. Recently we also worked on similar interests like [25,26]. We have also worked on using Artificial Intelligence and Multi-Modality in storytelling too [27,28]

2.2. Different Types of Fusion

Multimodal fusion techniques aim to combine data from multiple sources or modalities to generate more accurate and insightful representations [29]. The two fundamental strategies are Early Fusion and Late Fusion, differentiated by the stage at which data integration occurs [30,31]. Early Fusion, also referred to as feature-level fusion [29,30], combines raw data or low-level features from different modalities into a single feature set before inputting them into a single machine learning model [29,31]. This approach captures intricate relationships between modalities and yields rich feature representations [29,31,32]. However, early fusion can result in high-dimensional feature spaces, inflexibility, and significant challenges when dealing with heterogeneous or asynchronous data [29,31,32]. In contrast, Late Fusion, or decision-level fusion, processes each modality independently using separate models, combining the final predictions or outputs only at the decision stage, often using techniques like averaging or voting [29–31]. This method offers modularity and avoids the high dimensionality associated with early fusion [29,31], but it risks missing critical cross-modal interactions that are crucial for complex tasks, as the input modalities are processed separately [29,32]. In the context of vision-language transformers, early fusion may be related to Merged Attention, where unimodal representations are simply concatenated along the sequence dimension [33]. Hybrid Fusion (sometimes called intermediate fusion) combines aspects of both early and late strategies to mitigate their trade-offs [30,31]. Hybrid methods often apply feature-level fusion to modalities that are synchronous in time while using decision-level fusion for the remaining asynchronous modalities [30]. For example, intermediate fusion often utilizes cross-attention layers to dynamically integrate modality-specific representations [34]. Modern examples of advanced fusion include Progressive Fusion, which utilizes backward connections to feed late-stage fused representations back to the early layers of the unimodal feature generators, allowing progressive refinement and bridging the gap between early and late fusion advantages [32]. Another technique is Compound Tokens, which is generated via channel fusion concatenating the output of a cross-attention layer with the original query tokens along the feature (channel) dimension thus avoiding increased token length while benefiting from cross-attention [33]. Similarly, Depth-Breadth Fusion (DBFusion) is a novel feature-fusion architecture that concatenates visual features extracted from different depths (layers) and breadths (prompts) along the channel dimension, serving as a simple yet effective strategy [35].

2.3. Architecture Overview

3. General Transformer Architectures (LLMs)

These models primarily use variants of the original Transformer architecture [36].

3.1. Encoder–Decoder Architecture

This structure processes inputs through an encoder and feeds the representation to a decoder for output generation [37]. The encoder uses self-attention across the full input sequence, while the decoder uses cross-attention and generates tokens sequentially [37].

 doi:10.20944/preprints202602.0467.v2

T5 [38] is a prime example, using a unified text-to-text paradigm for all NLP tasks [38]. Another example is AlexaTM [39]. Some research suggests encoder–decoder models may be advantageous, though scaling decoder-only models can close this performance gap [37].

3.2. Causal Decoder Architecture

This architecture, often used for Natural Language Generation (NLG), lacks an encoder and relies solely on a decoder for output [37]. It employs causal attention, where the prediction of a token depends only on previous time steps [37]. Examples include PaLM [40], GPT-3 [41], BLOOM [42], and LLaMA [43].

3.3. Prefix Decoder Architecture (Non-Causal Decoder)

In this variant, the attention calculation is bidirectional and not strictly dependent only on prior context [37]. An example is U-PaLM, which is described as a non-causal decoder model [44].

3.4. Mixture-of-Experts (MoE)

This is an efficient sparse variation of the Transformer [37]. It incorporates parallel independent experts (typically feed-forward layers) and a router that directs tokens to specific experts [37].

Mathematically, the MoE layer can be expressed as:

$$y = \sum_{i=1}^{N} g_i(x) \cdot E_i(x)$$

where: - $E_i(x)$ denotes the output of the i^{th} expert, - $g_i(x)$ represents the gating function (probability of routing to each expert), - and only a sparse subset of experts is activated per input.

MoE architectures, like PanGu-Σ (1.085 trillion parameters), allow for massive model scaling without proportionate increases in computational cost, as only a fraction of experts are activated per input [37,45].

4. Multimodal Architectures (VLMs)

Vision-Language Models (VLMs), or Multimodal LLMs (MLLMs), combine vision encoders and LLMs to handle both image and text inputs [46]. They are primarily differentiated by their mechanism for multimodal fusion [46].

4.1. Classification by Fusion Mechanism

4.1.1. Dual Encoder Architectures

These models process modalities independently using dedicated encoders before interaction occurs. Fusion often happens via similarity comparison between global feature vectors [46].

CLIP (Contrastive Language–Image Pretraining) is a dual encoder that uses a contrastive objective [46,47], where fusion is achieved via the dot product between global image and text embeddings [47]:

$$s(I,T) = \frac{f_I(I) \cdot f_T(T)}{\|f_I(I)\| \|f_T(T)\|}$$

where f_I and f_T are the image and text encoders, respectively.

4.1.2. Fusion Encoders (Single-Stream)

These architectures perform multimodal interaction early by directly concatenating or summing image and text embeddings and feeding them into shared Transformer layers [48]. Examples include UniTER [48].

 doi:10.20944/preprints202602.0467.v2

4.1.3. Hybrid Methods

These models mix aspects of dual and fusion encoders. ViLBERT (Vision-and-Language BERT) uses a co-attention Transformer with two parallel streams that interact through cross-attentional layers [49].

CoCa (Contrastive Captioner) employs a minimalist encoder–decoder design pretrained jointly with a contrastive loss (as in CLIP) and a generative captioning loss [50]. The decoder layers omit cross-attention in the first half to maintain unimodal text representation before later layers cross-attend to the image encoder for multimodal representations [50].

4.2. Cross-Modal Interaction Mechanisms (Attention Variants)

Multimodal Transformers use various methods to integrate inputs X_A and X_B across modalities, represented by token embeddings $Z^{(A)}$ and $Z^{(B)}$ [51].

4.2.1. Early Summation

Weighted embeddings from different modalities are summed before entering the Transformer layers $Tf(\cdot)$:

$$Z = \alpha Z^{(A)} \oplus \beta Z^{(B)}$$

[51]

4.2.2. Early Concatenation

Token embedding sequences are concatenated and processed by subsequent Transformer layers:

$$Z = C(Z^{(A)}, Z^{(B)})$$

[51]

4.2.3. Cross-Attention (Co-Attention)

Used mainly in multi-stream or hybrid models, where queries from one modality attend to keys and values from another modality:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

[51,52]

4.2.4. Hierarchical Attention (Multi-Stream to One-Stream)

A late fusion method where independent Transformer streams encode inputs Tf_1, Tf_2, and their outputs are concatenated and fused by another Transformer:

$$Z = Tf_3([Tf_1(X_A); Tf_2(X_B)])$$

[51]

4.2.5. Hierarchical Attention (One-Stream to Multi-Stream)

An early interaction method where concatenated inputs pass through a shared Transformer before splitting into separate streams [53].

