Review Article

Benchmark Evaluations, Applications, and Challenges of Large Vision Language Models: A Survey

Zongxia Li[1], Xiyang Wu[1], Hongyang Du[1], Huy Nghiem[1], Guangyao Shi[2]

1. University of Maryland, United States; 2. University of Southern California, United States

Multimodal Vision Language Models (vlms) have emerged as a transformative technology at the intersection of computer vision and natural language processing, enabling machines to perceive and reason about the world through both visual and textual modalities. For example, models such as CLIP[1], Claude[2], and GPT-4V[3] demonstrate strong reasoning and understanding abilities on visual and textual data and beat classical single modality vision models on zero-shot classification[4]. Despite their rapid advancements in research and growing popularity in applications, a comprehensive survey of existing studies on vlms is notably lacking, particularly for researchers aiming to leverage vlms in their specific domains. To this end, we provide a systematic overview of vlms in the following aspects: [1] model information of the major vlms developed over the past five years (2019-2024); [2] the main architectures and training methods of these vlms; [3] summary and categorization of the popular benchmarks and evaluation metrics of vlms; [4] the applications of vlms including embodied agents, robotics, and video generation; [5] the challenges and issues faced by current vlms such as hallucination, fairness, and safety. Detailed collections including papers and model repository links are listed in https://github.com/zli12321/Awesome-VLM-Papers-And-Models.git.

Zongxia Li and Xiyang Wu equally contributed to this work.

1. Introduction

Pretrained large language models (llms), such as LLaMA[5], GPT-4[6] have achieved remarkable success across a wide range of nlp tasks[7][8]. However, as these models continue to scale[9], they face two challenges: (1) The finite supply of high-quality text data[10][11]; (2) The inherent limitations of single-modality architectures in capturing and processing real-world information that requires understanding the complex relationships between different modalities[12][13]. These limitations motivate the efforts to explore and develop vlms, which combine both visual (e.g., images, videos) and textual inputs, providing a more comprehensive understanding of visual spatial relationships, objects, scenes, and abstract concepts[14][15]. vlms expand the representational boundaries that have previous confined single-modality approaches, supporting a richer and more contextually informed view of the world[16][17][18], such as visual question answering (vqa)[19], autonomous driving[20]. Meanwhile, vlms encounter new challenges distinct from single-modality models, such as visual hallucination, which occurs when vlms generate responses without meaningful visual comprehension, instead relying primarily on parametric knowledge stored in the llm component[21][22]. There are already several reviews on single-modality models[23][24] while the multi-modality one is still missing. In this paper, we provide a critical examination of research results on vlms, offering a systematic review of current major architectures of vlms, evaluation and benchmarks, applications, and challenges faced by vlms.

2. State-of-the-Art vlms

In recent years, leading Artificial Intelligence (AI) organizations are consistently releasing new VLMs[25]. From OpenAI's CLIP[26], Salesforce's BLIP[27], DeepMind's Flamingo[28] to GPT-4V[3] and Gemini[29], these models are becoming larger and more interactive and illustrate the integration of chatbot functionality within VLM frameworks to support multimodality user interaction to improve user experience. The SoTA VLMs from 2019 to the end of 2024 are listed in Table 1 according to the following three principal research directions.

Model	Year	Architecture	Training Data	Parameters	Vision Encoder / Tokenizer	Pretrained Backbone Model
VisualBERT[30]	2019	Encoder-only	COCO[31]	110M	Faster R-CNN[32]	Pretrained from scratch
CLIP[1]	2021	Encoder-decoder	400M image-text pairs	63M-355M	ViT[33] / ResNet[34]	Pretrained from scratch
BLIP[27]	2022	Encoder-decoder	COCO[31], Visual Genome[35]	223M-400M	ViT-B/L/g[33]	Pretrained from scratch
Flamingo[28]	2022	Decoder-only	M3W[28], ALIGN[36]	80B	Custom	Chinchilla[37]
BLIP-2[38]	2023	Encoder-decoder	COCO[31], Visual Genome[35]	7B-13B	ViT-g[33]	Open Pretrained Transformer (opt)[39]
GPT-4V[3]	2023	Decoder-only	Undisclosed	Undisclosed	Undisclosed	Undisclosed
Gemini[29]	2023	Decoder-only	Undisclosed	Undisclosed	Undisclosed	Undisclosed
LLaVA-1.5[40]	2023	Decoder-only	COCO[31]	13B	CLIP ViT-L/14[33]	Vicuna[41]
PaLM-E[42]	2023	Decoder-only	All robots, WebLI[43]	562B	ViT[33]	PaLM[44]
CogVLM[45]	2023	Encoder-decoder	LAION-2B[46], COYO-700M[47]	18B	CLIP ViT-L/14[33]	Vicuna[41]
InstructBLIP[48]	2023	Encoder-decoder	CoCo[31], VQAv2[49]	13B	ViT[33]	Flan-T5[50], Vicuna[41]
InternVL[51]	2023	Encoder-decoder	LAION-en[52], LAION-	7B/20B	Eva CLIP ViT-g[33]	QLLaMA[53]

Model	Year	Architecture	Training Data	Parameters	Vision Encoder / Tokenizer	Pretrained Backbone Model
			multi[52]			
Claude 3[2]	2024	Decoder-only	Undisclosed	Undisclosed	Undisclosed	Undisclosed
Emu3[54]	2024	Decoder-only	Aquila[55]	7B	MoVQGAN[56]	LLaMA-2[5]
NVLM[57]	2024	Encoder-decoder	LAION-115M[58]	8B-24B	Custom ViT	Qwen-2-Instruct[59]
Qwen2-VL[60]	2024	Decoder-only	Undisclosed	7B-14B	EVA-CLIP ViT-L[33]	Qwen-2[59]
Pixtral[61]	2024	Decoder-only	Undisclosed	12B	CLIP ViT-L/14[33]	Mistral Large 2[62]
LLaMA 3.2-vision[63]	2024	Decoder-only	Undisclosed	11B-90B	CLIP[1]	LLaMA-3.1[63]
Baichuan Ocean Mini[64]	2024	Decoder-only	Image / Video / Audio / Text	7B	CLIP ViT-L/14[33]	Baichuan[65]
TransFusion[66]	2024	Encoder-decoder	Undisclosed	7B	VAE Encoder[67]	Pretrained from scratch on transformer architecture
DeepSeek-VL2[68]	2024	Decoder-only	WiT[69], WikiHow[70]	4.5B x 74	SigLIP[71] / SAMB[72]	DeepSeekMoE[73][74]
Molmo[75]	2024	Decoder-only	PixMo[75]	1B-72B	CLIP ViT-L/14[33]	OLMoE[76] / OLMo[77] / Qwen-2[59]
OLMo-2[78]	2024	Decoder-only	OLMo-mix-1124[78]	7B-13B	GPT-NeoX-20B[79]	Pretrained from scratch

Table 1. There is a growing number of vlms released in recent years, has expanded rapidly in recent years, with architectural variations enabling better and deeper integration between visual and textual representations. However, most current SoTA models use pretrained language models as the backbone model recently. DeepSeek-VL2 has a mixture of experts (MoE) architecture. The table only shows the primary sources/composition of the training data.

Vision-Language correlation

considers how training objectives or architectural design facilitate multimodal integration[80]. Training objectives such as contrastive learning are exemplified by approaches like SimCLR[81], which is originally developed for self-supervised vision tasks, adapts neatly to multimodal settings by bringing paired images and text closer together in the embedding space while pushing apart unpaired examples. Vision-language architecture considers how structural choices in model design facilitate or constrain multimodal integration[80]. Older architectural approaches primarily train models from scratch (CLIP[82]), whereas more recent methods (LLaMA 3.2-vision[63]) leverage the power of pre-trained LLMs as a backbone to improve the ability to correlate vision and language to better understand visual content (Section 3).

Benchmarks and evaluation

focuses on designing, collecting, and generating multimodal data, primarily in the format of question-answering (QA), to test VLMs on a variety of tasks such as visual text understanding, chart understanding, video understanding (Section 4).

Applications of VLMs

focuses on deploying VLM models in real-world scenarios. Virtual applications typically involve controlling personal device screens or simulated agent game playing (Section 5.1). Meanwhile, physical applications of VLMs primarily pertain to interactions with real-world physical objects, such as robotic human interaction or autonomous driving (Section 5.3).

These three directions provide a structured framework for analyzing, comparing, and guiding future progress in the rapidly evolving domain of vision-language modeling.[25][18]

3. Building Blocks and Training Methods

The architectures of VLMs are changing from pre-training from scratch to using pre-trained LLMs as a backbone to align the vision and textual information (Table 1). However, the fundamental components remain largely unchanged. We summarize the most foundational and widely adopted architectural components of VLMs, followed by an explanation of the popular pre-training and alignment methods. Details of SoTA VLM are given in Table 1 to show the shift in basic VLM architectures and newer architecture innovations that fuse visual features with textual features by treating visual features as tokens (Section 3.4).

3.1. Common Architecture Components

Vision Encoder

plays a crucial role in projecting visual components into embedding features that align with embeddings from large language models (LLMs) for tasks such as text or image generation[83]. It is trained to extract rich visual features from image or video data, enabling integration with language representations[84][85].

Specifically, vision encoders used in many VLMs[86][60][57][51], are pretrained on large-scale multimodal or image data: These encoders are jointly trained on image-text pairs, allowing them to capture visual and language relationships effectively. Notable examples include CLIP[1], which aligns images and text embeddings via contrastive learning, and BLIP[58], which leverages bootstrapped pretraining for robust language-image alignment. Pretrained on large scale ImageNet[87] or Similar Datasets: These encoders are trained on vast amounts of labeled visual data or through self-supervised training[88], enabling them to capture domain-specific visual features. While initially unimodal, these encoders, such as ResNet[34] or Vision Transformers (ViTs)[33], can be adapted for multimodal tasks. They excel at extracting meaningful object-level features and serve as a solid foundation for vision-language models. Many SoTA VLMs, such as Qwen2-VL[60] and LLaVA[89], commonly incorporate pretrained vision encoders. These encoders not only provide robust and meaningful visual representations but are also highly effective for transfer learning[90]. They outperform randomly initialized encoders[91] by leveraging learned vision knowledge from their training domains.

doi.org/10.32388/GXR68Q

Text Encoder

projects tokenized text sequences into an embedding space, similar to how vision encoders process images. Models such as CLIP[1], BLIP[58], and ALIGN[36] use both an image encoder and a text encoder. These models use contrastive learning to align image and text embeddings in a shared latent space, effectively capturing cross-modal relationships. However, newer models, such as LLaVA[89], often do not include a dedicated text encoder. Instead, they rely on large language models (LLMs) (e.g., LLaMA[5], Vicuna[92]) for text understanding, integrating visual inputs through projection layers or cross-attention mechanisms[93]. This shift shows a growing trend of using the capabilities of LLMs over vision components for more versatile and advanced multimodal reasoning and generation tasks.

