Article

Not peer-reviewed version

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

Gurpreet Singh *

Posted Date: 6 February 2026

doi: 10.20944/preprints202602.0467.v1

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

Article

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

Gurpreet Singh

Graduated - Endicott College of International Studies, Woosong University, Republic of Korea; gurpreetsinghmcse@gmail.com or gurpreetsingh@wsu.ac.kr

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 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 [?]. In education, these models enhance multimodal learning environments by generating visual explanations and captions that aid accessibility and support learners with visual impairments [?]. 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 [?]. 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 [3], ALIGN [4], and BLIP-2 [5] possess remarkable cross-modal alignment and generalization capabilities, enabling robust performance in zero-shot image classification and retrieval tasks [3,4]. These models further demonstrate strong potential in instruction-following scenarios, where natural language prompts are seamlessly mapped to visual tasks through multimodal reasoning [5,6]. Similarly, in visual question answering (VQA), transformer-based architectures such as LXMERT [7] and ViLBERT [8] have achieved state-of-the-art results in understanding fine-grained relationships between vision and language [7,8]. 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 [9,10].

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 [11,12]. 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 [12,13]. 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 [14]. 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 [15,16].

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 [17,18]. 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" [19]. 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" [17]. 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 diverse data streams, such as text, audio, images, or video, to achieve a more holistic understanding of complex inputs [18,20]. 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 [18,20]. 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 [21]. Recently we also worked on similar interests like [22] [23]. We have also worked on using Artificial Intelligence and Multi-Modality in storytelling too [24] [25]

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 [26]. The two fundamental strategies are Early Fusion and Late Fusion, differentiated by the stage at which data integration occurs [27,28]. Early Fusion, also referred to as feature-level fusion [26,27], combines raw data or low-level features from different modalities into a single feature set before inputting them into a single machine learning model [26,28]. This approach captures intricate relationships between modalities and yields rich feature representations [26,28,29]. However, early fusion can result in high-dimensional feature spaces, inflexibility, and significant challenges when dealing with heterogeneous or asynchronous data [26,28,29]. 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 [26–28]. This method offers modularity and avoids the high dimensionality associated with early fusion [26,28], but it risks missing critical cross-modal interactions that are crucial for complex tasks, as the input modalities are processed separately [26,29]. 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 [30]. Hybrid Fusion (sometimes called intermediate fusion) combines aspects of both early and late strategies to mitigate their trade-offs [27,28]. Hybrid methods often apply feature-level fusion to modalities that are synchronous in time while using decision-level fusion for the remaining asynchronous modalities [27]. For example, intermediate fusion often utilizes cross-attention layers to dynamically integrate modality-specific representations [31]. 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 [29]. 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 [30]. 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 [32].

2.3. Architecture Overview

3. General Transformer Architectures (LLMs)

These models primarily use variants of the original Transformer architecture [33].

3.1. Encoder–Decoder Architecture

This structure processes inputs through an encoder and feeds the representation to a decoder for output generation [34]. The encoder uses self-attention across the full input sequence, while the decoder uses cross-attention and generates tokens sequentially [34].

T5 [35] is a prime example, using a unified text-to-text paradigm for all NLP tasks [35]. Another example is AlexaTM [36]. Some research suggests encoder–decoder models may be advantageous, though scaling decoder-only models can close this performance gap [34].

doi:10.20944/preprints202602.0467.v1

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 [34]. It employs causal attention, where the prediction of a token depends only on previous time steps [34]. Examples include PaLM [37], GPT-3 [38], BLOOM [39], and LLaMA [40].

3.3. *Prefix Decoder Architecture (Non-Causal Decoder)*

In this variant, the attention calculation is bidirectional and not strictly dependent only on prior context [34]. An example is U-PaLM, which is described as a non-causal decoder model [41].

3.4. *Mixture-of-Experts (MoE)*

This is an efficient sparse variation of the Transformer [34]. It incorporates parallel independent experts (typically feed-forward layers) and a router that directs tokens to specific experts [34].

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 [34,42].

4. Multimodal Architectures (VLMs)

Vision-Language Models (VLMs), or Multimodal LLMs (MLLMs), combine vision encoders and LLMs to handle both image and text inputs [43]. They are primarily differentiated by their mechanism for multimodal fusion [43].

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 [43].

CLIP (Contrastive Language–Image Pretraining) is a dual encoder that uses a contrastive objective [43,44], where fusion is achieved via the dot product between global image and text embeddings [44]:

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

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 [45]. Examples include UniTER [45].

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 [46].

CoCa (Contrastive Captioner) employs a minimalist encoder–decoder design pretrained jointly with a contrastive loss (as in CLIP) and a generative captioning loss [47]. The decoder layers omit

doi:10.20944/preprints202602.0467.v1

cross-attention in the first half to maintain unimodal text representation before later layers cross-attend to the image encoder for multimodal representations [47].

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 $Z^{(B)}$ [48].

4.2.1. Early Summation

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

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

[48]

4.2.2. Early Concatenation

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

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

[48]

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:

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

[48,49]

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:

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

[48]

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 [50].

4.2.6. Cross-Attention to Concatenation

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

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

[48]

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 [51]. 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 [51]. The cross-attention layers are trained from scratch and attend to the Perceiver Resampler outputs [51].

5.2. LLaVA Architecture

LLaVA models (e.g., LLaVA-v1.5 [52]) 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 [52].