4.2.6. Cross-Attention to Concatenation

Combines outputs from cross-attention streams, concatenates them, and processes with a final Transformer layer:

 doi:10.20944/preprints202602.0467.v2

$$Z = Tf([Z_{CA}^{(A)}; Z_{CA}^{(B)}])$$

[51]

5. Specific Advanced VLM Architectures

5.1. Flamingo Architecture

Flamingo is designed for visually conditioned autoregressive text generation, capable of handling interleaved images/videos and text prompts [54]. It uses a Vision Encoder, a Perceiver Resampler (to produce fixed visual tokens), and pre-trained, frozen Language Model blocks interleaved with gated cross-attention dense blocks [54]. The cross-attention layers are trained from scratch and attend to the Perceiver Resampler outputs [54].

5.2. LLaVA Architecture

LLaVA models (e.g., LLaVA-v1.5 [55]) efficiently connect pre-trained LLMs (like Vicuna) and visual encoders (like CLIP) via a linear projection layer that maps image features into the LLM word embedding space [55].

6. Multimodal Datasets and Benchmarks

6.1. General and Comprehensive Multimodal Language Model (MLLM) Benchmarks

These benchmarks are designed to evaluate the broad perception, cognition, and integrated capabilities of MLLMs.

Table 1. Overview of General and Comprehensive MLLM Benchmarks.

Benchmark Name	Citation	Key Details & Data Sources
MMBench	[56]	A novel multi-modality benchmark utilizing a meticulously curated dataset and the CircularEval strategy with ChatGPT for robust evaluation.
MME	[57–59]	Measures both perception and cognition abilities across subtasks. It uses the MSCOCO dataset.
MM-Vet	[60]	Devised to study integrated vision-language capabilities, offering insights beyond overall model rankings. It covers 200 items in total.
SEED-Bench	[61]	A comprehensive benchmark featuring multiple-choice questions covering various evaluation dimensions for both image and video modalities.
SEED-Bench-2	[62]	Categorized MLLMs' capabilities into hierarchical levels from L0 to L4.
SEED-Bench-H	[62]	A comprehensive integration of previous SEED-Bench series (SEED-Bench, SEED-Bench-2, SEED-Bench-2-Plus) with 28,000 multiple-choice questions spanning 34 dimensions.
LLaVA-Bench	[62]	Constructed to examine a variety of MLLM capabilities.
LAMM	[63]	Provides a comprehensive assessment of MLLMs' capabilities, particularly in understanding visual prompting instructions.
MDVP-Bench	[64]	Created to provide a comprehensive assessment of MLLMs' capabilities, particularly in understanding visual prompting instructions.
ChEF	[65]	Constructed as a standardized and holistic evaluation framework.
UniBench	[66]	Constructed as a standardized and holistic evaluation framework.
TouchStone	[67]	Proposed to support open-ended answers, although its small scale introduces instability.
Open-VQA	[68]	Proposed to support open-ended answers.
VLUE	[69,70]	The first multi-task benchmark focusing on vision-language understanding, covering image-text retrieval, visual question answering, visual reasoning, and visual grounding, and includes a newly annotated private out-of-distribution (OOD) test set using images from MaRVL.

 doi:10.20944/preprints202602.0467.v2

6.2. Hallucination Evaluation Benchmarks

These benchmarks specifically target assessing hallucinations in Image-to-Text (I2T) and Text-to-Image (T2I) generation tasks.

A. I2T (Image-to-Text) Hallucination Benchmarks

Table 2. Overview of I2T (Image-to-Text) Hallucination Benchmarks.

Benchmark Name	Citation	Key Details & Data Sources
POPE	[71]	Discriminative task benchmark using MSCOCO [59]. Targets faithfulness hallucinations, specifically object hallucinations.
HallusionBench	[72]	Discriminative benchmark sourced from a website [72], targeting both faithfulness and factuality.
CHAIR	[73]	Generative task benchmark focusing on object hallucinations in image captioning, sourced from MSCOCO [59].
AMBER	[74,75]	Comprehensive, LLM-free multi-dimensional benchmark evaluating object existence, attributes, and relations using manually collected images.
MERLIM	[76]	Evaluates existence, relation, and counting hallucinations using edited and original images from MSCOCO [59].
HaELM	[77]	First benchmark to utilize LLMs for hallucination evaluation within MLLMs, sourced from MSCOCO [59].
R-Bench	[78]	Discriminative benchmark evaluating relationship hallucinations, using MSCOCO [59].
Hal-Eval	[79]	Comprehensive benchmark including both in-domain (MSCOCO [59]) and out-of-domain datasets to assess potential data leakage.
VHtest	[80]	Uses MSCOCO [59] and DALL-E-3 generated data to construct synthetic datasets.
LongHalQA	[81]	Discriminative benchmark using Visual Genome [82] and Object365 [83].
PhD	[84]	Discriminative benchmark using TDIUC [85] to evaluate faithfulness and factuality.
HallucinaGen	[86]	Generative benchmark using MSCOCO [59] and NIH Chest X-ray [87].
FactCheXcker	[88]	Pipeline detecting object and measurement hallucinations in radiology reports, leveraging the MIMIC-CXR dataset.
NOPE	[89]	Generative benchmark sourced from OpenImages [90].
CIEM	[91]	Discriminative benchmark leveraging LLMs for automated question generation, sourced from MSCOCO [59].
RAH-Bench	[92]	Discriminative benchmark leveraging LLMs for automated question generation, sourced from MSCOCO [59].
ROPE	[93]	Discriminative benchmark using MSCOCO [59] and ADE20K [94].
VisDiaHalBench	[95]	Discriminative benchmark sourced from GQA [96].
CC-Eval	[97]	Generative benchmark sourced from Visual Genome [82].
GAVIE	[98]	Generative benchmark sourced from Visual Genome [82].
MMHal-Bench	[99]	Generative benchmark sourced from OpenImages [90].
FGHE	[100]	Discriminative benchmark sourced from MSCOCO [59].
VHILT	[101]	Generative task benchmark sourced from a website.
Med-HallMark	[102]	Comprehensive medical benchmark sourced from Slake [103] and others.
AutoHallusion	[104]	Discriminative benchmark establishing automated pipelines, sourced from MSCOCO [59] and DALL-E-2 [105].

 doi:10.20944/preprints202602.0467.v2

B. T2I (Text-to-Image) Hallucination Benchmarks

Benchmark Name	Citation	Key Details & Data Sources
TIFA v1.0	[106]	Generative task benchmark sourced from MSCOCO [59].
T2I-FactualBench	[107]	Generative task benchmark evaluating factuality hallucinations, sourced from GPT.
T2I-CompBench	[108]	A comprehensive open-world benchmark for evaluating compositional T2I generation, sourced from MSCOCO [59], Template, and GPT.
WISE	[109]	Designed to evaluate factuality hallucinations through complex prompts across natural sciences, spatiotemporal reasoning, and cultural knowledge, sourced from LLM-Constructed data.
SR 2D	[110]	Generative task benchmark sourced from MSCOCO [59].
DrawBench	[111]	Generative task benchmark involving human evaluation, sourced from Human and DALL-E [105].
ABC-6K & CC-500	[112]	Generative task benchmark sourced from MSCOCO [59].
PaintSkills	[113]	Generative task benchmark sourced from Template.
HRS-Bench	[114]	Generative task benchmark sourced from GPT.
GenAI-Bench	[115]	Generative task benchmark sourced from Human input.
I-HallA v1.0	[116]	Generative task benchmark focusing on factuality hallucinations, sourced from Textbook data.
OpenCHAIR	[117]	Generative task benchmark using Stable Diffusion.
ODE	[118]	Comprehensive benchmark utilizing Stable Diffusion to construct synthetic datasets.