Text Decoder

leverages llms as the primary text generator, using visual encoders to project image features[94]. GPT-4V[6], Flamingo[95], and Kosmos-2[96] use this approach. These models typically use a minimal visual projection mechanism, allowing the powerful language decoder to generate contextually rich outputs. VisualBERT and VilBERT[97][30] provide the foundation to decoder architectures for multimodal pretraining. Training vlms from scratch typically requires a separate text decoder, whereas using llms as the backbone often uses the original decoders from the llm. (Figure 1).

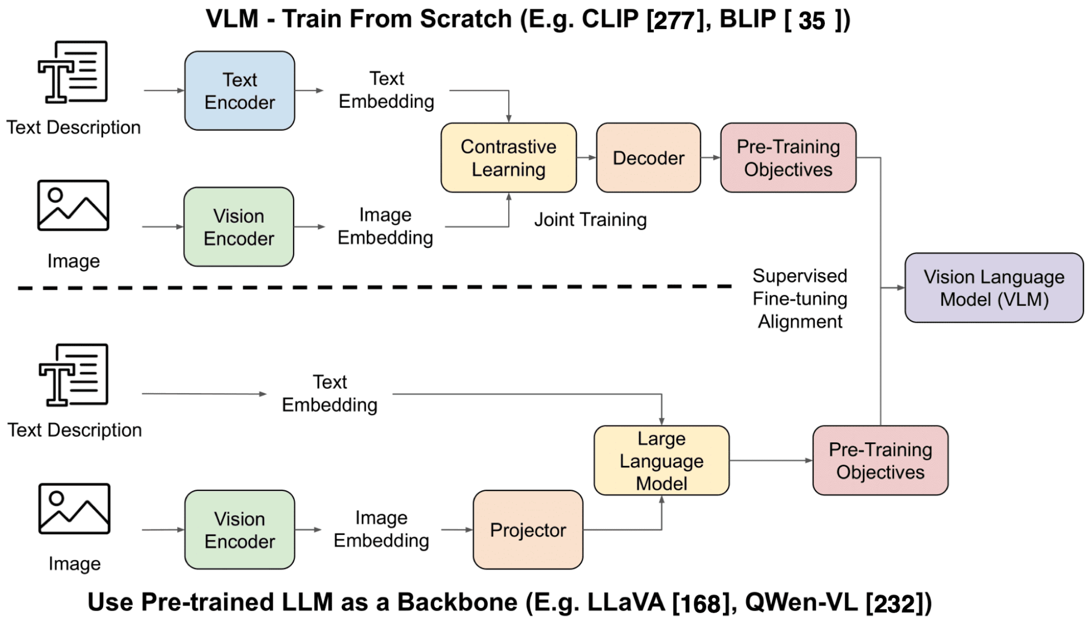


Figure 1. The basic components of common SoTA vlms.

Cross-Attention Mechanisms

enable dynamic interactions between visual and textual features by allowing tokens from one modality (vision) to influence tokens from the other modality (text)[93]. Cross-attention layers are commonly used to integrate information across modalities by computing attention scores between every pair of visual and textual tokens. Not all models use the cross-attention mechanism. For example, VisualBERT[97] and Flamingo[28] both have cross-attention mechanisms while CLIP[1] has no cross-attention..

3.2. Building Blocks of Training From Scratch

Training a vlm from scratch typically uses distinct training objectives and methodologies compared to using an llm as the backbone. Self-Supervised Learning (ssl) pre-trains without needing human labeled data to scale up pretraining[98]. Variants of ssl techniques include masked image modeling[99], contrastive learning[100], and image transformation prediction[101]. In this section, we delve into contrastive learning, a common pre-training process to scale up vlm training from scratch.

doi.org/10.32388/GXR68Q

Contrastive Learning

involves using separate encoders for visual and textual inputs, which are trained to map their respective modalities into a shared embedding space. The visual encoder processes images, generating feature embeddings from models like convolutional neural networks (cnn)[102] or vision transformers (ViTs)[103]. The text encoder processes textual inputs into embeddings. Contrastive learning aligns related image-text pairs by minimizing the distance between their visual and text embeddings in the shared space. At the same time, it maximizes the distance between embeddings of unrelated pairs. Pioneering models like CLIP[1], BLIP[104], and ALIGN[36] leverage this approach, pre-training on large-scale image-text datasets to develop robust, transferable representations for downstream tasks.

3.3. Building Blocks of Using llms as Backbone

Large Language Models

serve as the text generation component that processes encoded visual and textual inputs to produce text outputs autoregressively[105][5][6] for vlms. In the context of vlms, llms include their original text decoders. In this section, we list two common ways to align visual and pre-trained llm text features.

Projector

Projector maps visual features extracted by the vision encoder into a shared embedding space aligned with the text embeddings from the LLM. It typically consists of multi-layer perceptron (MLP) layers[106], which transform high-dimensional visual representations into compact embedding tokens compatible with the textual modality. The projector can be trained jointly with the rest of the model to optimize cross-modal objectives or freezing certain parts of the model, such as the LLM, to preserve pre-trained knowledge. Most cotemporary examples include LLaVA[86], QWen-2-VL[60], Nvidia VLM[57], Baichuan Ocean-mini[64], Emu3[54], and Pixtral (multimodal decoder)[61] .

Joint Training

is an end-to-end approach that updates weights of all components of the model in parallel without freezing any weights, including the LLM and projector layers. This approach has been used in models such as Flamingo[28].

Freeze Training Stages

involves selectively freezing model components during training, preserving pre-trained knowledge while adapting to new tasks[107]. Common strategies include freezing pre-trained vision encoders while fine-tuning projector layers, and implementing gradual unfreezing of components[108] or freezing LLM layers while only updating vision encoder weights[109].

3.4. Newer Architectures

Recent works have focused on enhancing the fusion of visual and textual features which we discuss in this section.

Treating all modalities as tokens

is a more recent approach that reads and encodes visual inputs (images and videos) as tokens similar to text tokens. Emu3[110] uses SBER-MoVQGAN to encode visual inputs into tokens and employs special separators, such as [SOT] and [EOV], to mark the start and end of visual tokens.[1] It still retains the LLMs architectures such as Llama[5], but comes with an expansion of the embedding layer to accommodate discrete vision tokens (Root Mean Square Layer Normalizatio layer[111] and Multi-query attention[112]). Additionally, it treats the generation of both visual and textual outputs as a token prediction task for a unified multimodal representation.

Transfusion

processes different modalities simultaneously within a single transformer architecture[66]. This method treats discrete text tokens and continuous image vectors in parallel by introducing strategic break points. While not yet perfected, this approach shows promising potential for developing more unified multimodal models that can handle diverse input types.

4. Benchmarks and Evaluation

The number of VLM benchmarks has grown rapidly with the quick development of new VLMs since 2022[113][114]. Comprehensive benchmarking is important for evaluating model performance and ensuring robust training across different capabilities various aspects such as math reasoning, scene recognition, etc[115][49]. Modern VLM benchmarks have moved beyond simple tasks like basic visual question answering to include a wider range of tests that better evaluate the models' multimodal abilities from more aspects[116]. In this section, we summarize and categorize existing 38 vision-language benchmarks for evaluating VLMs, including image-text and video-text benchmarks. We then summarize the commonly used evaluation metrics for these benchmarks, the typical methods for creating benchmark datasets, and the strengths and weaknesses of current benchmarks and evaluation practices. We highlight how most benchmarks prioritize data diversity and quantity while often overlooking improvements in evaluation quality, which hinders the effective assessment of VLMs.

Benchmark Categorization.

Benchmarks are designed with specific testing objectives, and we classify to ten primary categories (Table 2).

Category	Description	Datasets
Visual text understanding	Evaluates models' ability to extract and understand texts within visual components	TextVQA[117], DocVQA[118]
Multilingual multimodal understanding	Evaluates VLMs on different languages on different tasks such as question answering and reasoning	MM-En/CN[119], CMMLU[120], C-Eval[121], MTVQA[122]
Visual math reasoning	Tests models' ability to solve math problems in image forms	MathVista[115], MathVision[123], MM-Vet[124]
Optical Character Recognition (OCR)	Test models' ability to extract objects from visual inputs	MM-Vet[124], OCRBench[125], MME[126], MMTBench[127]
Chart graphic understanding	Evaluates models' ability to interpret graphic-related data	infographic VQA[128], AI2D[129], ChartQA[130], MMMU[131]
Text-to-Image generation	Evaluates models' ability to generate images	MSCOCO[31], GenEval[132], T2I-CompBench[133], DPG-Bench[134], VQAScore[135], GenAI-Bench[136]
Hallucination	Evaluates whether models are likely to hallucinate on certain visual and textual inputs	HallusionBench[21], POPE[137]
Multimodal general intelligence	Evaluates models' ability on diverse domains of tasks	MMLU[138], MMMU[131], MMStar[139], M3GIA[140], AGIEval[141]
Video understanding	Evaluates models' ability to understand videos (sequences of images)	EgoSchema[142], MLVU[143], MVBench[144], VideoMME[145], Perception-Test[146]
Visual reasoning, understanding, recognition, and question answering	Evaluate VLMs' ability to recognize objects, answer questions, and reason through both visual and textual information	MMTBench[127], GQA[147], MM-En/CN[119], VCR[148], VQAv2[49], MM-Vet[124], MMU[119], SEEDBench[149], Real World QA[150], MMMU-Pro[151], DPG[134] , MSCOCO-30K[31], MM-Vet[124], ST-VQA[152], NaturalBench[153]

Category	Description	Datasets
Robot simulator, web agent simulator	Evaluate the embodied vlms' abilities online in rule-based simulators	Habitat[154], Gibson[155], iGibson[156], Isaac Lab[157], WebArena[158], CALVIN[159], VLMBench[160], GemBench[161], VIMA-Bench[162]
Robotic benchmarks	Evaluate the embodied vlms' abilities using offline datasets recording collected interactions	Habitat[154], Gibson[155], iGibson[156]
Generative model, world model	Evaluate the embodied AI models' abilities with interactive models representing the environments	GAIA-1[163], UniSim[164], LWM[165], Genesis[166]

Table 2. The categories are surveyed from 15 SoTA vlm papers and collect the popular evaluation benchmarks used, categorized to 10 categories.