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	[53]	A novel multi-modality benchmark utilizing a meticulously curated dataset and the CircularEval strategy with ChatGPT for robust evaluation.
MME	[54], [55], [56]	Measures both perception and cognition abilities across subtasks. It uses the MSCOCO dataset.
MM-Vet	[57]	Devised to study integrated vision-language capabilities, offering insights beyond overall model rankings. It covers 200 items in total.
SEED-Bench	[58]	A comprehensive benchmark featuring multiple-choice questions covering various evaluation dimensions for both image and video modalities.
SEED-Bench-2	[59]	Categorized MLLMs' capabilities into hierarchical levels from L0 to L4.
SEED-Bench-H	[59]	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	[59]	Constructed to examine a variety of MLLM capabilities.
LAMM	[60]	Provides a comprehensive assessment of MLLMs' capabilities, particularly in understanding visual prompting instructions.
MDVP-Bench	[61]	Created to provide a comprehensive assessment of MLLMs' capabilities, particularly in understanding visual prompting instructions.
ChEF	[62]	Constructed as a standardized and holistic evaluation framework.
UniBench	[63]	Constructed as a standardized and holistic evaluation framework.
TouchStone	[64]	Proposed to support open-ended answers, although its small scale introduces instability.
Open-VQA	[65]	Proposed to support open-ended answers.
VLUE	[66], [67]	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.

6.2. II. 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

doi:10.20944/preprints202602.0467.v1

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

Benchmark Name	Citation	Key Details & Data Sources
POPE	[68]	Discriminative task benchmark using MSCOCO [56]. Targets faithfulness hallucinations, specifically object hallucinations.
HallusionBench	[69]	Discriminative benchmark sourced from a website [69], targeting both faithfulness and factuality.
CHAIR	[70]	Generative task benchmark focusing on object hallucinations in image captioning, sourced from MSCOCO [56].
AMBER	[71,72]	Comprehensive, LLM-free multi-dimensional benchmark evaluating object existence, attributes, and relations using manually collected images.
MERLIM	[73]	Evaluates existence, relation, and counting hallucinations using edited and original images from MSCOCO [56].
HaELM	[74]	First benchmark to utilize LLMs for hallucination evaluation within MLLMs, sourced from MSCOCO [56].
R-Bench	[75]	Discriminative benchmark evaluating relationship hallucinations, using MSCOCO [56].
Hal-Eval	[76]	Comprehensive benchmark including both in-domain (MSCOCO [56]) and out-of-domain datasets to assess potential data leakage.
VHtest	[77]	Uses MSCOCO [56] and DALL-E-3 generated data to construct synthetic datasets.
LongHalQA	[78]	Discriminative benchmark using Visual Genome [79] and Object365 [80].
PhD	[81]	Discriminative benchmark using TDIUC [82] to evaluate faithfulness and factuality.
HallucinaGen	[83]	Generative benchmark using MSCOCO [56] and NIH Chest X-ray [84].
FactCheXcker	[85]	Pipeline detecting object and measurement hallucinations in radiology reports, leveraging the MIMIC-CXR dataset.
NOPE	[86]	Generative benchmark sourced from OpenImages [87].
CIEM	[88]	Discriminative benchmark leveraging LLMs for automated question generation, sourced from MSCOCO [56].
RAH-Bench	[89]	Discriminative benchmark leveraging LLMs for automated question generation, sourced from MSCOCO [56].
ROPE	[90]	Discriminative benchmark using MSCOCO [56] and ADE20K [91].
VisDiaHalBench	[92]	Discriminative benchmark sourced from GQA [93].
CC-Eval	[94]	Generative benchmark sourced from Visual Genome [79].
GAVIE	[95]	Generative benchmark sourced from Visual Genome [79].
MMHal-Bench	[96]	Generative benchmark sourced from OpenImages [87].
FGHE	[97]	Discriminative benchmark sourced from MSCOCO [56].
VHILT	[98]	Generative task benchmark sourced from a website.
Med-HallMark	[99]	Comprehensive medical benchmark sourced from Slake [100] and others.
AutoHallusion	[101]	Discriminative benchmark establishing automated pipelines, sourced from MSCOCO [56] and DALL-E-2 [102].

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

Benchmark Name	Citation	Key Details & Data Sources
TIFA v1.0	[103]	Generative task benchmark sourced from MSCOCO [56].
T2I-FactualBench	[104]	Generative task benchmark evaluating factuality hallucinations, sourced from GPT.

T2I-CompBench	[105]	A comprehensive open-world benchmark for evaluating compositional T2I generation, sourced from MSCOCO [56], Template, and GPT.
WISE	[106]	Designed to evaluate factuality hallucinations through complex prompts across natural sciences, spatiotemporal reasoning, and cultural knowledge, sourced from LLM-Constructed data.
SR 2D	[107]	Generative task benchmark sourced from MSCOCO [56].
DrawBench	[108]	Generative task benchmark involving human evaluation, sourced from Human and DALL-E [102].
ABC-6K & CC-500	[109]	Generative task benchmark sourced from MSCOCO [56].
PaintSkills	[110]	Generative task benchmark sourced from Template.
HRS-Bench	[111]	Generative task benchmark sourced from GPT.
GenAI-Bench	[112]	Generative task benchmark sourced from Human input.
I-HallA v1.0	[113]	Generative task benchmark focusing on factuality hallucinations, sourced from Textbook data.
OpenCHAIR	[114]	Generative task benchmark using Stable Diffusion.
ODE	[115]	Comprehensive benchmark utilizing Stable Diffusion to construct synthetic datasets.