6.3. Domain-Specific and Focused Benchmarks

These benchmarks evaluate capabilities in specialized fields (e.g., medical, finance, robotics) or focused tasks (e.g., visual reasoning, long context).

A. Expert-Level and Reasoning Benchmarks

Benchmark Name	Citation	Key Details & Data Sources
MMMU	[119,119]	Massive Multi-discipline Multimodal Understanding and Reasoning benchmark, featuring 11.5K college-level questions across 6 disciplines, sourced from Textbooks and the Internet.
MMMU-Pro	[119]	A more robust version of the MMMU benchmark, introduced in September 2024.
MathVista	[120]	Evaluates mathematical reasoning in visual contexts, limited exclusively to the mathematical domain.
SCIENCEQA	[121]	Assesses multimodal reasoning via thought chains for science question answering.
GAIA	[122]	A benchmark testing fundamental abilities such as reasoning, multimodality handling, or tool use.
Visual CoT	[123]	Constructed with visual chain-of-thought prompts, requiring comprehensive recognition and understanding of image text content.
MMStar	[124]	A vision-indispensable benchmark covering a wide range of tasks and difficulty levels.
CLEVR	[125]	A diagnostic dataset for compositional language and elementary visual reasoning, relying on synthetic images.

 doi:10.20944/preprints202602.0467.v2

B. Medical and Healthcare Benchmarks

Benchmark Name	Citation	Key Details & Data Sources
CARES	[126]	A benchmark for evaluating the trustworthiness of medical vision-language models (Med-LVLMs) across five dimensions (trustfulness, fairness, safety, privacy, robustness).
OmniMedVQA	[127]	A large-scale comprehensive evaluation benchmark for medical LVLM, collected from 73 different medical datasets and 12 modalities, used as a source for CARES.
MIMIC-CXR	[128]	A large publicly available database of labeled chest radiographs. Used to construct CARES.
IU-Xray	[129]	A dataset including chest X-ray images and corresponding diagnostic reports, used to construct CARES.
Harvard-FairVLMed	[130]	Focuses on fairness in multimodal fundus images, used to construct CARES.
PMC-OA	[131,132]	Contains biomedical images extracted from open-access publications, used to construct CARES.
HAM10000	[133]	A dataset of dermatoscopic images of skin lesions for classification, used to construct CARES.
OL3I	[134]	A multimodal dataset for opportunistic CT prediction of ischemic heart disease (IHD), used to construct CARES.
VQA-RAD	[135]	An early-released VQA dataset, generally avoided in new medical benchmarks like CARES to prevent data leakage.
SLAKE	[103]	A semantically-labeled knowledge-enhanced dataset for medical VQA, generally avoided in new medical benchmarks like CARES to prevent data leakage.

C. Long Context and Document Understanding Benchmarks

Benchmark Name	Citation	Key Details & Data Sources
Document Haystack	[136]	A novel benchmark evaluating VLMs' ability to retrieve key multimodal information from long, visually complex documents (5 to 200 pages).
MM-NIAH (Multimodal Needle in a Haystack)	[137]	Benchmarking long-context capability, although its prompt length limitations make it less suitable for very long documents.
M-LongDoc	[138]	Benchmark for multimodal super-long document understanding, featuring documents spanning hundreds of pages.
Needle in a Haystack	[139]	Tests models' ability to retrieve information (the "needle") embedded within an extended context window (the "haystack").
LongBench	[140]	The first bilingual, multi-task framework for assessing long-form text understanding.
MileBench	[141]	Benchmarking MLLMs in long context.
DUDE	[142]	Document Understanding Dataset and Evaluation benchmark, attempting to tackle multi-page document comprehension.
Loong		Benchmark dealing with extended multi-document question answering.
SlideVQA	[143]	A dataset for document visual question answering on multiple images.
MMLongBench-Doc	[144]	Benchmarking long-context document understanding with visualizations.

 doi:10.20944/preprints202602.0467.v2

D. Specialized Datasets/Benchmarks (Perception, Retrieval, etc.)

Dataset/Benchmark Name	Citation	Key Details & Data Sources
MS COCO (Common Objects in Context)	[59]	Widely used dataset (330,000+ images) for object detection, segmentation, VQA, and captioning.
Visual Genome	[82]	Provides dense annotations (3.8M objects, 2.3M relationships) to bridge images and language, enabling reasoning tasks.
Flickr30K Entities	[145]	Extends Flickr30K with bounding box annotations and coreference chains for phrase grounding.
ImageBind (Meta AI)	[146]	Large-scale dataset linking images with six modalities (text, audio, depth, thermal, IMU) for unified multimodal embeddings.
LAION-5B	[147]	One of the largest open multimodal datasets (5.85 billion image-text pairs) for training foundation models.
Conceptual Captions (CC3M)	[148]	Contains ∼3.3 million image-caption pairs extracted and filtered from the web, designed for automatic image captioning.
VizWiz	[149]	Benchmark consisting of visual questions originating from blind people.
GQA	[96]	Developed to address the limitations of VQAv2, offering rich semantic and visual complexity for real-world visual reasoning.
VQAv2	[150]	A benchmark using pairs of similar images leading to different answers to compel models to prioritize visual data.
OCRBench	[151]	Focuses on Optical Character Recognition tasks.
TallyQA	(Contextual citation)	A Visual Question Answering dataset specifically designed to address counting questions in images.
RF100-VL (Roboflow100-VL)	[152]	Large-scale multimodal benchmark evaluating VLMs on out-of-distribution object detection, covering seven domains.
NLVR	[153]	A corpus for reasoning about natural language grounded in photographs (NLVR2 is the related task in VLUE [69]).
Massive Multitask Language Understanding (MMLU)		Crucial benchmark for evaluating general knowledge and reasoning across 57 diverse subjects.

6.4. Other Modalities (Video, Audio, 3D)

Dataset/Benchmark Name	Citation	Key Details & Data Sources
MVBench	[154]	A comprehensive multi-modal video understanding benchmark focusing on temporal perception.
Perception Test	[155]	A diagnostic benchmark for multimodal video models, covering Memory, Abstraction, Physics, and Semantics.
MSR VTT	[156]	A large video captioning dataset (10,000 video clips, 200,000 clip–sentence pairs) bridging video content and natural language.
VaTeX (Video And Text)	[157]	A multilingual video captioning dataset (English and Chinese) with 41,250 videos and 825,000 captions.
Dynamic-SUPERB	[158]	A benchmark assessing MLLMs' ability to follow instructions in the audio domain, focusing on human speech processing.
AIR-Bench	[159]	A comprehensive benchmark designed to evaluate MLLMs' ability to comprehend various audio signals (speech, natural sounds, music) and interact according to instructions.
MuChoMusic	[160]	The first benchmark for evaluating music understanding in audio MLLMs.
MCUB (Multimodal Commonality Understanding Benchmark)	[161]	Includes four modalities image, audio, video, and point cloud measuring the model's ability to identify commonalities among input entities.
M3DBench	[162]	Focuses on 3D instruction following.
ScanQA	[163]	3D question answering for spatial scene understanding.
AVQA	[164]	Designed for audio-visual question answering on general videos of real-life scenarios.
MMT-Bench	[165]	A comprehensive benchmark assessing MLLMs across massive multimodal tasks toward multitask AGI.

 doi:10.20944/preprints202602.0467.v2

6.5. Text-to-Audio Generation

Text-to-Audio (TTA) generation has emerged as a significant cross-modal research direction that extends the vision-language paradigm to the auditory domain. Unlike Text-to-Speech (TTS), which focuses exclusively on synthesizing spoken words from text, TTA encompasses the broader generation of general sound effects, music, and environmental audio from natural language descriptions [166]. This capability is central to achieving holistic multimodal understanding, as auditory signals provide complementary semantic information to visual and textual modalities. The rapid development of TTA systems has been catalyzed by advances in Large Language Models (LLMs), diffusion-based generative frameworks, and contrastive audio-language pretraining, mirroring the architectural trends observed in vision-language models discussed in preceding sections.