4.1. How Are Benchmark Data Collected

Benchmark datasets are typically created using one of three common data collection pipelines: fully human-annotated datasets; partially human-annotated datasets scaled up with synthetic data generation and partially validated by humans; and partially human-annotated datasets scaled up with synthetic data and fully validated by humans.

Fully human-annotated datasets

are created by having humans collect or generate adversarial or challenging test questions from diverse subjects and fields. For example, MMMU[131] has 50 college students from various disciplines to collect existing test questions from textbooks and lecture materials, often in multiple choice format. Another approach involves humans creating questions and having annotators provide answers to these questions. In VCR[148], Mechanical Turks are tasked with using contexts, detected objects, and images to write one to three questions about each image, along with reasonable answers and explanations. Fully human annotated datasets are time-consuming and hard to scale up, which brings inspiration to automatic question generation with human validation.

Synthetic question generation

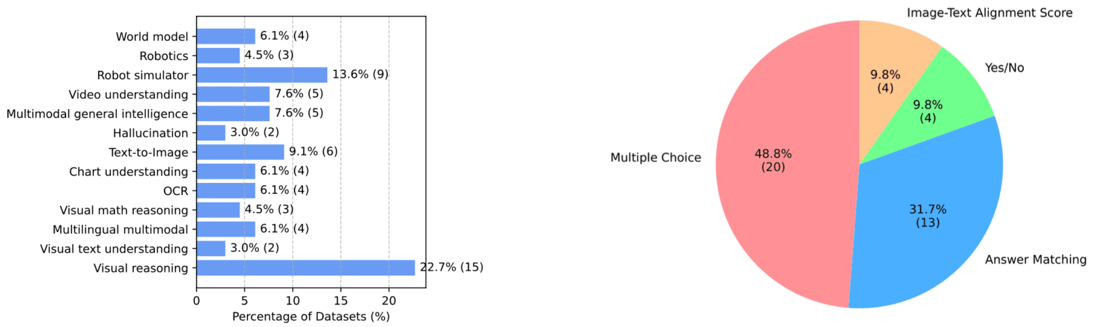
has become a more popular part of benchmark generation pipeline on various disciplines such as chart understanding[130], video understanding[142] to quickly scale up dataset sizes. Common practices include using human written examples as seed examples, giving a powerful llm to generate more adversarial example questions and answers[149]. Often, the generation process is only involved with texts. Chart and video data are often paired with visual content and captions, which are often used by authors as context to prompt llms to extract answers and generate questions[142][144]. However, llms are not always accurate and may produce unfaithful content or hallucinations[167]. To address this, pipelines typically include automatic filters to remove low-quality outputs, followed by crowdworker validation of either randomly sampled or all generated examples[130][149][142]. Automatic benchmark generation helps scale dataset size with reduced human effort. However, current automatic question-generation methods primarily rely on captions and textual contexts, which can lead to the creation of questions that are easy to answer without requiring significant visual reasoning[21], which undermines the benchmark's primary goal—evaluating a vlm's ability to comprehend and reason about visual content.

Interaction in the Simulator

is mainly targeted at VLM benchmarks in robotics . It gathers data for training and evaluation by assessing the VLM-powered agents online. As a data generation method stemming from reinforcement learning, such a data generation method is applicable for those scenarios that human-labeled datasets or synthetic datasets are hard and expensive to acquire, while the data construction follows some common rules like the physical law or some other common senses. With this rule-based data acquisition method, the outcome VLMs are more robust to the deviation within the multimodal inputs. During recent years, many works focus on realistic simulators for either robotics[154][155][156][157][159][160][161] and web agents[158] to simulator human agents or robots' interactions with the physical world. Nonetheless, benchmarks[154][155][156] based on the interaction data records from the simulator are also widely used for VLM agents training and evaluation. Notably, more efforts have been used for generative model[164] or even world model[165][163][166] to replace the previous simulators or datasets in generating more practical and better-quality datasets for VLMs. Though simulators are widely used in training and evaluating the VLM-power agents, the potential sim2real gap might exist when transplanting the terminal VLM into real-world applications, *i.e.* the VLM-powered agents might not be able to handle some real-world situations. More efforts towards the mitigation of these issues are still expected in this direction.

4.2. Evaluation Metrics

Benchmarks are designed for evaluation, with metrics established during their creation. VLM evaluation metrics are automatic to support repeated use at scale, and they often influence the question formats used in the benchmarks. We show the common evaluation metrics used in our surveyed benchmarks (Figure 2b, Figure 3).



(a) Most of our surveyed data tests VLMs' visual reasoning abilities.

(b) Majority of the benchmarks are designed in multiple choice format for ease of evaluations.

Figure 2. Our surveyed benchmark dataset categories and common evaluation practices.

doi.org/10.32388/GXR68Q

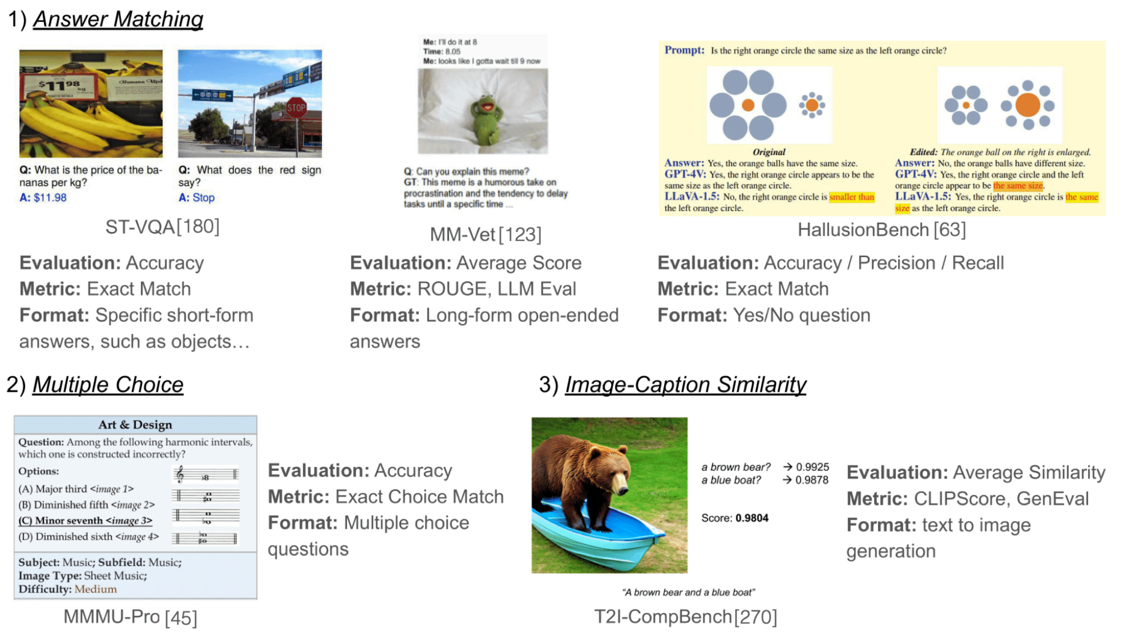

1) Answer Matching
Q: What is the price of the bananas per kg?
A: $11.98
Q: What does the red sign say?
A: Stop
ST-VQA[180]
Evaluation: Accuracy
Metric: Exact Match
Format: Specific short-form answers, such as objects…
Me: I'll do it at 8
Time: 8.05
Me: looks like I gotta wait till 9 now
Q: Can you explain this meme?
GT: This meme is a humorous take on procrastination and the tendency to delay tasks until a specific time …
MM-Vet[123]
Evaluation: Average Score
Metric: ROUGE, LLM Eval
Format: Long-form open-ended answers
Prompt: Is the right orange circle the same size as the left orange circle?
Original
Answer: Yes, the orange balls have the same size.
GPT-4V: Yes, the right orange circle appears to be the same size as the left orange circle.
LLaVA-1.5: No, the right orange circle is smaller than the left orange circle.
Edited: The orange ball on the right is enlarged.
Answer: No, the orange balls have different size.
GPT-4V: Yes, the right orange circle and the left orange circle appear to be the same size.
LLaVA-1.5: Yes, the right orange circle is the same size as the left orange circle.
HallusionBench[63]
Evaluation: Accuracy / Precision / Recall
Metric: Exact Match
Format: Yes/No question
2) Multiple Choice
Art & Design
Question: Among the following harmonic intervals, which one is constructed incorrectly?
Options:
(A) Major third <image 1>
(B) Diminished fifth <image 2>
(C) Minor seventh <image 3>
(D) Diminished sixth <image 4>
Subject: Music; Subfield: Music;
Image Type: Sheet Music;
Difficulty: Medium
MMMU-Pro[45]
Evaluation: Accuracy
Metric: Exact Choice Match
Format: Multiple choice questions
3) Image-Caption Similarity
a brown bear? → 0.9925
a blue boat? → 0.9878
Score: 0.9804
"A brown bear and a blue boat"
T2I-CompBench[270]
Evaluation: Average Similarity
Metric: CLIPScore, GenEval
Format: text to image generation

Figure 3. Common benchmark evaluation metrics restrict the formats of most benchmarks, which mostly evaluates whether a VLM can generate a short-form answer that matches the correct answers.