III. 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	[116], [116]	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	[116]	A more robust version of the MMMU benchmark, introduced in September 2024.
MathVista	[117]	Evaluates mathematical reasoning in visual contexts, limited exclusively to the mathematical domain.
SCIENCEQA	[118]	Assesses multimodal reasoning via thought chains for science question answering.
GAIA	[119]	A benchmark testing fundamental abilities such as reasoning, multimodality handling, or tool use.
Visual CoT	[120]	Constructed with visual chain-of-thought prompts, requiring comprehensive recognition and understanding of image text content.
MMStar	[121]	A vision-indispensable benchmark covering a wide range of tasks and difficulty levels.
CLEVR	[122]	A diagnostic dataset for compositional language and elementary visual reasoning, relying on synthetic images.

6.3. B. Medical and Healthcare Benchmarks

Benchmark Name	**Citation**	**Key Details & Data Sources**
CARES	[123]	A benchmark for evaluating the trustworthiness of medical vision-language models (Med-LVLMs) across five dimensions (trustfulness, fairness, safety, privacy, robustness).
OmniMedVQA	[124]	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	[125]	A large publicly available database of labeled chest radiographs. Used to construct CARES.

IU-Xray	[126]	A dataset including chest X-ray images and corresponding diagnostic reports, used to construct CARES.
Harvard-FairVLMed	[127]	Focuses on fairness in multimodal fundus images, used to construct CARES.
PMC-OA	[128], [129]	Contains biomedical images extracted from open-access publications, used to construct CARES.
HAM10000	[130]	A dataset of dermatoscopic images of skin lesions for classification, used to construct CARES.
OL3I	[131]	A multimodal dataset for opportunistic CT prediction of ischemic heart disease (IHD), used to construct CARES.
VQA-RAD	[132]	An early-released VQA dataset, generally avoided in new medical benchmarks like CARES to prevent data leakage.
SLAKE	[100]	A semantically-labeled knowledge-enhanced dataset for medical VQA, generally avoided in new medical benchmarks like CARES to prevent data leakage.

6.4. C. Long Context and Document Understanding Benchmarks

Benchmark Name	Citation	Key Details & Data Sources
Document Haystack	[133]	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)	[134]	Benchmarking long-context capability, although its prompt length limitations make it less suitable for very long documents.
M-LongDoc	[135]	Benchmark for multimodal super-long document understanding, featuring documents spanning hundreds of pages.
Needle in a Haystack	[136]	Tests models' ability to retrieve information (the "needle") embedded within an extended context window (the "haystack").
LongBench	[137]	The first bilingual, multi-task framework for assessing long-form text understanding.
MileBench	[138]	Benchmarking MLLMs in long context.
DUDE	[139]	Document Understanding Dataset and Evaluation benchmark, attempting to tackle multi-page document comprehension.
Loong		Benchmark dealing with extended multi-document question answering.
SlideVQA	[140]	A dataset for document visual question answering on multiple images.
MMLongBench-Doc	[141]	Benchmarking long-context document understanding with visualizations.

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

Dataset/Benchmark Name	Citation	Key Details & Data Sources
MS COCO (Common Objects in Context)	[56]	Widely used dataset (330,000+ images) for object detection, segmentation, VQA, and captioning.
Visual Genome	[79]	Provides dense annotations (3.8M objects, 2.3M relationships) to bridge images and language, enabling reasoning tasks.
Flickr30K Entities	[142]	Extends Flickr30K with bounding box annotations and coreference chains for phrase grounding.
ImageBind (Meta AI)	[143]	Large-scale dataset linking images with six modalities (text, audio, depth, thermal, IMU) for unified multimodal embeddings.
LAION-5B	[144]	One of the largest open multimodal datasets (5.85 billion image-text pairs) for training foundation models.
Conceptual Captions (CC3M)	[145]	Contains ∼3.3 million image-caption pairs extracted and filtered from the web, designed for automatic image captioning.
VizWiz	[146]	Benchmark consisting of visual questions originating from blind people.

GQA	[93]	Developed to address the limitations of VQAv2, offering rich semantic and visual complexity for real-world visual reasoning.
VQAv2	[147]	A benchmark using pairs of similar images leading to different answers to compel models to prioritize visual data.
OCRBench	[148]	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)	[149]	Large-scale multimodal benchmark evaluating VLMs on out-of-distribution object detection, covering seven domains.
NLVR	[150]	A corpus for reasoning about natural language grounded in photographs (NLVR2 is the related task in VLUE [66]).
Massive Multitask Language Understanding (MMLU)		Crucial benchmark for evaluating general knowledge and reasoning across 57 diverse subjects.

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

Dataset/Benchmark Name	**Citation**	**Key Details & Data Sources**
MVBench	[151]	A comprehensive multi-modal video understanding benchmark focusing on temporal perception.
Perception Test	[152]	A diagnostic benchmark for multimodal video models, covering Memory, Abstraction, Physics, and Semantics.
MSR VTT	[153]	A large video captioning dataset (10,000 video clips, 200,000 clip–sentence pairs) bridging video content and natural language.
VaTeX (Video And Text)	[154]	A multilingual video captioning dataset (English and Chinese) with 41,250 videos and 825,000 captions.
Dynamic-SUPERB	[155]	A benchmark assessing MLLMs' ability to follow instructions in the audio domain, focusing on human speech processing.
AIR-Bench	[156]	A comprehensive benchmark designed to evaluate MLLMs' ability to comprehend various audio signals (speech, natural sounds, music) and interact according to instructions.
MuChoMusic	[157]	The first benchmark for evaluating music understanding in audio MLLMs.
MCUB (Multimodal Commonality Understanding Benchmark)	[158]	Includes four modalities image, audio, video, and point cloud measuring the model's ability to identify commonalities among input entities.
M3DBench	[159]	Focuses on 3D instruction following.
ScanQA	[160]	3D question answering for spatial scene understanding.
AVQA	[161]	Designed for audio-visual question answering on general videos of real-life scenarios.
MMT-Bench	[162]	A comprehensive benchmark assessing MLLMs across massive multimodal tasks toward multitask AGI.