6.5.1. Architectural Taxonomy of TTA Models

Current TTA approaches can be broadly categorized into three architectural families.

Diffusion-based models constitute the dominant paradigm, employing probabilistic generative processes that learn to reverse a gradual noise-adding procedure to produce realistic audio samples [167]. Notable examples include AudioLDM [168], which operates in a compressed latent space using latent diffusion models to generate diverse audio types including sound effects, music, and speech from natural language prompts. Make-An-Audio [169] addresses data scarcity through pseudo prompt enhancement via expert distillation and dynamic reprogramming, while operating on spectrogram autoencoders that predict self-supervised representations. PicoAudio2 [170] introduces fine-grained temporal control by combining coarse text descriptions with temporal matrices through a Diffusion Transformer backbone, enabling precise timing of audio events. DualSpec [171] extends TTA to the spatial audio domain, generating 3D sound from text descriptions by leveraging dual spectrogram representations (Mel and STFT) to balance sound quality with directional accuracy.

Transformer-based models adapt transformer architectures to handle audio tokens conditioned on text inputs. VinTAGe [172] exemplifies this approach as a flow-based transformer that leverages a Visual-Text Encoder for cross-modal interaction, enabling holistic audio generation that is both temporally synchronized with video and semantically aligned with text.

Audio Language Models represent a third category, utilizing transformer-based language models to predict audio tokens autoregressively or non-autoregressively. This paradigm includes AudioLM [173], SPEAR-TTS [174], MusicLM [175], and MusicGen [176], which treat audio generation as a sequence modeling task analogous to text generation. VALL-E [177] further demonstrates this approach by framing TTS as a conditional language modeling task.

A related emerging direction involves **joint video-audio conditioning models** such as T2AV [178], which generates audio from text while maintaining temporal alignment with video content, and ReWaS [179], which uses energy patterns as a bridge between video and audio modalities.

6.5.2. TTA Datasets and Benchmarks

Several datasets have been developed to support TTA research, though the field still faces significant data scarcity compared to vision-language tasks. Table 9 summarizes the key datasets used in current TTA research.

 doi:10.20944/preprints202602.0467.v2

Table 9. Overview of Key Text-to-Audio Datasets.

Dataset	Scale	Domain	Key Characteristics
AudioCaps	∼46K clips	Natural sounds	Human-annotated captions for audio events
Clotho	∼5K clips	Environmental sounds	Crowdsourced captions; 5 captions per clip
WavCaps	∼403K clips	Multi-source audio	Machine-labeled via LLMs; sources include AudioSet [180] and FreeSound [181]
Audio-FLAN [182]	∼100M instances	Speech, music, sound	Large-scale instruction-tuning dataset; 80 tasks spanning understanding and generation
VGGSound	∼200K clips	Audio-visual	YouTube video clips; 309 sound classes
VinTAGe-Bench [172]	636 pairs	Video-text-audio	212 videos with on-screen/offscreen captions; 14 onscreen and 24 offscreen categories
ESC-50	2,000 clips	Environmental sounds	50 categories; 5-second clips from FreeSound

6.5.3. Comparative Analysis of TTA Models

Table 10 presents a comparative overview of representative TTA models across key evaluation dimensions. The CLAP (Contrastive Language-Audio Pretraining) score measures audio-text alignment, while the Fréchet Audio Distance (FAD) quantifies the distributional similarity between generated and reference audio, with lower values indicating higher quality.

Table 10. Comparative Analysis of Text-to-Audio Generation Models

Model	Dataset	Architecture	Approach	CLAP	FAD↓	Year
AudioGen [166]	AudioCaps	Transformer	Auto-reg.	0.72	2.45	2023
AudioLDM [168]	AudioCaps	Latent Diffusion	Diffusion	—	—	2023
Make-An-Audio [169]	Multi-source	Latent Diffusion + CLAP	Diffusion	—	—	2023
V2A Mapper	ESC-50 + AudioCaps	CLIP-based Mapping	Mapping	0.80	1.35	2023
DreamAudio [183]	Custom	Diffusion Model	Diffusion	0.84	0.46	2025
PicoAudio2 [170]	AudioCaps	Diffusion Transformer	Diffusion	—	—	2025
VinTAGe [172]	Multi-modal	Flow Transformer	Hybrid	**0.86**	0.72	2024

As shown in Table 10, diffusion-based architectures generally demonstrate superior generation quality (lower FAD), with DreamAudio [183] achieving the best FAD score of 0.46. Hybrid models such as VinTAGe [172] achieve the highest CLAP alignment score (0.86), suggesting that incorporating visual conditioning alongside text improves semantic alignment. The progression from pure transformer-based autoregressive models (AudioGen) to latent diffusion and hybrid architectures mirrors the broader trend observed in vision-language models, where increasingly sophisticated fusion mechanisms yield improved cross-modal alignment.

 doi:10.20944/preprints202602.0467.v2

6.5.4. Evaluation Metrics for TTA

TTA evaluation employs both objective and subjective metrics, many of which parallel those used in vision-language evaluation. The **CLAP Score** measures audio-text alignment analogous to CLIP-based image-text alignment. The **Fréchet Audio Distance (FAD)** quantifies distributional similarity between generated and reference audio features (typically extracted via VGGish), serving a role comparable to FID in image generation. **Kullback-Leibler (KL) Divergence** measures the logarithmic difference between probability distributions of audio features extracted via PANNs [184]. **Mean Opinion Score (MOS)** provides subjective evaluation across dimensions including overall quality (MOS-Q), faithfulness to text description (MOS-F), and temporal alignment (MOS-T). Additional task-specific metrics include the **Inception Score (IS)** for assessing generation quality and diversity, defined as:

$$\text{IS} = \exp(\mathbb{E}_{\mathbf{x}}[\text{KL}(p(y|\mathbf{x}) \parallel p(y))]) \tag{1}$$

and **Segment-F1** for evaluating temporal accuracy of audio event detection, which compares predicted audio event segments with ground-truth segments and computes F1 based on segment-level matches. The **Structural Similarity Index (SSIM)** is also employed for spectrogram-level comparison:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

6.5.5. Open Challenges and Connections to Vision-Language Research

Several open challenges in TTA research directly parallel those identified in vision-language modeling. First, **semantic alignment between text and audio** remains difficult, as text encoders often fail to capture precise temporal and acoustic cues from natural language descriptions, echoing the fine-grained cross-modal alignment challenges discussed in Section 8. Second, **data scarcity** persists as a fundamental limitation; general audio datasets lack the fine-grained temporal annotations necessary for training models with precise timing control, and there exists a significant distribution gap between simulated and real-world audio data. Third, **computational efficiency** for latent diffusion models and the difficulty of modeling long continuous waveforms impose constraints analogous to those faced by large-scale VLMs. Fourth, connecting video information to audio generation models across different representation spaces presents a cross-modal bridging challenge that mirrors the modality alignment problems observed in vision-language fusion (Section **??**). Finally, balancing sound quality with directional accuracy in spatial audio generation [171] exemplifies the broader trade-off between generation fidelity and semantic precision that pervades multimodal generative systems.