Benchmark	Evaluation	Category	Annotation	Size (K)
MMTBench	Multiple Choice	Visual reasoning	AI Experts	30.1
MM-Vet	LLM Eval	Visual reasoning	Human	0.2
MM-En/CN	Multiple Choice	Visual reasoning / Multilingual understanding	Human	3.2
GQA	Answer Matching	Visual reasoning	Seed with Synthetic	22,000
VCR	Multiple Choice	Visual reasoning,	MTurks	290
VQAv2	Answer Matching Yes/No	Chart graphic understanding	MTurks	1,100
MMMU	Answer Matching Multiple Choice	Chart graphic understanding	College Students	11.5
SEEDBench	Multiple Choice	Visual reasoning	Synthetic	19
RealWorld QA	Multiple Choice	Visual reasoning, understanding, recognition, and question answering	Human	0.765
MMMU-Pro	Multiple Choice	Visual reasoning	Human	3.64
DPG-Bench	Semantic Alignment	Visual reasoning / Text-to-Image generation	Synthetic	1.06
MSCOCO-30K	BLEU Rouge Similarity	Visual reasoning / Text-to-Image generation	MTurks	30
TextVQA	Answer Matching	Visual text understanding	CrowdSource	45
DocVQA	Answermatching	Visual text understanding	CrowdSource	50
CMMLU	Multiple Choice	Multilingual multi-modal understanding	College Students	11.5
C-Eval	Multiple Choice	Multilingual multi-modal understanding	Human	13.9
TextVQA	Answer Matching	Visual text understanding	Expert Human	28.6

Benchmark	Evaluation	Category	Annotation	Size (K)
MathVista	Answer Matching Multiple Choice	Visual math reasoning	Human	6.15
MathVision	Answer Matching Multiple Choice	Visual math reasoning	College Students	3.04
OCRBench	Answer Matching	OCR	Human	1
MME	Yes/No	OCR	Human	2.8
InfographicVQA	Answer Matching	Chart graphic understanding	CrowdSource	30
AI2D	Answer Matching	Chart graphic understanding	CrowdSource	1
ChartQA	Answer Matching	Chart graphic understanding	CrowdSource and synthetic	32.7
GenEval	CLIPScore GenEval	Text-to-Image generation	MTurks	1.2
T2I-CompBench	Multiple Metrics	Text-to-Image generation	Synthetic	6
HallusionBench	Yes/No	Hallucination	Human	1.13
POPE	Yes/No	Hallucination	Human	9
MMLU	Multiple Choice	Multimodal general intelligence	Human	15.9
MMStar	Multiple Choice	Multimodal general intelligence	Human	1.5
M3GIA	Multiple Choice	Multimodal general intelligence	Human	1.8
InternetAGIEval	Multiple Choice	Multimodal general intelligence	Human	8.06
EgoSchem	Multiple Choice	Video understanding	Synthetic/Human	5
MVBench	Multiple Choice	Video understanding	Synthetic/Human	4
MLVU	Multiple Choice	Video understanding	Synthetic/Human	2.6
VideoMME	Multiple Choice	Video understanding	Experts	2.7
Perception-Test	Multiple Choice	Video understanding	CrowdSource	11.6
VQAScore	Yes/No	Vision-Language Alignment	AI	665

Benchmark	Evaluation	Category	Annotation	Size (K)
GenAI-Bench	Human Ratings	Generative AI Evaluation	Human	80.0
NaturalBench	Yes/No Multiple Choice	Vision-Language Adversarial Testing	Human	10.0

Table 3. Benchmarks and evaluations, along with their annotation and data source.

Answer matching

is widely used for open-ended and closed-ended question types, which are the answers are *short-form entities, long-form answers, numbers, or yes/no*. Generative vlms are more verbose than extractive llms and vlms, where they often generate verbose but correct answers[168], containment exact match[169] is a more practical version used more often in the evaluation, which includes removing articles and space of predicted answers and check whether the normalized predicted answer is contained in the normalized gold answer[170][171]. However, exact match tends to have high recall, which often fails to account for semantic equivalence between the gold and predicted answers, frequently misjudging human-acceptable correct answers as incorrect[172][173][168] and becomes impossible for benchmarks that seek long-form answers[174]. Prior to the instruction following success of llm period, standard token overlapping socres such as F_1, ROUGE[175], BLEU[176] to measure the similarity score between the gold and predicted answers, but start failing when generative models are generating more complex and diverse but correct answers[174][173][168][172].

Thus, some of the benchmarks like MM-Vet[124] adopts llms to evaluate generated responses when the responses are long-form answers that requires semantic understanding to judge correctness. llm evaluations are shown to have the highest correlations to human evaluation, but they also face the struggles of producing consistent outputs with internal model updates or changing prompt instructions[177][178][179]. While no current answer-matching evaluation method is perfect, yes/no questions are the easiest to evaluate compared to open-ended ones. As a result, most benchmarks rely on a multiple-choice format to assess vlms (Figure 2b).

Multiple Choice

format involves selecting an answer from a set of options, including distractors, for a given visual question[127][148][149][150]. This format provides definitive answers and is among the easiest to evaluate, as it measures the percentage of questions a vlm answers correctly. However, llms have demonstrated an unusual ability to select correct answers even without access to the actual questions[180]. Since vlms incorporate an llm component for generating responses (Section 3), further research is required to assess the robustness and reliability of current vlm benchmarks.

Image/text similarity scores

are commonly used in image generation benchmarks like T2I-CompBench, GenEval[133][132] to evaluate the alignment between generated images and their corresponding textual descriptions. They often rely on measures such as CLIPScore[181] for image-text alignment or ROUGE for caption matching to assess the semantic and lexical similarity between the outputs and the references.

In summary, vlm benchmarks encompass a wide range of question types, fields of expertise, and tasks, with MMLU[138] alone covering 57 distinct tasks. However, popular evaluations remain largely confined to simple answer matching or multiple choice formats, far from the broader definition of general intelligence of the Turing test[182].

5. Applications

VLMs are adopted to a wide variety of tasks, from virtual world applications such as virtual embodied agents to real world applications such as robotics and autonomous driving.

5.1. Embodied VLM Agents

Visual question answering (VQA) is a foundational task that involves answering questions based on visual and textual content[19]. It requires extracting meaningful information from images or video sequences, such as identifying objects, scenes, and activities. In practice, embodied VLM agents[183] is a popular application of VQA, ranging from embodied personal device chatbot assistance to visual chart interpretation and diagram generation for low-vision users[184][185].

Embodied agents are AI models with virtual or physical bodies that can interact with their environment[186]. Pure textual agents such as Apple Intelligence[187] can process, reason, and execute user requests by converting them to executable code to control phone applications, but lacks visual reasoning abilities. In this context, we focus specifically on embodied agents with virtual bodies, particularly in relation to the application of VQA models for personal assistance and accessibility.

Embodied VLM agents as assistive applications and accessibility aims at helping users perform actions on devices or providing on-screen answers to assist individuals with low vision. Recent developments include: ScreenAI[188] specializes in understanding user interface (UI) components and answering questions about screen elements. Smartphone assistant[189] extends this capability by using an end-to-end VLM that directly reads visual screen inputs and user requests and converts into executable code sequences to fulfill user request actions. Similar to Smartphone assistant, ScreenAgent[190] uses a three-step approach (planning, acting, reflecting) to process user requests. It first understands UI components through natural language descriptions, then decomposes user requests into subtasks, and finally generates mouse and keyboard operations in a function-call format to execute actions on user screens. In addition, some of these VLM agents might also require chart understanding or generation capabilities to tell a user what the graphics, diagrams or charts are about. VLMs are often prone to hallucination, especially for chart understanding that often extracts wrong numbers. ChartLLaMA[184] is finetuned specifically for understanding various chart or plot visual inputs with more accuracy number extraction and interpretation. Nonetheless, these VLM applications serve as an assistant to help users automatically execute actions without user involving and help disabled people access and understand UI pages better to improve accessibility[191].

Despite the advancements of embodied virtual VLM agents, there is a limitation of their reliance on language models, often using vision as a supplementary role rather than fully integrating the two modalities[21]. These models often use language reasoning as the primary driver, with visual input playing a secondary role, leading to insufficient visual understanding to inform decision-making effectively.[192][193]. Besides virtual applications, embodied VLMagents are also used to perform real physical world applications such as surgical planning and simulation to reduce risks[194], with more physical details in Section 5.3.

5.2. Generative Visual Media Applications

Generative **vlm** models, including generative adversarial networks (gan)[195], diffusion models[196], and newer frameworks like Transfusion are widely used in media applications to aid art and content creations. One notable application of generative **vlm** models is in the creation of memes, a universal language of the Internet. Platforms like Supermeme.ai[197] uses **vlm** models to generate customized memes in over 110 languages, enabling users to express emotions or ideas effectively through humorous or relatable visual content. In addition, generative **vlm** models are used in cinematic and visual effects. For instance, MovieGen[198] allows users to create dynamic movie scenes by transforming static images into visually stunning video effects based on user input.

5.3. Robotics and Embodied AI

The integration of vision-language models with robotics is a very heated topic that bridges the foundation models residing in cyberspace and the physical world[199]. An enormous amount of research work has emerged in the last few years, focusing on using **vlm**s' abilities on visual reasoning[200][201], complicated scene understanding[202][203], planning[204][205] over various tasks across manipulation[162][206], navigation[207][208][209], human-robot interaction[210][211], multi-robot coordination[212][213], motion planning[214][215], reward function design[216][217][218], etc. The revolutionary development in this area triggers many unexplored research problems that gather much attention from the robotics community, while also revealing many hidden limitations during implementation (Section 5.3.5).

5.3.1. Manipulation

The application of **vlm** in robot manipulation tasks focuses on improving robots' abilities to manipulate out-of-domain objects or perform more demanding, expensive action planning using their language priors. VIMA[162] designs a transformer-based robot agent that processes these prompts and outputs motor actions autoregressive. Instruct2Act[206] uses an **llm** model to generate Python programs that constitute a comprehensive perception, planning, and action loop for robotic tasks. RoboVQA[219] proposes an approach for the efficient collection of robotics data, with a large and diverse dataset for robotics visual question answering and a single model with embodied reasoning. Robotool[220] proposes a system developed to enable robots to employ creative tools use through the integration of foundation models. The RT series[221][222][223] purpose a vision-language action model that encodes visual observations and text prompts and computes the target positions and orientation for robot manipulation tasks. Though the current **vlm** applications in robotics show impressive abilities in visual reasoning and scene understanding in manipulation tasks, their abilities are still constrained by their generalization levels, given the diversity of the robot manipulators.