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) [163–165]

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).

Key Models and Architectures

1. **DeViSE (Deep Visual-Semantic Embedding Model) [165]**
   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) [163,166]**

   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 [167,167] and GoogLeNet [168,168] were commonly used, often employing transfer learning by leveraging knowledge learned on large vision datasets like ImageNet [169,169]. The fused output was then typically passed to a classifier or generator.

3. **NeuralTalk / Neural-Image-QA [164]**

   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) [33,170–172]

This period is defined by the proliferation of the Transformer architecture [33], 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 [170,170]**

   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) [173,173].

   Training & Contributions: Served as a highly performant and relatively simple baseline for vision and language tasks.

2. **ViLBERT (Pretraining Task-Agnostic Visiolinguistic Representations) [171,171]**

   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) [172,172]**

   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.

Recent Large-Scale MLLMs (2021–2025) [44,51,174–176]

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) [44,44]**

   Year: 2021.

   Architecture: Encoder–decoder model, using Vision Transformers (ViT) [177,178] or ResNets as the vision encoder.

   Training & Contributions: Trained using a contrastive learning objective on 400M image-text pairs [44], 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 [51]**

   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 [176,179]**

   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 [180?]**
   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) [174,181]**
   Year: 2023.
   Architecture & Training: Details are undisclosed.
   Unique Contributions: Recognized as a pioneering MLLM [174,175], 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 [175,182]**
   Year: 2023.
   Architecture & Training: A family of models utilizing a decoder-only architecture [175,175]. Details are undisclosed.
   Unique Contributions: Gemini excels in providing detailed, expansive answers, often incorporating relevant imagery and links, showcasing sophisticated multimodal capabilities [182].
7. **CogVLM [183?]**
   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. Conclusions

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.

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.

References

1. Ryu, J.S.; Kang, H.; Chu, Y.; Yang, S. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters* **2025**, *15*, 809–830. https://doi.org/10.1007/s13534-025-00484-6.

2. Liu, W.; Wu, G.; Wang, H.; Ren, F. Cross-Modal Data Fusion via Vision-Language Model for Crop Disease Recognition. *Sensors* **2025**, *25*, 4096. https://doi.org/10.3390/s25134096.
3. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision, 2021, [arXiv:cs.CV/2103.00020].
4. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021, [arXiv:cs.CV/2102.05918].
5. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023, [arXiv:cs.CV/2301.12597].
6. et al., O. GPT-4 Technical Report, 2024, [arXiv:cs.CL/2303.08774].
7. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers, 2019, [arXiv:cs.CL/1908.07490].
8. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, 2019, [arXiv:cs.CV/1908.02265].
9. Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O.K.; Patra, B.; et al. Language Is Not All You Need: Aligning Perception with Language Models, 2023, [arXiv:cs.CL/2302.14045].
10. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning, 2023, [arXiv:cs.CV/2304.08485].
11. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*, 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.
12. Qin, R.; Institutes, A. Tiny-Align: Bridging Automatic Speech Recognition and Large Language Model on Edge, 2024, [arXiv:cs.CL/2411.13766]. Accessed: 2025-11-01.
13. Han, X.; Chen, S.; Fu, Z.; Feng, Z.; Fan, L.; An, D.; Wang, C.; Guo, L.; Meng, W.; Zhang, X.; et al. Multimodal fusion and vision–language models: A survey for robot vision. *Information Fusion* **2026**, *126*, 103652. https://doi.org/10.1016/j.inffus.2025.103652.
14. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023, [arXiv:cs.RO/2307.15818].
15. Zong, Y.; Aodha, O.M.; Hospedales, T. Self-Supervised Multimodal Learning: A Survey, 2024, [arXiv:cs.LG/2304.01008].
16. Zhong, C.; Zeng, S.; Zhu, H. Adaptive Multimodal Fusion with Cross-Attention for Robust Scene Segmentation and Urban Economic Analysis. *Applied Sciences* **2025**, *15*, 438. https://doi.org/10.3390/app15010438.
17. Kress, G. *Multimodality: A social semiotic approach to contemporary communication*; Routledge: London, 2010. Definition of 'mode' in source [4], cited in [14].
18. Saleh, M.; Tabatabaei, A. 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.
19. Van Leeuwen, T. *Introducing social semiotics*; Psychology Press, 2005. Definition of Multimodal Discourse in source [1].
20. 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]).
21. Milvus. How is multimodal AI used in robotics? **2025**. Discusses multimodal AI integration in robotics (Source [13]).
22. Singh, G. A Review of Multimodal Vision–Language Models: Foundations, Applications, and Future Directions. *Preprints* **2025**. https://doi.org/10.20944/preprints202510.2511.v1.
23. Singh, G.; Banerjee, T.; Ghosh, N. Tracing the Evolution of Artificial Intelligence: A Review of Tools, Frameworks, and Technologies (1950–2025). *Preprints* **2025**. https://doi.org/10.20944/preprints202511.0637.v1.
24. Singh, G. AI-Assisted Storytelling: Enhancing Narrative Creation in Digital Media. *International Journal of Engineering Development and Research* **2026**, *14*, 882–894.
25. Singh, G.; Naaz, A.; Syed, A.; Akhila, V. AI-Assisted Storytelling: Enhancing Narrative Creation in Digital Media. *Preprints* **2026**. https://doi.org/10.20944/preprints202601.0330.v1.
26. GeeksforGeeks. Early Fusion vs. Late Fusion in Multimodal Data Processing **2025**. Last Updated: 23 Jul, 2025.
27. Karani, R.; Desai, S. Review on Multimodal Fusion Techniques for Human Emotion Recognition. *The Science and Information (SAI) Organization* **2022**, *13*.