Future directions for TTA include developing unified Audio-LLMs with zero-shot generalization across understanding and generation tasks, expanding instruction-tuning datasets such as Audio-FLAN [182], integrating conversational capabilities for real-time dialogue, and improving temporal control for overlapping sound events. These directions are closely aligned with the broader research trajectories for multimodal foundation models outlined in Section 8, particularly regarding self-supervised pretraining, adaptive fusion mechanisms, and efficient cross-modal interaction.

7. Evolution of Multimodal Vision Models

The evolution of Multimodal Vision Models (VLM/MLLM) can be systematically categorized into three major eras, moving from early systems focused on task-specific feature engineering to modern, large-scale foundational models that leverage generalized pre-training and transformer architectures.

Early Models (2007–2015) [185–187]

This initial phase saw the introduction of foundational tasks for combining vision and language, primarily relying on convolutional neural networks (CNNs) for vision and recurrent neural networks (RNNs) or long short-term memory (LSTM) for language. This era predates the widespread adoption of large-scale, unified vision-language pre-training (VLP).

 doi:10.20944/preprints202602.0467.v2

Key Models and Architectures

1. **DeViSE (Deep Visual-Semantic Embedding Model) [187]**
   Architecture & Training: Introduced in 2013, DeViSE focused on learning a shared embedding space between visual and semantic modalities.
   Unique Contributions: This approach enabled zero-shot classification, allowing the model to detect unseen object classes by leveraging purely textual descriptions.
2. **VQA (Visual Question Answering) [185,188]**
   Unique Contributions: While VQA refers primarily to the task and dataset (introduced in 2015 by Antol et al.), it drove the development of early VLM architectures, defining the goal of answering questions based on visual input.
   Architecture & Training (Early Methods): The earliest deep learning approaches for VQA relied on CNN–RNN pairs. For vision feature extraction, models like VGGNet [189,189] and GoogLeNet [190,190] were commonly used, often employing transfer learning by leveraging knowledge learned on large vision datasets like ImageNet [191,191]. The fused output was then typically passed to a classifier or generator.
3. **NeuralTalk / Neural-Image-QA [186]**
   Architecture & Training: Neural-Image-QA (2015) was one of the first deep learning-based approaches for image question answering. It often used components like GoogLeNet for the image encoder and LSTM for the text encoder.
   Unique Contributions: These models marked the shift towards deep learning for image understanding and question answering tasks.

Transformer Revolution (2016–2020) [36,192–194]

This period is defined by the proliferation of the Transformer architecture [36], leading to the emergence of Vision-Language Pre-training (VLP) techniques that treat vision and language jointly, often pre-trained on large image-text pair datasets.

Key Models and Architectures

1. **VisualBERT [192,192]**
   Architecture: A single-stream model that processes both vision and language sequences jointly within a single encoder, usually based on BERT. The visual features were typically extracted using Faster R-CNN (FR-CNN) [195,195].
   Training & Contributions: Served as a highly performant and relatively simple baseline for vision and language tasks.
2. **ViLBERT (Pretraining Task-Agnostic Visiolinguistic Representations) [193,193]**
   Architecture: A dual-stream model architecture that encodes the visual and textual sequences separately before joining them in a Cross-Modal Transformer for fusion. It used BERT for the text encoder and FR-CNN for the visual encoder.
   Unique Contributions: ViLBERT was an early example of dual-stream models, proposed to account for the differences in abstraction levels between the two modalities. It aimed to pre-train task-agnostic representations for vision-and-language tasks.
3. **LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [194,194]**
   Architecture: A dual-stream framework based on Transformer encoders, featuring three components: a language encoder, an object relationship encoder, and a dedicated cross-modality encoder. It uses Cross-Modal Transformer technology.
   Training & Contributions: LXMERT utilized a comprehensive pre-training strategy involving five diverse tasks, including masked language modeling, masked object prediction (feature regression and label classification), cross-modality matching, and image question answering. This resulted in strong generalization capabilities across multiple visual reasoning tasks.

 doi:10.20944/preprints202602.0467.v2

Recent Large-Scale MLLMs (2021–2025) [47,54,196–198]

This era is characterized by the convergence of massive, pre-trained Large Language Models (LLMs) and advanced vision encoders, resulting in Multimodal Large Language Models (MLLMs). These models often utilize frozen LLMs as a backbone and focus on efficient alignment strategies.

Key Models and Architectures

1. **CLIP (Contrastive Language-Image Pre-training) [47,47]**
   Year: 2021.
   Architecture: Encoder–decoder model, using Vision Transformers (ViT) [199,200] or ResNets as the vision encoder.
   Training & Contributions: Trained using a contrastive learning objective on 400M image-text pairs [47], aligning vision and language encoders into a shared representation space. This training method enables remarkable transferability and strong zero-shot classification capabilities, surpassing classical single-modality models.
2. **Flamingo [54]**
   Year: 2022.
   Architecture: Decoder-only structure, designed to bridge powerful pretrained vision-only models (like NFNet) and language-only models (like Chinchilla-70B). It incorporates Cross-Attention (XAttn LLM) modules within the language model layers to fuse visual features.
   Training & Contributions: Flamingo was the first VLM to explore in-context few-shot learning at scale. It introduced architectural innovations to handle interleaved visual and textual data sequences. The model uses a resampling strategy to fix the number of visual tokens presented to the LLM.
3. **BLIP and BLIP-2 [198,201]**
   Year: BLIP (2022), BLIP-2 (2023).
   Architecture: BLIP used an Encoder–decoder architecture trained from scratch. BLIP-2 introduced the Q-Former (Querying Transformer). The Q-Former acts as a flexible, trainable adapter module between a frozen visual encoder (like EVA ViT-g) and a frozen LLM (like FlanT5).
   Training & Contributions: BLIP used bootstrapping for unified V–L understanding and generation. BLIP-2 revolutionized VLM training by decoupling the visual encoder and the LLM, enabling the leverage of powerful, frozen pre-trained LLMs to bootstrap language-image pre-training.
4. **LLaVA-1.5 [202]**
   Year: 2023.
   Architecture: Decoder-only model, typically using a frozen CLIP ViT-L/14 visual encoder and a Vicuna LLM backbone. It uses a simple MLP projection (a two-layer multilayer perceptron) to connect visual features to the textual embedding space.
   Training & Contributions: A primary example of utilizing visual instruction tuning (VIT) to enhance multimodal capabilities and promote conversation skills.
5. **GPT-4V (GPT-4 Vision) [196,203]**
   Year: 2023.
   Architecture & Training: Details are undisclosed.
   Unique Contributions: Recognized as a pioneering MLLM [196,197], GPT-4V demonstrates strong reasoning and understanding across visual and textual data. Qualitatively, it is noted for its precision and succinctness in responses compared to competitors.
6. **Gemini [197,204]**
   Year: 2023.
   Architecture & Training: A family of models utilizing a decoder-only architecture [197,197]. Details are undisclosed.
   Unique Contributions: Gemini excels in providing detailed, expansive answers, often incorporating relevant imagery and links, showcasing sophisticated multimodal capabilities [204].

 doi:10.20944/preprints202602.0467.v2

7. **CogVLM [205?]**
   Year: 2023.
   Architecture: Encoder–decoder model, utilizing a visual expert (CLIP ViT-L/14) and combining projection (MLP) with a modality experts fusion strategy.
   Training: It is visually instructed tuned. CogVLM is designed as a visual expert for pretrained language models.