5.3.2. Navigation

The incorporation of VLMs in robot navigation tasks focuses on open-world zero-shot or few-shot object-goal navigation or semantic cue-driven navigation. ZSON[207] trains agents on image-goal navigation using a multimodal semantic embedding space, enabling zero-shot ObjectNav from natural language instructions and robust generalization to complex, inferred instructions. LOC-ZSON[208] introduces a Language-driven Object-Centric image representation and LLM-based augmentation techniques for zero-shot object navigation. LM-Nav[224] is a system for robotic navigation that combines pre-trained models to enable natural language-based long-horizon navigation in real-world outdoor environments without requiring fine-tuning or language-annotated data. NaVILA[225] proposes a vision-language-action (VLA) model for legged robot navigation under challenging and cluttered scenes. VLFM[209] builds occupancy maps from depth observations to identify frontiers, and leverages RGB observations and a pre-trained vision-language model to generate a language-grounded value map to identify the most promising frontier and explore for finding the given target object. LFG-Nav[226] uses the language model to bias exploration of novel real-world environments by incorporating the semantic knowledge stored in language models as a search heuristic for planning. Many existing works follow the Task and Motion Planning (TAMP) [227] pipeline, a framework convenient in segmenting the entire task into feasible subgoals that are execrable by low-level planners, though its adaptability is constrained by planners and in lack of flexibility in handling unexpected situations.

5.3.3. Human-robot Interaction

Human-robot interaction (HRI) is a sub-field demanding cognition and adaptation, as well as the ability to interpret human intentions in reality and take actions accordingly. vlm-powered HRI has shown much better ability in understanding human intentions and adaptability during interaction. MUTEX[228] is a transformer-based approach for policy learning and human-robot collaboration from multimodal task specifications, enabling robots to interpret and follow tasks across six modalities (video, images, text, and speech). LaMI[229] revolutionizes multi-modal human-robot interaction by enabling intuitive, guidance-driven regulation of robot behavior, dynamically coordinating actions and expressions to assist humans while simplifying traditional state-and-flow design processes. Wang et al.[211] designs a pipeline that uses vlms to interpret human demonstration videos and generate robot task plans by integrating keyframe selection, visual perception, and VLM reasoning, demonstrating superior performance on long-horizon pick-and-place tasks across diverse categories. vlm-Social-Nav[210] leverages Vision-Language Models to enable socially compliant navigation by detecting social entities and guiding robot actions in human-centered environments.

5.3.4. Autonomous Driving

Autonomous driving is a very intensive research area in robotics, while the long-tail corner cases covering out-of-domain objects and traffic events have been a long-lasting problem in this field. The on-board vlm agents for autonomous driving have revealed abilities to overcome both problems with the better abilities in object recognition[230][231], navigation and planning[232][233], and decision-making[234][235]. VLPD[230] leverages self-supervised segmentation and contrastive learning to model explicit semantic contexts like small or occluded pedestrians without additional annotations. MotionLM[214] reframes multi-agent motion prediction as a language modeling task by using discrete motion tokens and autoregressive decoding, enabling efficient and temporally causal joint trajectory forecasting. DiLU[236] combines reasoning and reflection modules to enable the system to perform decision-making based on common-sense knowledge and evolve continuously in traffic. Recently, more efforts have been made towards end-to-end autonomous driving models that produce actions from vlms without generating intermediate tasks. VLP[237] introduces a Vision-Language-Planning framework that integrates language models to enhance reasoning, contextual understanding, and generalization in autonomous driving. DriveGPT4[238] proposes the first interpretable end-to-end autonomous driving system leveraging multimodal large language models, capable of processing video inputs, textual queries, and predicting vehicle control signals.

5.3.5. Limitations

Despite the success of vlms' applications in virtual agents, robotics, and autonomous driving, they still face several limitations.

1. **Generalization vs. Flexibility.** Many existing works depend on the TAMP[227] pipeline that uses the vision-language methods to procedure programming code-like workflows[206][217] constructed by pre-defined executable modules, or produce waypoints for external low-level planners to execute actions. Such a pipeline allows efficient modulized robot action

planning, but its upper-bound is constrained by the scope of available executable modules or low-level planners that are vulnerable to out-of-domain (OOD) scenarios. On the other hand, many efforts[214][215] have been made to tokenize the robot's motions as language-like tokens, and outputs the low-level actionable trajectories directly. Such methods, though reconciling with the nature of robot planning, their abilities are highly constrained by the robots models or datasets encountered in their training procedure, which could be highly diverse in the real world.

2. **Intelligence vs. Safety.** Though the applications of vlms improves the abilities of robots, but they also introduce potential risks that may not be encountered before in robotics research. Risks may be inherited from the jail-breaking[239] and biases of vlms[240][241], and robot malfunctioning when executing vlm-determined actions[242] when applying robot-specific attacks or performing reward-hacking. These risks must gather more attraction in revealing and resolving as robots have the access to the physical worlds that could perceive uncensored information in their routine, incorporate it into their internet-level databases, and execute those hazardous actions.
3. **Embodiment vs. Effectiveness.** The current difference in the developing trends of general-purpose vlms and micro-electronics enlarges the gap between the two, which introduces the trade-off issues in embodying the state-of-the-art models onboard and the computational constraints for robots. Many prior works use relatively old models like CLIP, BLIP, ViT or other small vlms for fine-tuning, or merely use the inference functions of close-source vlm like GPT-4v and Gemini[202][206][205], without adaptation to domain-specific data. More discussions are expected in applying large-scale vlms in robotics to show their abilities to enhance the robots' performance with language priors of Large vlms.

5.4. Human-Centered AI

One important and promising application of vlms is to use their understanding and reasoning abilities for human intentions and behaviors during human interaction with AI agents. LVLMs help to perform sentiment analysis[243], predict human intentions[244], and assist human interaction with the real world[245] across many applications for social goodness like AI4Science[246][247], agriculture[248], education[249][250], accessibility[251][252], healthcare[253][254], climate change[255], etc. VLMs show impressive potential in all these fields and help the widespread AI revolutions have a broad impact on every corner of society.

doi.org/10.32388/GXR68Q

5.4.1. *Web Agents*

Web agent[256] is designed to assist human's daily interaction and activities on the webpages. Empowered by VLMs, web agents show enhanced abilities in understanding human behaviors and better adaptation and generalization abilities for human assistance. CogAgent[257] excels in GUI understanding and navigation by utilizing high-resolution image encoding. WebVoyager[258] demonstrates complete user instructions end-to-end by interacting with real-world websites by leveraging multimodal understanding abilities. ShowUI[259] introduces UI-Guided visual token selection to reduce computational costs, interleaved Vision-Language-Action streaming for flexible task handling, and curated GUI instruction-following datasets. ScreenAgent[260] is a vlm that utilizes a planning-acting-reflecting control pipeline. This VLM agent is trained to interact with real computer screens by observing screenshots and executing GUI actions.

5.4.2. *Accessibility*

Accessibility intends to help those disabilities living more conveniently, while VLMs help to interpret the visual contexts to those with vision impairment during their interaction with the webpages and the physical world. X-World[252] is an accessibility-focused environment generating annotated simulation data with dynamic agents using mobility aids, enabling analysis of challenges like occlusion and interaction. Oliveira et al.[251] explores using Multimodal Large Language Models (MLLMs) to generate high-quality text descriptions for 360 VR scenes based on Speech-to-Text prompts, enhancing accessibility and dynamic experiences, as demonstrated in educational VR museum settings. Mohanbabu et al.[261]introduces a Chrome Extension that incorporates webpage context into GPT-4V-generated image descriptions, showing that context-aware descriptions significantly enhance quality, imaginability, relevance, and plausibility.

5.4.3. *Healthcare*

AI for Healthcare is a sub-field that requires much expertise knowledge in information interpretation and very demanding in the accuracy level, due to the severe outcomes. During the recent few years, given the rapid development of LVLMs, AI for healthcare has been increasingly investigated with many exciting breakthroughs, helping it become much more practical in the real-world applications. VisionUnite[254] introduces a vision-language foundation model pretrained on extensive ophthalmology datasets, exceling in multi-disease diagnosis, clinical explanations, and patient interactions. Yildirim et al.[253] explores the clinical utility of vlms in radiology through various clinical applications, revealing high potential from assessment from multiple radiologists and clinicians. M-FLAG[262] presents a novel method for pre-training medical vision-language models, utilizing a frozen language model for efficiency and an orthogonality loss to optimize latent space geometry with significantly fewer parameters and exceptional performance even on limited data. Medclip[263] introduces a decoupled approach to multimodal contrastive learning, scaling training data combinatorially and addressing false negatives with a semantic matching loss based on medical knowledge. Med-Flamingo[264] introduces a multimodal few-shot learner adapted to the medical domain enabling few-shot adaptations like rationale generation and excelling in clinician-reviewed evaluations on challenging datasets.

5.4.4. Social Goodness

The strong abilities of vlms help a wide range of applications for social goodness. In K-12 education, the recent works help to reason mathematically over educational content using VLMs[250], or simulate students with various personalities to improve teachers' teaching skills[249]. VLMs help to diagnosis disease for plants[248] and optimize the utilization of farmlands[265] in agriculture applications. VLMs are also used for promoting fundamental science research like chemistry[266], mathematics[267][268], etc., and for other impactful field like climate change[255], mitigating social biases[269], and urban planning[270].

6. Challenges

This section focuses on efforts on 3 challenging areas in vlm evaluation: hallucination, safety, and fairness. While recent improvements have enabled vlms to continuously attain SOTA performance (subsubsection 5.4.4), understanding the risks from their misapplication is paramount to assess and prevent harms to end users, especially those who belong in marginalized groups. The following discussion serves to highlight current limitations and ongoing research to ensure the reliable and ethical use of vlms.