28. Milvus. What fusion strategies work best for combining results from different modalities? **2025**. AI Reference.
29. Shankar, S.; Thompson, L.; Fiterau, M. Progressive Fusion for Multimodal Integration. In Proceedings of the arXiv:2209.00302v2 [cs.LG], 2022.
30. Aladago, M.M.; Piergiovanni, A. COMPOUND TOKENS: CHANNEL FUSION FOR VISION-LANGUAGE REPRESENTATION LEARNING. In Proceedings of the OpenReview: ICLR 2023 Tiny Papers Track, 2023.
31. Wikipedia contributors. Multimodal learning. *Wikipedia, The Free Encyclopedia* **2024**. Retrieved on YYYY-MM-DD.
32. Chen, J.; Yang, J.; Wu, H.; Li, D.; Gao, J.; Zhou, T.; Xiao, B. Florence-VL: Enhancing Vision-Language Models with Generative Vision Encoder and Depth-Breadth Fusion. *CVF Open Access* **2024**.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. *Advances in Neural Information Processing Systems* **2017**, *30*.
34. Zhao, W.C.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv preprint arXiv:2303.18223* **2023**.
35. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **2020**, *21*, 5485–5551.
36. Soltan, S.; Ananthakrishnan, S.; FitzGerald, J.; Gupta, R.; Hamza, W.; Khan, H.; Peris, C.; Rawls, S.; Rosenbaum, A.; Rumshisky, A.; et al. AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448* **2022**.
37. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* **2022**.
38. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
39. Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilic, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100* **2022**.
40. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **2023**.
41. Tay, Y.; Wei, J.; Chung, H.W.; Tran, V.Q.; So, D.R.; Shakeri, S.; Garcia, X.; Zheng, H.S.; Rao, J.; Chowdhery, A.; et al. Transcending scaling laws with 0.1% extra compute. *arXiv preprint arXiv:2210.11399* **2022**.
42. Ren, X.; Zhou, P.; Meng, X.; Huang, X.; Wang, Y.; Wang, W.; Li, P.; Zhang, X.; Podolskiy, A.; Arshinov, G.; et al. Pangu-Σ: Towards trillion parameter language model with sparse heterogeneous computing. *arXiv preprint arXiv:2303.10845* **2023**.
43. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* **2023**.
44. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* **2021**.
45. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 61–68.
46. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the Conference on Neural Information Processing Systems, 2019.
47. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv preprint arXiv:2205.01917* **2022**.
48. Xu, P.; Zhu, X.; Clifton, D.A.; et al. Multimodal Learning with Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **2023**.
49. Chen, G.; Liu, F.; Meng, Z.; Liang, S. Revisiting parameter-efficient tuning: Are we really there yet? *arXiv preprint arXiv:2202.07962* **2022**.

50. Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. Gpt understands, too. In Proceedings of the arXiv preprint arXiv:2103.10385, 2021.
51. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **2022**, *35*, 23716–23736.
52. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *Advances in Neural Information Processing Systems* **2023**, *36*, 34892–34916.
53. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281* **2023**.
54. Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR* **2023**, *2306.13394*.
55. Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR* **2024**, *2306.13394*.
56. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the ECCV. Springer, 2014, pp. 740–755.
57. Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* **2023**.
58. Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* **2023**.
59. Li, B.; Ge, Y.; Ge, Y.; Wang, G.; Wang, R.; Zhang, R.; Shan, Y. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.16911* **2023**.
60. Yin, Z.; Wang, J.; Cao, J.; Shi, Z.; Liu, D.; Li, M.; Sheng, L.; Bai, L.; Huang, X.; Wang, Z.; et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *NeurIPS Datasets and Benchmarks* **2023**.
61. Lin, W.; Wei, X.; An, R.; Gao, P.; Zou, B.; Luo, Y.; Huang, S.; Zhang, S.; Li, H. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2404.18029* **2024**.
62. Shi, Z.; Wang, Z.; Fan, H.; Yin, Z.; Sheng, L.; Qiao, Y.; Shao, J. Chef: A comprehensive evaluation framework for standardized assessment of multimodal large language models. *arXiv preprint arXiv:2310.11585* **2023**.
63. Al-Tahan, H.; Garrido, Q.; Balestriero, R.; Bouchacourt, D.; Hazirbas, C.; Ibrahim, M. UniBench: Visual Reasoning Requires Rethinking Vision-Language Beyond Scaling. *arXiv preprint arXiv:2401.12781* **2024**.
64. Bai, S.; Yang, S.; Bai, J.; Wang, P.; Zhang, X.; Lin, J.; Wang, X.; Zhou, C.; Zhou, J. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2310.15053* **2023**.
65. Zeng, Y.; Zhang, H.; Zheng, J.; Xia, J.; Wei, G.; Wei, Y.; Zhang, Y.; Kong, T. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2310.00794* **2023**.
66. Zhou, W.; Zeng, Y.; Diao, S.; Zhang, X. VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models. In Proceedings of the ICML, 2022, Vol. 162.
67. Liu, F.; Bugliarello, E.; Ponti, E.M.; Reddy, S.; Collier, N.; Elliott, D. Visually grounded reasoning across languages and cultures. *EMNLP* **2021**, pp. 10467–10485.
68. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W.X.; Wen, J.R. Evaluating object hallucination in large vision-language models. *EMNLP* **2023**.
69. Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the CVPR, 2023, pp. 14375–14385.
70. Rohrbach, A.; Hendricks, L.A.; Burns, K.; Darrell, T.; Saenko, K. Object hallucination in image captioning. In Proceedings of the EMNLP, 2018, pp. 4035–4045.
71. Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR* **2023**, *2311.07397*.
72. Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR* **2024**, *2311.07397*.
73. Villa, A.; Léon, J.; Soto, A.; Ghanem, B. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. *CVPR* **2025**, pp. 492–502.
74. Wang, J.; Zhou, Y.; Xu, G.; Shi, P.; Zhao, C.; Xu, H.; Ye, Q.; Yan, M.; Zhang, J.; Zhu, J.; et al. Evaluation and analysis of hallucination in large vision-language models. *CoRR* **2023**, *2308.15126*.