8. Conclusion

Vision-based multimodal learning has undergone a profound transformation over the past decade, evolving from early visual–semantic embedding approaches to large-scale vision-language models (VLMs) and instruction-tuned multimodal large language models (MLLMs). In this review, we presented a comprehensive, task-oriented, and historically grounded analysis of this evolution, systematically categorizing multimodal vision models across major architectural paradigms, including dual-encoder contrastive frameworks, fusion-based transformer architectures, and unified generative models. By examining representative models and their applications across core vision-centric tasks such as image captioning, visual question answering, visual grounding, and cross-modal generation, we highlighted how advances in multimodal pretraining and transformer-based design have reshaped the capabilities of visual-language systems.

Our analysis demonstrates that large-scale multimodal pretraining has fundamentally shifted the field from task-specific multimodal fusion toward more generalizable, instruction-following, and zero-shot capable models. Compared to traditional early, late, and hybrid fusion strategies, modern VLMs and MLLMs benefit from stronger cross-modal alignment, emergent multimodal reasoning abilities, and improved transfer across downstream tasks. However, this progress has also introduced new challenges related to computational cost, data efficiency, interpretability, and robustness in real-world scenarios. Through a detailed examination of widely used multimodal datasets and benchmarks, we further revealed limitations in dataset diversity, annotation bias, and representativeness, which continue to constrain the generalization and evaluation of multimodal models. Furthermore, our analysis extends beyond the vision-language boundary to examine Text-to-Audio (TTA) generation as an increasingly important cross-modal research direction. However, TTA research faces its own distinct challenges, including limited large-scale annotated datasets, difficulty in capturing precise temporal cues from natural language, and the inherent trade-off between generation fidelity and semantic alignment.

Despite their impressive performance, current VLMs and MLLMs still struggle with fine-grained cross-modal reasoning, reliable grounding between visual entities and linguistic concepts, and adaptation to dynamic or low-resource environments. These challenges are particularly evident in embodied and robotic settings, where real-time constraints, multimodal synchronization, and domain-specific variability demand more efficient and adaptive fusion mechanisms. Addressing these limitations requires future research efforts that move beyond scale alone, emphasizing self-supervised and weakly supervised multimodal learning, dynamic and task-aware fusion strategies, and more principled approaches to multimodal reasoning and generalization.

Looking forward, promising research directions include self-supervised multimodal pretraining to reduce reliance on large annotated datasets, adaptive attention and routing mechanisms for efficient cross-modal interaction, and the integration of symbolic reasoning and world knowledge into multimodal foundation models. Furthermore, as multimodal systems become increasingly deployed in real-world applications, incorporating ethical, human-centered, and safety-aware design principles will be critical to ensuring responsible and trustworthy multimodal AI. By consolidating past developments, clarifying current limitations, and outlining future research trajectories, this survey aims to serve as a unified reference framework for advancing vision-language understanding and multimodal intelligence in the next generation of AI systems.

 doi:10.20944/preprints202602.0467.v2

References

1. J. S. Ryu, H. Kang, Y. Chu, and S. Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15(5):809–830, 2025.
2. W. Liu, G. Wu, H. Wang, and F. Ren. Cross-modal data fusion via vision-language model for crop disease recognition. *Sensors*, 25(13):4096, 2025.
3. S. Singh. Everything you need to know about vision language models (vlms), July 8 2025. Accessed: 2025-11-01.
4. D. Garcia. 5 ways vision-language models are transforming ai applications, September 22 2025. Accessed: 2025-11-01.
5. Wikipedia contributors. Vision-language-action model, October 26 2025. Accessed: 2025-11-01.
6. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
7. Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.
8. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
9. OpenAI et al. Gpt-4 technical report, 2024.
10. Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019.
11. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.
12. Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models, 2023.
13. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
14. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.
15. Ruiyang Qin and Authors Institutes. Tiny-align: Bridging automatic speech recognition and large language model on edge, 2024. Accessed: 2025-11-01.
16. Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, Rongtao Xu, and Shibiao Xu. Multimodal fusion and vision–language models: A survey for robot vision. *Information Fusion*, 126:103652, 2026.
17. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.
18. Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey, 2024.
19. Chen Zhong, Shuo Zeng, and Hao Zhu. Adaptive multimodal fusion with cross-attention for robust scene segmentation and urban economic analysis. *Applied Sciences*, 15(1):438, 2025.
20. Gunther Kress. *Multimodality: A social semiotic approach to contemporary communication*. Routledge, London, 2010. Definition of 'mode' in source [4], cited in [14].
21. Mohammad Saleh and Azadeh Tabatabaei. Building trustworthy multimodal ai: A review of fairness, transparency, and ethics in vision-language tasks. *arXiv preprint arXiv:2501.02189*, 2025. Source [15] provides technical context for multimodality in AI.
22. Theo Van Leeuwen. *Introducing social semiotics*. Psychology Press, 2005. Definition of Multimodal Discourse in source [1].
23. Wikipedia. Multimodal learning. A type of deep learning that integrates and processes multiple types of data, such as text, audio, images, or video. (Source [7]).
24. Milvus. How is multimodal ai used in robotics? 2025. Discusses multimodal AI integration in robotics (Source [13]).
25. G. Singh. A review of multimodal vision–language models: Foundations, applications, and future directions. *Preprints*, 2025.
26. G. Singh, T. Banerjee, and N. Ghosh. Tracing the evolution of artificial intelligence: A review of tools, frameworks, and technologies (1950–2025). *Preprints*, 2025.

 doi:10.20944/preprints202602.0467.v2

27. G. Singh. Ai-assisted storytelling: Enhancing narrative creation in digital media. *International Journal of Engineering Development and Research*, 14(1):882–894, 2026.
28. G. Singh, A. Naaz, A. Syed, and V. Akhila. Ai-assisted storytelling: Enhancing narrative creation in digital media. *Preprints*, 2026.
29. GeeksforGeeks. Early fusion vs. late fusion in multimodal data processing. 2025. Last Updated: 23 Jul, 2025.
30. Ruhina Karani and Sharmishta Desai. Review on multimodal fusion techniques for human emotion recognition. *The Science and Information (SAI) Organization*, 13(10), 2022.
31. Milvus. What fusion strategies work best for combining results from different modalities? 2025. AI Reference.
32. Shiv Shankar, Laure Thompson, and Madalina Fiterau. Progressive fusion for multimodal integration. In *arXiv:2209.00302v2 [cs.LG]*, 2022.
33. Maxwell Mbabilla Aladago and AJ Piergiovanni. Compound tokens: Channel fusion for vision-language representation learning. In *OpenReview: ICLR 2023 Tiny Papers Track*, 2023.
34. Wikipedia contributors. Multimodal learning. *Wikipedia, The Free Encyclopedia*, 2024. Retrieved on YYYY-MM-DD.
35. Jiuhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jianfeng Gao, Tianyi Zhou, and Bin Xiao. Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion. *CVF Open Access*, 2024.
36. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017.
37. Weng C Zhao, Kun Zhou, Jun Li, Tianyi Tang, Xi Wang, Yuxiao Hou, Ying Min, Beichen Zhang, Junjie Zhang, Zhipeng Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
38. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
39. Salman Soltan, Sonal Ananthakrishnan, Jon FitzGerald, Rohit Gupta, Wael Hamza, Hitesh Khan, Carlos Peris, Scott Rawls, Andrew Rosenbaum, Anna Rumshisky, et al. AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*, 2022.
40. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
41. Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
42. Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Rémi Castagné, Alexandra S Luccioni, François Yvon, Matthieu Gallé, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*, 2022.
43. Hugo Touvron, Thibaut Lavril, G Izacard, Xavier Martinet, Marie-Anne Lachaux, Thomas Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
44. Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q Tran, David R So, Siamak Shakeri, Xavier Garcia, Hong S Zheng, Jinfeng Rao, Aakanksha Chowdhery, et al. Transcending scaling laws with 0.1% extra compute. *arXiv preprint arXiv:2210.11399*, 2022.
45. Xiaozhe Ren, Peng Zhou, Xinzhou Meng, Xinyu Huang, Yue Wang, Wenbin Wang, Peng Li, Xinchao Zhang, Alexey Podolskiy, Gleb Arshinov, et al. Pangu-Σ: Towards trillion parameter language model with sparse heterogeneous computing. *arXiv preprint arXiv:2303.10845*, 2023.
46. Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
47. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
48. Xiao Liu, Kaipeng Ji, Yuxian Fu, Wayne Tam, Zhilin Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022.