6.1. Hallucination

Hallucination refers to the vlm's tendency to refer to objects and/or artifacts that do not appear in the relevant image[271]. Despite benchmark-setting performance, hallucination is still a pervasive issue especially in visual-text application tasks. Researchers have proposed datasets and metrics to quantify hallucination, with early efforts tend to require human annotation. For image-captioning, Rorhrbach et el.[271] proposed CHAIR, a metric that calculated the proportion of words generated that appeared in the image based on ground-truth captions. CHAIR consists of 2 variants: per-instance, which measures the fraction of hallucinated instances, and per-sentence, which measures the fraction of sentences that include a hallucinated object. Li et al.[272] developed POPE, which assessed the amount of hallucination via a series of Yes-No questions about existent and non-existent objects given an image. Gunjal et al.[273] released M-HalDetect, a fine-grained annotated dataset of 16,000 samples on visual QA that can be used to train vlms to detect and prevent hallucination.

Subsequent research investigated hallucination in finer details. Halle-Switch evaluates hallucination from the perspective of data amount, quality and granularity; which incorporates both contextual and parametric knowledge to control hallucination rather than outright removal[274]. Hallu-Pi[275] contains 1260 images of 11 object types with detailed annotation to detect various hallucination types that occur in perturbed input.[276] focuses on before-and-after changes to image while proposing new metrics to analyze hallucination: true understanding, ignorance, stubbornness, indecision. Guan et al.[21] proposed HallusionBench to investigate vlm's visual reasoning via dependent questions that have no affirmative answers without visual content to on diverse topics (e.g.: food, math, meme) and image formats (e.g.: logo, poster, chart) to detect hallucination. [277] develops an automatic benchmark generation approach that harnesses a few principal strategies to create diverse hallucination examples by probing the language modules in vlms for context cues.

The advent of more sophisticated llms has also assisted the development of larger benchmark datasets in this area. GAIVE[278] uses GPT-4 to generate 400,000 samples in the form of open-ended instruction that covers 16 vision-and-language tasks. They account for various semantic levels of hallucination, such as nonexistent object manipulation and knowledge manipulation[278]. Jiang et al. [279] constructed Hal-Eval using GPT-4 to induce fine-grained hallucination and tailored prompts for 2 million image-caption sample pairs. On the other hand, AMBER[280] is an llm-free multi-dimensional benchmark designed for both generative and discriminative tasks with annotation for 4 types of hallucination.

6.2. Safety

Due to vlms' tremendous versatility, it becomes even more important to safeguard them against unethical and harmful usage. Malicious actors may utilize vlms to deleterious effects by jailbreaking, defined as "deliberately circumventing the ethical and operational boundaries" of the models[281], which might be harmful for both vlms and their applications in downstream tasks like robotics[282][200][242]. Yinbg et al.[283] SafeBench, a dataset of harmful queries on 23 risks scenarios generated by llms, along with a jury deliberation protocol using multiple llm collaborative framework. Similarly, MM-Safetybench is another benchmark dataset that uses queries of images paired with malicious texts to assess vlms' behaviors in unsafe scenarios.

Luo et al.[284] released JailbreakV with 28,000 malicious queries as image-based attack that vlms should not respond. This dataset also enables the verification of transferability between models of jailbreak attacks. Shi et al.[285] developed SHIELD, which uses True-False queries to evaluate vlms' performance on face spoofing and forgery detection in zero- and few-shot settings. Other research investigate attacks that can reverse prior efforts to align models towards ethical use. For instance, HADES by Li et al.[286] exploits gradient updates and adversarial methods to hide and amplify image-based harmfulness and destroy multimodal alignment. Niu et al.[287] proposed imgJP, which uses specific image instead of prompts to bypass refusal guardrails. imgJP has been shown to be highly transferrable across a wide range of vlms[287].

6.3. Fairness

Extensive literature has discussed the inequity propagated by llms and vlms[288][289]. Similar to their unimodal llm counterpart, vlms have exhibited disparate performance in downstream applications particularly towards certain marginalized groups[290][240][241]. Janghorbani and Gerard[291] introduced MMBias, an human-annotated datasets of images based on target concepts (religion, nationality, disability, sexual orientation) with a dichotomous grouping on pleasantness. Wu et al.[292] proposed FMBench, a framework that uses annotated medical images for both direct and single-choice visal QA to measure bias with respect to gender, skin tone and age. Also in the medical domain, Luo et al.[293] released Harvard-FairVL, a dataset of SLO fundus images paired with clinical notes with demographic attributes. Empirical results on CLIP and BLIP2 show preference Asian, Male, Non-Hispanic groups compared to other attributes[293]. Jin et al.'s FairmedFM integrates 17 medical image datasets to evaluate fairness on classification and segmentation on downstream tasks[294]. In other veins, CulturalVQA by Nayak et al.[295] was constructed with 2,378 image-question pairs with multiple human-annotated answers per question drawn from different cultures, with results showing better performance for North American cultures and worse on African and Islamic ones.

6.4. Multi-modality Alignment

The alignment issue within the multi-modality models refers to the contextual deviation between the different modalities. The mis-alignment issue of vlms may cause hallucinations[296]. Many efforts have been made to mitigate this issue by either utlizting the reasoning abilities of vlms to perform self-reflection[297] or designing projectors to bridge over different modalities. SIMA[297] enhances alignment between visual and language modalities in large vision-language models (Lvlms) through self-improvement, using self-generated responses and an in-context self-critic mechanism with vision metrics. SAIL[298] introduces an efficient transfer learning framework that aligns pretrained unimodal vision and language models for vision-language tasks, enhancing the language-compatibility of vision encoders to improve multimodal large language models. Ex-MCR[299] introduces a training-efficient, paired-data-free approach to multi-modal contrastive representation (MCR) by extending one modality's space into another, enabling emergent semantic alignment between extended modalities. OneLLM[300] is a unified Multimodal Large Language Model (mllm) that aligns eight modalities to language through a unified encoder and progressive multimodal alignment.

6.5. Efficient Training and Fine-Tuning

The efficient training and fine-tuning for vision language models has been a very heated research topic, as the current large-scale vlms are hard and expensive to train. An increasing number of recent works draw their attentions onto the pre-training procedure of vision language model that tries to understand the effect of different settings over modules[301] or supervision on the ultimate performance of vlms. Meanwhile, specific purposes that require the application of vlms do not necessarily require the versatile multi-task performance of vlms, but outstanding one or two expertise of tasks. Typically, the Low-Rank Adaptation (LoRa) methods[302][303] helps to manipulate the Lvlms by changing fewer parameters and lowering the computational resources. Methods like reinforcement learning with human or AI feedback (RLHF)[304][305] are also widely used in fine-tuning vlms by integrating human or other Lvlms' knowledge into the fine-tuning procedure.

6.6. Scarce of High-quality Dataset

The abilities and reliabilities of VLMs are highly depending on the availability and diversity of the training datasets. However, the massive scale of current advanced VLMs and the scarce of high-quality training datasets add up to the difficulty in continuously improving the performance of the future VLMs. One potential method to mitigate this issue is to use self-supervised learning (SSL) [306] that learns the representation automatically from the unlabelled dataset. Another major direction is to use the synthetic data generated by following some rules[307] or utilizing some third-party tools[308]. In VLM specifically designed for physical world-related purposes, like robotics[309] or web agents[310], another option is to gather datasets from the interactions with the physical simulators or world model. Though a lot of efforts have been made in all three directions, more insights are still expected into the breakthrough of the mass-scale training for LVLMs and the alternatives to the internet-scale data, given Ilya Sutskever's quote that "Pre-training as we know it will unquestionably end."

7. Conclusion

Developments of VLMs and LLMS are happening at a breakneck pace with more sophisticated applications and use cases being introduced in quick succession. This paper aims to capture the most notable architectures, tends, applications along with prominent challenges in this area. We hope that our survey provides a solid general overview for practitioners as a road map for future works.

Footnotes

[1] https://github.com/ai-forever/MoVQGAN

References

1. [a, b, c, d, e, f, g]*Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021). "Learning Transferable Visual Models From Natural Language Supervision". arXiv. arXiv:2103.00020 [cs.CV].*
2. [a, b]*Anthropic. Claude: An AI Assistant by Anthropic, 2024. Accessed: 2024-12-23. Available from: https://www.anthropic.com/claude.*
3. [a, b, c]*Yang Z, Li L, Lin K, Wang J, Lin C, Liu Z, Wang L (2023). "The dawn of lmms: Preliminary explorations with gpt-4v (ision)". arXiv preprint arXiv:2309.17421.* **9** *(1): 1.*
4. [^]*Islam A, Biswas MR, Zaghouani W, Belhaouari SB, Shah Z (2023). "Pushing Boundaries: Exploring Zero Shot Object Classification with Large Multimodal Models". arXiv. arXiv:2401.00127 [cs.CV].*
5. [a, b, c, d, e]*Touvron H, Martin L, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. Available from: https://arxiv.org/abs/2307.09288.*
6. [a, b, c]*OpenAI (2024). "GPT-4 Technical Report". arXiv. Available from: https://arxiv.org/abs/2303.08774.*
7. [^]*Makridakis S, Petropoulos F, Kang Y (2023). "Large language models: Their success and impact". Forecasting.* **5** *(3): 536–549.*
8. [^]*Mohammad AF, Clark B, Hegde R. Large Language Model (LLM) & GPT, A Monolithic Study in Generative AI. In: 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE). IEEE; 2023. p. 383–388.*
9. [^]*Nayab S, Rossolini G, Buttazzo G, Manes N, Giacomelli F (2024). "Concise thoughts: Impact of output length on llm reasoning and cost". arXiv preprint arXiv:2407.19825.*
10. [^]*Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. "Position: Will we run out of data? Limits of LLM scaling based on human-generated data." In: Forty-first International Conference on Machine Learning.*
11. [^]*Li Y, Ren S, Deng W, Xu Y, Gao Y, Ngai E, Wang H (2024). "Beyond Finite Data: Towards Data-free Out-of-distribution Generalization via Extrapola". arXiv preprint arXiv:2403.05523. Available from: https://arxiv.org/abs/2403.05523.*
12. [^]*Goodwin C, Bjørndahl JS (2018). "Why multimodality? Why co-operative action?(transcribed by J. Philipsen)". Social Interaction. Video-Based Studies of Human Sociality.* **1** *(2).*

13. ^Hong H, Wang S, Huang Z, Wu Q, Liu J (2024). "Why Only Text: Empowering Vision-and-Language Navigation with Multi-modal Prompts". arXiv preprint arXiv:2406.02208. Available from: https://arxiv.org/abs/2406.02208.