75. Wu, M.K.; Ji, J.; Huang, O.; Li, J.; Wu, Y.; Sun, X.; Ji, R. Evaluating and analyzing relationship hallucinations in large vision-language models. *ICML* **2024**.
76. Jiang, C.; Ye, W.; Dong, M.; Jia, H.; Xu, G.; Yan, M.; Zhang, J.; Zhang, S. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *ACM MM* **2024**.
77. Huang, W.; Liu, H.; Guo, M.; Gong, N.Z. Visual hallucinations of multi-modal large language models. *Findings of the ACL* **2024**, pp. 9614–9631.
78. Qiu, H.; Huang, J.; Gao, P.; Qi, Q.; Zhang, X.; Shao, L.; Lu, S. Longhalqa: Long-context hallucination evaluation for multimodal large language models. *CoRR* **2024**, *2410.09962*.
79. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* **2016**, *123*, 32–73.
80. Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; Sun, J. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the ICCV, 2019, pp. 8430–8439.
81. Liu, J.; Fu, Y.; Xie, R.; Xie, R.; Sun, X.; Lian, F.; Kang, Z.; Li, X. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *CVPR* **2025**, pp. 19857–19866.
82. Kafle, K.; Kanan, C. An analysis of visual question answering algorithms. In Proceedings of the ICCV, 2017, pp. 1965–1973.
83. Seth, A.; Manocha, D.; Agarwal, C. Hallucinogen: A benchmark for evaluating object hallucination in large visual-language models. *CoRR* **2024**, *2412.20622*.
84. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CVPR* **2017**, pp. 2097–2106.
85. Chen, X.; Wang, C.; Xue, Y.; Zhang, N.; Yang, X.; Li, Q.; Shen, Y.; Liang, L.; Gu, J.; Chen, H. Unified hallucination detection for multimodal large language models. *ACL* **2024**.
86. Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; Fung, P. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ALVR Workshop* **2024**, pp. 37–58.
87. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* **2020**, *128*, 1956–1981.
88. Hu, H.; Zhang, J.; Zhao, M.; Sun, Z. Ciem: Contrastive instruction evaluation method for better instruction tuning. *NeurIPS Workshop* **2023**.
89. Chen, Z.; Zhu, Y.; Zhan, Y.; Li, Z.; Zhao, C.; Wang, J.; Tang, M. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479* **2023**.
90. Chen, X.; Ma, Z.; Zhang, X.; Xu, S.; Qian, S.; Yang, J.; Fouhey, D.; Chai, J. Multi-object hallucination in vision language models. *NeurIPS* **2024**, *37*, 44393–44418.
91. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ade20k dataset. *CVPR* **2017**, pp. 633–641.
92. Cao, Q.; Cheng, J.; Liang, X.; Lin, L. VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In Proceedings of the ACL, 2024, pp. 12161–12176.
93. Hudson, D.A.; Manning, C.D. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In Proceedings of the CVPR, 2019, pp. 6700–6709.
94. Zhai, B.; Yang, S.; Xu, C.; Shen, S.; Keutzer, K.; Li, C.; Li, M. Halle-control: controlling object hallucination in large multimodal models. *CoRR* **2023**, *2310.01779*.
95. Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. *ICLR* **2023**.
96. Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.Y.; Wang, Y.X.; Yang, Y.; et al. Aligning large multimodal models with factually augmented rlhf. *Findings of the ACL* **2024**, pp. 13088–13110.
97. Wang, L.; He, J.; Li, S.; Liu, N.; Lim, E.P. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *MMM* **2023**.
98. Rani, A.; Rawte, V.; Sharma, H.; Anand, N.; Rajbangshi, K.; Sheth, A.; Das, A. Visual hallucination: Definition, quantification, and prescriptive remediations. *CoRR* **2024**, *2403.17306*.
99. Chen, J.; Yang, D.; Wu, T.; Jiang, Y.; Hou, X.; Li, M.; Wang, S.; Xiao, D.; Li, K.; Zhang, L. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185* **2024**.
100. Liu, B.; Zhan, L.M.; Xu, L.; Ma, L.; Yang, Y.; Wu, X.M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *ISBI* **2021**, pp. 1650–1654.