 doi:10.20944/preprints202602.0467.v2

49. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Conference on Neural Information Processing Systems*, 2019.
50. Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv preprint arXiv:2205.01917*, 2022.
51. Peng Xu, Xiatian Zhu, David A Clifton, et al. Multimodal Learning with Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.
52. Gang Chen, Feng Liu, Zhiliang Meng, and Sheng Liang. Revisiting parameter-efficient tuning: Are we really there yet? *arXiv preprint arXiv:2202.07962*, 2022.
53. Xiao Liu, Yuxian Zheng, Zhilin Du, Ming Ding, Yujia Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. In *arXiv preprint arXiv:2103.10385*, 2021.
54. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
55. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.
56. Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
57. Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, 2306.13394, 2023.
58. Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, 2306.13394, 2024.
59. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.
60. Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
61. Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
62. Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.16911*, 2023.
63. Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Wanli Ouyang, and Jing Shao. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *NeurIPS Datasets and Benchmarks*, 2023.
64. Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2404.18029*, 2024.
65. Zhelun Shi, Zhipin Wang, Hongxing Fan, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. Chef: A comprehensive evaluation framework for standardized assessment of multimodal large language models. *arXiv preprint arXiv:2310.11585*, 2023.
66. Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *arXiv preprint arXiv:2401.12781*, 2024.
67. Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2310.15053*, 2023.
68. Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2310.00794*, 2023.
69. Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. In *ICML*, volume 162, 2022.
70. Fangyu Liu, Enrico Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. *EMNLP*, pages 10467–10485, 2021.

 doi:10.20944/preprints202602.0467.v2

71. Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *EMNLP*, 2023.
72. Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385, 2023.
73. Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pages 4035–4045, 2018.
74. Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2311.07397, 2023.
75. Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2311.07397, 2024.
76. Alexander Villa, Jesús Léon, Alfonso Soto, and Bernard Ghanem. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. *CVPR*, pages 492–502, 2025.
77. Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, 2308.15126, 2023.
78. Min-Ku Wu, Jian Ji, Olivia Huang, Jinsong Li, Yu Wu, Xiaojun Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. *ICML*, 2024.
79. Conghui Jiang, Wenqian Ye, Min Dong, Haiyun Jia, Guohai Xu, Ming Yan, Ji Zhang, and Sheng Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *ACM MM*, 2024.
80. Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. *Findings of the ACL*, pages 9614–9631, 2024.
81. Hao Qiu, Jing Huang, Peng Gao, Qi Qi, Xiangliang Zhang, Ling Shao, and Sheng Lu. Longhalqa: Long-context hallucination evaluation for multimodal large language models. *CoRR*, 2410.09962, 2024.
82. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Saqib Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2016.
83. Siyuan Shao, Zhifeng Li, Tianliang Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019.
84. Jiali Liu, Yuzheng Fu, Ruifei Xie, Rui Xie, Xiaowei Sun, Fan Lian, Zhaoli Kang, and Xiaofeng Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *CVPR*, pages 19857–19866, 2025.
85. Krishna Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, pages 1965–1973, 2017.
86. Aashay Seth, Dinesh Manocha, and Chetan Agarwal. Hallucinogen: A benchmark for evaluating object hallucination in large visual-language models. *CoRR*, 2412.20622, 2024.
87. Xiaosong Wang, Yuxing Peng, Le Lu, Zhiyong Lu, Mahdi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CVPR*, pages 2097–2106, 2017.
88. Xingjian Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. *ACL*, 2024.
89. Holy Lovenia, Wenfei Dai, Samuel Cahyawijaya, Zhisheng Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ALVR Workshop*, pages 37–58, 2024.
90. Alina Kuznetsova, Hassan Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020.
91. Honghao Hu, Jiannan Zhang, Mingwei Zhao, and Zhiwei Sun. Ciem: Contrastive instruction evaluation method for better instruction tuning. *NeurIPS Workshop*, 2023.
92. Zhiyuan Chen, Yuxin Zhu, Yang Zhan, Zhilin Li, Chenlin Zhao, Jiaqi Wang, and Min Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023.

 doi:10.20944/preprints202602.0467.v2

93. Xinyuan Chen, Zongyao Ma, Xinyu Zhang, Shuyuan Xu, Shijia Qian, Jizhao Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *NeurIPS*, 37:44393–44418, 2024.
94. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *CVPR*, pages 633–641, 2017.
95. Qi Cao, Jianjun Cheng, Xiaodan Liang, and Liang Lin. Visdiahalbench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In *ACL*, pages 12161–12176, 2024.
96. Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019.
97. Boyi Zhai, Sheng Yang, Chenyu Xu, Shu Shen, Kurt Keutzer, Cong Li, and Ming Li. Halle-control: controlling object hallucination in large multimodal models. *CoRR*, 2310.01779, 2023.
98. Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *ICLR*, 2023.
99. Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *Findings of the ACL*, pages 13088–13110, 2024.
100. Lin Wang, Jie He, Sixing Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *MMM*, 2023.
101. Anita Rani, Vaibhav Rawte, Hritik Sharma, Nitish Anand, Koustuv Rajbangshi, Amit Sheth, and Abhijeet Das. Visual hallucination: Definition, quantification, and prescriptive remediations. *CoRR*, 2403.17306, 2024.
102. Jin Chen, Di Yang, Tianyi Wu, Ye Jiang, Xiaoyan Hou, Mengxue Li, Shuming Wang, Dong Xiao, Kai Li, and Li Zhang. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024.
103. Bingyao Liu, Li-Ming Zhan, Lu Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *ISBI*, pages 1650–1654, 2021.
104. Xiyang Wu, Tianrui Guan, Dahu Li, Shu Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Boyd-Graber, et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *Findings of the EMNLP*, pages 8395–8419, 2024.
105. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ICML*, pages 8821–8831, 2021.
106. Yongjing Hu, Bo Liu, Jungo Kasai, Yizhi Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, pages 20349–20360, 2023.
107. Zhen Huang, Wentao He, Qinghao Long, Yali Wang, Hongyang Li, Ziyu Yu, Fu Shu, Lillian Chan, Hanyuan Jiang, Li Gan, et al. T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. *CoRR*, 2412.04300, 2024.
108. Kuan-Chieh Huang, Kyle Sun, Enze Xie, Zhili Li, and Xiao Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, volume 36, pages 78723–78747, 2023.
109. Yitong Niu, Mengfan Ning, Ming Zheng, Bin Lin, Peng Jin, Jian Liao, Kang Ning, Bin Zhu, and Lu Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *CoRR*, 2503.07265, 2025.
110. Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhor Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *CoRR*, 2212.10015, 2022.
111. Chitwan Saharia, William Chan, Saurabh Saxena, Liyuan Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Rafael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022.
112. Weili Feng, Xin He, Tung-Jui Fu, Varun Jampani, Adithya Akula, Pavan Narayana, Sudipto Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ICLR*, 2023.
113. Biao Li, Ziyang Lin, Dilip Pathak, Jifei Li, Yu Fei, Kun Wu, Xiao Xia, Peng Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. *CVPR*, pages 5290–5301, 2024.
114. Elmahdi M. Bakr, Peng Sun, Xingqian Shen, Faisal F. Khan, L. E. Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *ICCV*, pages 20041–20053, 2023.
115. Jaemin Cho, Aman Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, pages 3043–3054, 2023.