14. ^Bordes F, Pang RY, Ajay A, Li AC, Bardes A, Petryk S, Mañas O, Lin Z, Mahmoud A, Jayaraman B, et al. An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247. 2024.

15. ^Hartsock I, Rasool G (2024). "Vision-language models for medical report generation and visual question answering: A review". Frontiers in Artificial Intelligence. 7: 1430984.

16. ^Doveh S, Perek S, Mirza MJ, Lin W, Alfassy A, Arbelle A, Ullman S, Karlinsky L (2024). "Towards multimodal in-context learning for vision & language models". arXiv preprint arXiv:2403.12736.

17. ^Wang F, Ding L, Rao J, Liu Y, Shen L, Ding C (2024). "Can linguistic knowledge improve multimodal alignment in vision-language pretraining?" ACM Transactions on Multimedia Computing, Communications and Applications. **20** (12): 1–22.

18. a, bLymperaiou M, Stamou G (2024). "A survey on knowledge-enhanced multimodal learning". Artificial Intelligence Review. **57** (10): 284.

19. a, bAgrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Batra D, Parikh D (2016). "VQA: Visual Question Answering". arXiv. Available from: https://arxiv.org/abs/1505.00468.

20. ^Tian X, Gu J, Li B, Liu Y, Wang Y, Zhao Z, Zhan K, Jia P, Lang X, Zhao H (2024). "DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models". arXiv. arXiv:2402.12289 [cs.CV].

21. a, b, c, d, eGuan 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024:14375-14385.

22. ^Liu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, Hou L, Li R, Peng W (2024). "A Survey on Hallucination in Large Vision-Language Models". arXiv. Available from: https://arxiv.org/abs/2402.00253.

23. ^Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N, Mian A (2023). "A comprehensive overview of large language models". arXiv preprint arXiv:2307.06435.

24. ^Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. **15**(3):1-45, 2024.

25. a, bLiu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, Hou L, Li R, Peng W (2024). "A survey on hallucination in large vision-language models". arXiv preprint arXiv:2402.00253.

26. ^Radford A, Kim JW, 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. In: International conference on machine learning. PMLR; 2021. p. 8748-8763.

27. a, bLi J, Li D, Xiong C, Hoi S. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In: International conference on machine learning. PMLR; 2022. p. 12888–12900.

28. a, b, c, d, eAlayrac JB, 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. 35:23716–23736, 2022.

29. a, bAnil R, Borgeaud S, Wu Y, Alayrac JB, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. 1, 2023.

30. a, bLi LH, Yatskar M, Yin D, Hsieh CJ, Chang KW (2019). "VisualBERT: A Simple and Performant Baseline for Vision and Language". arXiv. arXiv:1908.03557 [cs.CV].

31. a, b, c, d, e, f, gLin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P (2015). "Microsoft COCO: Common Objects in Context". arXiv. arXiv:1405.0312 [cs.CV].

32. ^Ren S, He K, Girshick R, Sun J (2016). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". arXiv. arXiv:1506.01497 [cs.CV].

33. a, b, c, d, e, f, g, h, i, j, k, l, mDosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". arXiv. arXiv:2010.11929 [cs.CV].

34. a, bHe K, Zhang X, Ren S, Sun J (2015). "Deep Residual Learning for Image Recognition". arXiv. arXiv:1512.03385 [cs.CV].

35. a, bKrishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA, Bernstein MS, Li F-F (2016). "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". arXiv. Available from: https://arxiv.org/abs/1602.07332.

36. a, b, cJia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung Y, Li Z, Duerig T (2021). "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision". arXiv. arXiv:2102.05918 [cs.CV].

37. ^Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks LA, Welbl J, Clark A, Hennigan T, Noland E, Millican K, van den Driessche G, Damoc B, Guy A, Osindero S, Si

monyan K, Elsen E, Rae JW, Vinyals O, Sifre L (2022). "Training Compute-Optimal Large Language Models". arXiv. Available from: https://arxiv.org/abs/2203.15556.

38. ^Li J, Li D, Savarese S, Hoi S (2023). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". arXiv. arXiv:2301.12597 [cs.CV].
39. ^Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin XV, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura PS, Sridhar A, Wang T, Zettlemoyer L (2022). "OPT: Open Pre-trained Transformer Language Models". arXiv. arXiv:2205.01068.
40. ^Liu H, Li C, Li Y, Lee YJ (2024). "Improved Baselines with Visual Instruction Tuning". arXiv. arXiv:2310.03744 [cs.CV].
41. a, b, cVicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality [Internet]. 2023 [cited 2024 Dec 23]. Available from: https://vicuna.lmsys.org/.
42. ^Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I, Florence P (2023). "PaLM-E: An Embodied Multimodal Language Model". arXiv. arXiv:2303.03378 [cs.LG].
43. ^Chen X, Wang X, Changpinyo S, Piergiovanni AJ, Padlewski P, Salz D, Goodman S, Grycner A, Mustafa B, Beyer L, Kolesnikov A, Puigcerver J, Ding N, Rong K, Akbari H, Mishra G, Xue L, Thapliyal A, Bradbury J, Kuo W, Seyedhosseini M, Jia C, Karagol Ayan B, Riquelme C, Steiner A, Angelova A, Zhai X, Houlsby N, Soricut R (2023). "PaLI: A Jointly-Scaled Multilingual Language-Image Model". arXiv. arXiv:2209.06794.
44. ^Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, Schuh P, Shi K, Tsvyashchenko S, Maynez J, Rao A, Barnes P, Tay Y, Shazeer N, Prabhakaran V, Reif E, Du N, Hutchinson B, Pope R, Bradbury J, Austin J, Isard M, Gur-Ari G, Yin P, Duke T, Levskaya A, Ghemawat S, Dev S, Michalewski H, Garcia X, Misra V, Robinson K, Fedus L, Zhou D, Ippolito D, Luan D, Lim H, Zoph B, Spiridonov A, Sepassi R, Dohan D, Agrawal S, Omernick M, Dai AM, Pillai TS, Pellat M, Lewkowycz A, Moreira E, Child R, Polozov O, Lee K, Zhou Z, Wang X, Saeta B, Diaz M, Firat O, Catasta M, Wei J, Meier-Hellstern K, Eck D, Dean J, Petrov S, Fiedel N (2022). "PaLM: Scaling Language Modeling with Pathways". arXiv. Available from: https://arxiv.org/abs/2204.02311.
45. ^Wang W, Lv Q, Yu W, Hong W, Qi J, Wang Y, Ji J, Yang Z, Zhao L, Song X, Xu J, Xu B, Li J, Dong Y, Ding M, Tang J (2024). "CogVLM: Visual Expert for Pretrained Language Models". arXiv. Available from: https://arxiv.org/abs/2311.03079.

46. ^Webster R, Rabin J, Simon L, Jurie F (2023). "On the De-duplication of LAION-2B". arXiv. arXiv:2303.12733 [cs.CV].

47. ^Byeon M, Park B, Kim H, Lee S, Baek W, Kim S (2022). "COYO-700M: Image-Text Pair Dataset". Available from: https://github.com/kakaobrain/coyo-dataset.

48. ^Dai W, Li J, Li D, Tiong AMH, Zhao J, Wang W, Li B, Fung P, Hoi S (2023). "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning". arXiv. 2305.06500.

49. a, b, cGoyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017). "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". arXiv. arXiv:1612.00837 [cs.CV].

50. ^Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, Webson A, Gu SS, Dai Z, Suzgun M, Chen X, Chowdhery A, Castro-Ros A, Pellat M, Robinson K, Valter D, Narang S, Mishra G, Yu A, Zhao V, Huang Y, Dai A, Yu H, Petrov S, Chi EH, Dean J, Devlin J, Roberts A, Zhou D, Le QV, Wei J (2022). "Scaling Instruction-Finetuned Language Models". arXiv. arXiv:2210.11416.

51. a, bChen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, Li B, Luo P, Lu T, Qiao Y, Dai J (2024). "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks". arXiv. arXiv:2312.14238 [cs.CV].

52. a, bSchuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R, Jitsev J (2022). "LAION-5B: An open large-scale dataset for training next generation image-text models". arXiv. arXiv:2210.08402 [cs.CV].

53. ^Cui Y, Yang Z, Yao X (2024). "Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca". arXiv. arXiv:2304.08177 [cs.CL].

54. a, bWang X, Zhang X, Luo Z, Sun Q, Cui Y, Wang J, Zhang F, Wang Y, Li Z, Yu Q, Zhao Y, Ao Y, Min X, Li T, Wu B, Zhao B, Zhang B, Wang L, Liu G, He Z, Yang X, Liu J, Lin Y, Huang T, Wang Z (2024). "Emu3: Next-Token Prediction is All You Need". arXiv. Available from: https://arxiv.org/abs/2409.18869.

55. ^Zhang BW, Wang L, Li J, Gu S, Wu X, Zhang Z, Gao B, Ao Y, Liu G (2024). "Aquila2 Technical Report". arXiv. Available from: https://arxiv.org/abs/2408.07410.

56. ^Zheng C, Vuong LT, Cai J, Phung D (2022). "MoVQ: Modulating Quantized Vectors for High-Fidelity Image Generation". arXiv. arXiv:2209.09002 [cs.CV].

57. a, b, cDai W, Lee N, Wang B, Yang Z, Liu Z, Barker J, Rintamaki T, Shoeybi M, Catanzaro B, Ping W (2024). "NVLM: Open Frontier-Class Multimodal LLMs". arXiv. arXiv:2409.11402 [cs.CL].