101. Wu, X.; Guan, T.; Li, D.; Huang, S.; Liu, X.; Wang, X.; Xian, R.; Shrivastava, A.; Huang, F.; Boyd-Graber, J.; et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *Findings of the EMNLP* **2024**, pp. 8395–8419.
102. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. *ICML* **2021**, pp. 8821–8831.
103. Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; Smith, N.A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In Proceedings of the ICCV, 2023, pp. 20349–20360.
104. Huang, Z.; He, W.; Long, Q.; Wang, Y.; Li, H.; Yu, Z.; Shu, F.; Chan, L.; Jiang, H.; Gan, L.; et al. T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. *CoRR* **2024**, *2412.04300*.
105. Huang, K.C.; Sun, K.; Xie, E.; Li, Z.; Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In Proceedings of the NeurIPS, 2023, Vol. 36, pp. 78723–78747.
106. Niu, Y.; Ning, M.; Zheng, M.; Lin, B.; Jin, P.; Liao, J.; Ning, K.; Zhu, B.; Yuan, L. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *CoRR* **2025**, *2503.07265*.
107. Gokhale, T.; Palangi, H.; Nushi, B.; Vineet, V.; Horvitz, E.; Kamar, E.; Baral, C.; Yang, Y. Benchmarking spatial relationships in text-to-image generation. *CoRR* **2022**, *2212.10015*.
108. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Lopes, R.G.; Ayan, B.K.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* **2022**, *35*, 36479–36494.
109. Feng, W.; He, X.; Fu, T.J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X.E.; Wang, W.Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ICLR* **2023**.
110. Li, B.; Lin, Z.; Pathak, D.; Li, J.; Fei, Y.; Wu, K.; Xia, X.; Zhang, P.; Neubig, G.; Ramanan, D. Evaluating and improving compositional text-to-visual generation. *CVPR* **2024**, pp. 5290–5301.
111. Bakr, E.M.; Sun, P.; Shen, X.; Khan, F.F.; Li, L.E.; Elhoseiny, M. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. In Proceedings of the ICCV, 2023, pp. 20041–20053.
112. Cho, J.; Zala, A.; Bansal, M. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In Proceedings of the ICCV, 2023, pp. 3043–3054.
113. Lim, Y.; Choi, H.; Shim, H. Evaluating image hallucination in text-to-image generation with question-answering. *AAAI* **2025**, *39*, 26290–26298.
114. Ben-Kish, A.; Yanuka, M.; Alper, M.; Giryes, R.; Averbuch-Elor, H. Mitigating open-vocabulary caption hallucinations. *EMNLP* **2024**, pp. 22680–22698.
115. Tu, Y.; Hu, R.; Sang, J. Ode: Open-set evaluation of hallucinations in multimodal large language models. *CVPR* **2025**, pp. 19836–19845.
116. Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *CVPR* **2024**, pp. 9556–9567.
117. Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.W.; Galley, M.; Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* **2023**.
118. Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.W.; Zhu, S.C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Neurips* **2022**, *35*, 2507–2521.
119. Mialon, G.; Fourrier, C.; Swift, C.; Wolf, T.; LeCun, Y.; Scialom, T. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983* **2023**.
120. Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; Li, H. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2407.10657* **2024**.
121. Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. Are we on the right way for evaluating large vision-language models? *NeurIPS* **2024**, *37*, 27056–27087.
122. Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C.L.; Girshick, R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In Proceedings of the CVPR, 2016, pp. 1988–1997.
123. Xia, P.; Chen, Z.; Tian, J.; Gong, Y.; Hou, R.; Xu, Y.; Wu, Z.; Fan, Z.; Zhou, Y.; Zhu, K.; et al. CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models. *arXiv preprint arXiv:2410.19830* **2024**.

124. Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; Luo, P. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. *arXiv preprint arXiv:2402.09181* **2024**.
125. Johnson, A.E.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Lu, Y.; Mark, R.G.; Berkowitz, S.J.; Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* **2019**.
126. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **2016**, *23*, 304–310.
127. Luo, Y.; Shi, M.; Khan, M.O.; Afzal, M.M.; Huang, H.; Yuan, S.h.; Tian, Y.; Song, L.; Kouhana, A.; Elze, T.; et al. Fairclip: Harnessing fairness in vision-language learning. *arXiv preprint arXiv:2403.19949* **2024**.
128. Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *MICCAI* **2023**, pp. 525–536.
129. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* **2023**.
130. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018, Vol. 5, pp. 1–9.
131. Zambrano Chaves, J.M.; Wentland, A.L.; Desai, A.D.; Banerjee, I.; Kaur, G.; Correa, R.; Boutin, R.D.; Maron, D.J.; Rodriguez, F.; Sandhu, A.T.; 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* **2023**, *13*, 21034.
132. Lau, J.J.; Gayen, S.; Abacha, A.B.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* **2018**, *5*, 1–10.
133. Huybrechts, G.; Ronanki, S.; Jayanthi, S.M.; Fitzgerald, J.; Veeravanallur, S. Document Haystack: A Long Context Multimodal Image/Document Understanding Vision LLM Benchmark. *Amazon Science* **2024**.
134. Wang, H.; Shi, H.; Tan, S.; Qin, W.; Wang, W.; Zhang, T.; Nambi, A.; Ganu, T.; Wang, H. Needle in a multimodal haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.07230* **2024**.
135. Chia, Y.K.; Cheng, L.; Chan, H.P.; Liu, C.; Song, M.; Aljunied, S.M.; Poria, S.; Bing, L. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176* **2024**.
136. Kamradt, G. Needle in a haystack-pressure testing llms. *Github Repository* **2023**, p. 28.
137. Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508* **2023**.
138. Song, D.; Chen, S.; Chen, G.H.; Yu, F.; Wan, X.; Wang, B. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532* **2024**.
139. Van Landeghem, J.; Tito, R.; Borchmann, Ł.; Pietruszka, M.; Jóźiak, P.; Powalski, R.; Jurkiewicz, D.; Coustaty, M.; Ackaert, B.; Valveny, E.; et al. Document understanding dataset and evaluation (dude). *ICCV* **2023**, pp. 19528–19540.
140. Tanaka, R.; Nishida, K.; Nishida, K.; Hasegawa, T.; Saito, I.; Saito, K. Slidevqa: A dataset for document visual question answering on multiple images. *AAAI* **2023**, *37*, 13636–13645.
141. Ma, Y.; Zang, Y.; Chen, L.; Chen, M.; Jiao, Y.; Li, X.; Lu, X.; Liu, Z.; Ma, Y.; Dong, X.; et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523* **2024**.
142. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the ICCV, 2015, pp. 2641–2649.
143. Girdhar, R.; El-Nouby, A.; Mangalam, K.; Singh, P.; Han, X.; Kopuluru, A.; Joulin, A.; Taveres, I. Imagebind: One embedding space to bind them all. In Proceedings of the CVPR, 2023, pp. 15180–15190.
144. Schuhmann, C.; Beaumont, R.; Vencovsky, R.; Gordon, R.; Wightman, M.; Jitsev, A.; et al. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. *arXiv preprint arXiv:2210.16084* **2022**.
145. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the ACL, 2018, pp. 2556–2565.
146. Gurari, D.; Li, Q.; Stangl, A.J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; Bigham, J.P. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218* **2018**.

147. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the CVPR, 2017, pp. 6904–6913.
148. Liu, Y.; Li, Z.; Yang, B.; Li, C.; Yin, X.; Liu, C.l.; Jin, L.; Bai, X. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895* **2024**.
149. Roboflow. RF100-VL: A Benchmark for Few-Shot Generalization in Vision-Language Models. *Research paper (Contextual Citation)* **2025**.
150. Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; Artzi, Y. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491* **2018**.
151. Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2312.00985* **2024**.
152. Pătrăucean, V.; Smaira, L.; Gupta, A.; Recasens Continente, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Heyward, J.; Malinowski, M.; Yang, Y.; et al. Perception test: A diagnostic benchmark for multimodal video models. *arXiv preprint arXiv:2303.13380* **2023**.
153. Xu, J.; Mei, T.; Yao, T.; Zhang, Y. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the CVPR, 2016, pp. 2601–2610.
154. Wang, X.; Wu, W.; Li, J.; Wang, X.; Liu, L.; Wu, Z.; Wang, J.; Wang, J. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. *ICCV* **2019**, pp. 5710–5719.
155. Huang, C.y.; Lu, K.H.; Wang, S.H.; Hsiao, C.Y.; Kuan, C.Y.; Wu, H.; Arora, S.; Chang, K.W.; Shi, J.; Peng, Y.; et al. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *arXiv preprint arXiv:2404.09068* **2024**.
156. Yang, Q.; Xu, J.; Liu, W.; Chu, Y.; Jiang, Z.; Zhou, X.; Leng, Y.; Lv, Y.; Zhao, Z.; Zhou, C.; et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2405.02384* **2024**.
157. Weck, B.; Manco, I.; Benetos, E.; Quinton, E.; Fazekas, G.; Bogdanov, D. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2405.01358* **2024**.
158. Chen, C.; Du, Y.; Fang, Z.; Wang, Z.; Luo, F.; Li, P.; Yan, M.; Zhang, J.; Huang, F.; Sun, M.; et al. Model composition for multimodal large language models. *arXiv preprint arXiv:2404.03212* **2024**.
159. Li, M.; Chen, X.; Zhang, C.; Chen, S.; Zhu, H.; Yin, F.; Yu, G.; Chen, T. M3dbench: Let's instruct large models with multi-modal 3d prompts. *arXiv preprint arXiv:2312.01255* **2023**.
160. Azuma, D.; Miyanishi, T.; Kurita, S.; Kawanabe, M. Scanqa: 3d question answering for spatial scene understanding. *arXiv preprint arXiv:2208.06456* **2022**.
161. Yang, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.; Zhu, W. Avqa: A dataset for audio-visual question answering on videos. In Proceedings of the ACM MM, 2022, pp. 3480–3491.
162. Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; Zhang, W.; Lin, Y.; Liu, S.; et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2407.13532* **2024**.
163. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.
164. Malinowski, M.; Rohrbach, M.; Fritz, M. Ask your neurons: A neural-based approach to answering questions about images. *Proceedings of the IEEE international conference on computer vision* **2015**, pp. 1–9.
165. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems* **2013**, pp. 2121–2129.
166. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in VQA matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.
167. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
168. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
169. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
170. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* **2019**.

doi:10.20944/preprints202602.0467.v1

171. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the Advances in Neural Information Processing Systems, 2019, Vol. 32, pp. 13–23.
172. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.
173. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*, 91–99.
174. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* **2023**.
175. Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* **2023**, *1*.
176. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597* **2023**.
177. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR, 2021.
178. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
179. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International conference on machine learning, 2022, pp. 12888–12900.
180. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744* **2023**.
181. Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.C.; Liu, Z.; Wang, L. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* **2023**, *9*, 1.
182. Qi, Z.; Fang, Y.; Zhang, M.; Sun, Z.; Wu, T.; Liu, Z.; Lin, D.; Wang, J.; Zhao, H. Gemini vs GPT-4V: A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases. *arXiv preprint arXiv:2312.15011* **2023**.
183. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079* **2023**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.