 doi:10.20944/preprints202602.0467.v2

116. Yoojin Lim, Hyewon Choi, and Hwanjong Shim. Evaluating image hallucination in text-to-image generation with question-answering. *AAAI*, 39(25):26290–26298, 2025.
117. Adam Ben-Kish, Moran Yanuka, Michael Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. *EMNLP*, pages 22680–22698, 2024.
118. Yichen Tu, Renguang Hu, and Jingkuan Sang. Ode: Open-set evaluation of hallucinations in multimodal large language models. *CVPR*, pages 19836–19845, 2025.
119. Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *CVPR*, pages 9556–9567, 2024.
120. Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
121. Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Neurips*, 35:2507–2521, 2022.
122. Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.
123. Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2407.10657*, 2024.
124. Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *NeurIPS*, 37:27056–27087, 2024.
125. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997, 2016.
126. Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2410.19830*, 2024.
127. Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. *arXiv preprint arXiv:2402.09181*, 2024.
128. Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
129. Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
130. Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuai-hang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. *arXiv preprint arXiv:2403.19949*, 2024.
131. Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *MICCAI*, pages 525–536, 2023.
132. Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
133. Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. volume 5, pages 1–9, 2018.
134. Juan M Zambrano Chaves, Andrew L Wentland, Arjun D Desai, Imon Banerjee, Gurkiran Kaur, Ramon Correa, Robert D Boutin, David J Maron, Fatima Rodriguez, Alexander T Sandhu, et al. Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach. *Scientific Reports*, 13(1):21034, 2023.
135. Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
136. Goeric Huybrechts, Srikanth Ronanki, Sai Muralidhar Jayanthi, Jack Fitzgerald, and Srinivasan Veeravanallur. Document haystack: A long context multimodal image/document understanding vision llm benchmark. *Amazon Science*, 2024.

 doi:10.20944/preprints202602.0467.v2

137. Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Needle in a multimodal haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.07230*, 2024.
138. Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176*, 2024.
139. Greg Kamradt. Needle in a haystack-pressure testing llms. *Github Repository*, page 28, 2023.
140. Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.
141. Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.
142. Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). *ICCV*, pages 19528–19540, 2023.
143. Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. *AAAI*, 37:13636–13645, 2023.
144. Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024.
145. Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.
146. Rohit Girdhar, Alaessia El-Nouby, Karttikeya Mangalam, Piyush Singh, Xinlei Han, Angjoo Kopuluru, Armand Joulin, and Ishan Taveres. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023.
147. Christoph Schuhmann, Romain Beaumont, Richard Vencovsky, Robert Gordon, Melissa Wightman, A. Jitsev, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.16084*, 2022.
148. Piyush Sharma, Nan Ding, S. Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018.
149. Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218*, 2018.
150. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.
151. Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2024.
152. Roboflow. Rf100-vl: A benchmark for few-shot generalization in vision-language models. *Research paper (Contextual Citation)*, 2025.
153. Alane Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.
154. Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2312.00985*, 2024.
155. Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, et al. Perception test: A diagnostic benchmark for multimodal video models. *arXiv preprint arXiv:2303.13380*, 2023.
156. Junnan Xu, Tao Mei, Ting Yao, and Yongdong Zhang. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 2601–2610, 2016.
157. Xin Wang, Wenlu Wu, Jianfeng Li, Xiaokang Wang, Lei Liu, Zili Wu, Junxing Wang, and Jian Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *ICCV*, pages 5710–5719, 2019.
158. Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chun-Yi Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *arXiv preprint arXiv:2404.09068*, 2024.

 doi:10.20944/preprints202602.0467.v2

159. Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2405.02384*, 2024.
160. Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2405.01358*, 2024.
161. Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. Model composition for multimodal large language models. *arXiv preprint arXiv:2404.03212*, 2024.
162. Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. M3dbench: Let's instruct large models with multi-modal 3d prompts. *arXiv preprint arXiv:2312.01255*, 2023.
163. Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *arXiv preprint arXiv:2208.06456*, 2022.
164. Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *ACM MM*, pages 3480–3491, 2022.
165. Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2407.13532*, 2024.
166. Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2023.
167. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
168. Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.
169. Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023.
170. Zeyu Zheng, Zhaojun Xie, Xiaobin Xu, Wenjie Wu, Chaofan Zhang, and Meng Wu. Picoaudio2: Temporal controllable text-to-audio generation with natural language description. *arXiv preprint arXiv:2509.00683*, 2025.
171. Ling Zhao, Shuai Chen, Li Feng, Jie Zhang, Xiao-Lei Zhang, Chaofan Zhang, and Xiaolin Li. Dualspec: Text-to-spatial-audio generation via dual-spectrogram guided diffusion model. *arXiv preprint arXiv:2502.18952*, 2025.
172. Surya Shankar Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. *arXiv preprint arXiv:2412.10768*, 2024.
173. Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2023.
174. Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718, 2023.
175. Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.
176. Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2024.
177. Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.
178. Shentong Mo, Jing Shi, and Yapeng Tian. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*, 2024.
179. Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Joonhyuk Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024.

 doi:10.20944/preprints202602.0467.v2

180. Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.
181. Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 411–412, 2013.
182. Liumeng Xue, Ziya Zhou, Jiahui Pan, Zixuan Li, Shuai Fan, Yinghao Ma, Sitong Cheng, Dongchao Yang, Haohan Guo, Yijin Xiao, Xinhu Wang, Zhuo Shen, Chaofan Zhu, Xinchao Zhang, Ting Liu, Ruibin Yuan, Zhaoxiang Tian, Haohe Liu, Emmanouil Benetos, Ge Zhang, Yike Guo, and Wei Xue. Audio-flan: A preliminary release. *arXiv preprint arXiv:2502.16584*, 2025.
183. Yi Yuan, Xubo Liu, Haohe Liu, Xinyue Kang, Zehua Chen, Yuping Wang, Mark D Plumbley, and Wenwu Wang. Dreamaudio: Customized text-to-audio generation with diffusion models. *arXiv preprint arXiv:2509.06027*, 2025.
184. Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.
185. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
186. Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.
187. Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomáš Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, pages 2121–2129, 2013.
188. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
189. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
190. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
191. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
192. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
193. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, pages 13–23, 2019.
194. Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
195. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.
196. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
197. Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1, 2023.
198. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
199. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

 doi:10.20944/preprints202602.0467.v2

200. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
201. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900, 2022.
202. Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
203. Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.
204. Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases. *arXiv preprint arXiv:2312.15011*, 2023.
205. Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, and et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.