58. a, b, c *Li J, Li D, Xiong C, Hoi S. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". arXiv [cs.CV]. 2022. Available from: https://arxiv.org/abs/2201.12086.*

59. a, b, c *Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, Li C, Li C, Liu D, Huang F, Dong G, Wei H, Lin H, Tang J, Wang J, Yang J, Tu J, Zhang J, Ma J, Yang J, Xu J, Zhou J, Bai J, He J, Lin J, Dang K, Lu K, Chen K, Yang K, Li M, Xue M, Ni N, Zhang P, Wang P, Peng R, Men R, Gao R, Lin R, Wang S, Bai S, Tan S, Zhu T, Li T, Liu T, Ge W, Deng X, Zhou X, Ren X, Zhang X, Wei X, Ren X, Liu X, Fan Y, Yao Y, Zhang Y, Wan Y, Chu Y, Liu Y, Cui Z, Zhang Z, Guo Z, Fan Z. "Qwen2 Technical Report". arXiv. 2024. Available from: https://arxiv.org/abs/2407.10671.*

60. a, b, c, d *Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, Wang J, Ge W, Fan Y, Dang K, Du M, Ren X, Men R, Liu D, Zhou C, Zhou J, Lin J (2024). "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution". arXiv. 2409.12191.*

61. a, b *Agrawal P, Antoniak S, Bou Hanna E, Bout B, Chaplot D, Chudnovsky J, Costa D, De Monicault B, Garg S, Gervet T, Ghosh S, Héliou A, Jacob P, Jiang AQ, Khandelwal K, Lacroix T, Lample G, Las Casas D, Lavril T, Le Scao T, Lo A, Marshall W, Martin L, Mensch A, Muddireddy P, Nemychnikova V, Pellat M, Von Platen P, Raghuraman N, Rozière B, Sablayrolles A, Saulnier L, Sauvestre R, Shang W, Soletskyi R, Stewart L, Stock P, Studnia J, Subramanian S, Vaze S, Wang T, Yang S (2024). "Pixtral 12B". arXiv. Available from: https://arxiv.org/abs/2410.07073.*

62. ^ *Mistral AI Team. Mistral Large 2: A New Generation of Advanced Language Models. https://www.mistral.ai/. July 2024. Accessed: 2024-12-23.*

63. a, b, c *Grattafiori A, et al. (2024). "The Llama 3 Herd of Models". arXiv. Available from: https://arxiv.org/abs/2407.21783.*

64. a, b *Li Y, Sun H, Lin M, Li T, Dong G, Zhang T, Ding B, Song W, Cheng Z, Huo Y, Chen S, Li X, Pan D, Zhang S, Wu X, Liang Z, Liu J, Zhang T, Lu K, Zhao Y, Shen Y, Yang F, Yu K, Lin T, Xu J, Zhou Z, Chen W (2024). "Ocean-omni: To Understand the World with Omni-modality". arXiv. Available from: https://arxiv.org/abs/2410.08565.*

65. ^ *Yang A, Xiao B, Wang B, Zhang B, Bian C, Yin C, Lv C, Pan D, Wang D, Yan D, Yang F, Deng F, Wang F, Liu F, Ai G, Dong G, Zhao H, Xu H, Sun H, Zhang H, Liu H, Ji J, Xie J, Dai J, Fang K, Su L, Song L, Liu L, Ru L, Ma L, Wang M, Liu M, Lin M, Nie N, Guo P, Sun R, Zhang T, Li T, Li T, Cheng W, Chen W, Zeng X, Wang X, Chen X, Men X, Yu X, Pan X, Shen Y, Wang Y, Li Y, Jiang Y, Gao Y, Zhang Y, Zhou Z, Wu Z (2023). "Ba*

ichuan 2: Open Large-scale Language Models". arXiv. Available from: https://arxiv.org/abs/2309.10305.

66. ^a, b^Zhou C, Yu L, Babu A, Tirumala K, Yasunaga M, Shamis L, Kahn J, Ma X, Zettlemoyer L, Levy O (2024). "Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model". arXiv. arXiv:2408.11039 [cs.AI].
67. ^Kingma DP, Welling M (2022). "Auto-Encoding Variational Bayes". arXiv. arXiv:1312.6114 [stat.ML].
68. ^Wu Z, Chen X, Pan Z, Liu X, Liu W, Dai D, Gao H, Ma Y, Wu C, Wang B, Xie Z, Wu Y, Hu K, Wang J, Sun Y, Li Y, Piao Y, Guan K, Liu A, Xie X, You Y, Dong K, Yu X, Zhang H, Zhao L, Wang Y, Ruan C. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv [Internet]. 2024 [cited 2024]. Available from: https://arxiv.org/abs/2412.10302.
69. ^Srinivasan K, Raman K, Chen J, Bendersky M, Najork M. "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning." In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21. ACM; 2021. p. 2443–2449. doi:10.1145/3404835.3463257.
70. ^Koupaee M, Wang WY (2018). "WikiHow: A Large Scale Text Summarization Dataset". arXiv. arXiv:1810.09305 [cs.CL].
71. ^Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R (2023). "Segment Anything". arXiv. arXiv:2304.02643 [cs.CV].
72. ^Zhai X, Mustafa B, Kolesnikov A, Beyer L (2023). "Sigmoid Loss for Language Image Pre-Training". arXiv. arXiv:2303.15343 [cs.CV].
73. ^Wang L, Gao H, Zhao C, Sun X, Dai D (2024). "Auxiliary-Loss-Free Load Balancing Strategy for Mixture-of-Experts". arXiv. arXiv:2408.15664 [cs.LG].
74. ^Dai D, Deng C, Zhao C, Xu RX, Gao H, Chen D, Li J, Zeng W, Yu X, Wu Y, Xie Z, Li YK, Huang P, Luo F, Ruan C, Sui Z, Liang W (2024). "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models". arXiv. Available from: https://arxiv.org/abs/2401.06066.
75. ^a, b^Deitke M, Clark C, Lee S, Tripathi R, Yang Y, Park JS, Salehi M, Muennighoff N, Lo K, Soldaini L, Lu J, Anderson T, Bransom E, Ehsani K, Ngo H, Chen Y, Patel A, Yatskar M, Callison-Burch C, Head A, Hendrix R, Bastani F, VanderBilt E, Lambert N, Chou Y, Chheda A, Sparks J, Skjonsberg S, Schmitz M, Sarnat A, Bischoff B, Walsh P, Newell C, Wolters P, Gupta T, Zeng KH, Borchardt J, Groeneveld D, Nam C, Lebrecht S, Wittlif C, Schoenick C, Michel O, Krishna R, Weihs L, Smith NA, Hajishirzi H, Girshick R, Farhadi A, Ke

mbhavi A. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. 2024. Available from: https://arxiv.org/abs/2409.17146.

76. ^*Muennighoff N, Soldaini L, Groeneveld D, Lo K, Morrison J, Min S, Shi W, Walsh P, Tafjord O, Lambert N, Gu Y, Arora S, Bhagia A, Schwenk D, Wadden D, Wettig A, Hui B, Dettmers T, Kiela D, Farhadi A, Smith NA, Koh PW, Singh A, Hajishirzi H. OLMoE: Open Mixture-of-Experts Language Models. 2024. Available from: https://arxiv.org/abs/2409.02060.*

77. ^*Groeneveld D, Beltagy I, Walsh P, Bhagia A, Kinney R, Tafjord O, Jha A, Ivison H, Magnusson I, Wang Y, Arora S, Atkinson D, Authur R, Chandu KR, Cohan A, Dumas J, Elazar Y, Gu Y, Hessel J, Khot T, Merrill W, Morrison JD, Muennighoff N, Naik A, Nam C, Peters ME, Pyatkin V, Ravichander A, Schwenk D, Shah S, Smith W, Strubell E, Subramani N, Wortsman M, Dasigi P, Lambert N, Richardson K, Zettlemoyer L, Dodge J, Lo K, Soldaini L, Smith NA, Hajishirzi H. Olmo: Accelerating the science of language models. arXiv preprint. 2024.*

78. a, b*Team OLMo, Walsh P, Soldaini L, Groeneveld D, Lo K, Arora S, Bhagia A, Gu Y, Huang S, Jordan M, Lambert N, Schwenk D, Tafjord O, Anderson T, Atkinson D, Brahman F, Clark C, Dasigi P, Dziri N, Guerquin M, Ivison H, Koh PW, Liu J, Malik S, Merrill W, Miranda LJV, Morrison J, Murray T, Nam C, Pyatkin V, Rangapur A, Schmitz M, Skjonsberg S, Wadden D, Wilhelm C, Wilson M, Zettlemoyer L, Farhadi A, Smith NA, Hajishirzi H. "2 OLMo 2 Furious". arXiv [cs.CL]. 2024. Available from: https://arxiv.org/abs/2501.00656.*

79. ^*Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, He H, Leahy C, McDonell K, Phang J, Pieler M, Prashanth USVSN S, Purohit S, Reynolds L, Tow J, Wang B, Weinbach S. GPT-NeoX-20B: An Open-Source Autoregressive Language Model, 2022. arXiv [cs.CL]. Available from: https://arxiv.org/abs/2204.06745.*

80. a, b*Zhang J, Huang J, Jin S, Lu S (2024). "Vision-language models for vision tasks: A survey". IEEE Transactions on Pattern Analysis and Machine Intelligence.* **2024**.

81. ^*Chen T, Kornblith S, Norouzi M, Hinton G (2020). "A simple framework for contrastive learning of visual representations". In: International conference on machine learning. PMLR. pp. 1597--1607.*

82. ^*Yao L, Han J, Wen Y, Liang X, Xu D, Zhang W, Li Z, Xu C, Xu H (2022). "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection". Advances in Neural Information Processing Systems.* **35***: 9125–9138.*

83. ^*Fini E, Shukor M, Li X, Dufter P, Klein M, Haldimann D, Aitharaju S, Béthune L, Gan Z, Turrisi V, Toshev A, Eichner M, Yang Y, Nabi M, Susskind J, El-Nouby A. "Multimodal Autoregressive Pre-Training of La*

rge Vision Encoders". 2024. Available from: https://arxiv.org/abs/2411.14402.

84. ^*Maaz M, Rasheed H, Khan S, Khan F (2024). "VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding". arXiv. arXiv:2406.09418 [cs.CV].*
85. ^*Zhao L, Gundavarapu NB, Yuan L, Zhou H, Yan S, Sun JJ, Friedman L, Qian R, Weyand T, Zhao Y, Hornung R, Schroff F, Yang MH, Ross DA, Wang H, Adam H, Sirotenko M, Liu T, Gong B (2024). "VideoPrism: A Foundational Visual Encoder for Video Understanding". arXiv. arXiv:2402.13217 [cs.CV].*
86. a, b*Liu H, Li C, Wu Q, Lee YJ (2023). "Visual Instruction Tuning". In: NeurIPS.*
87. ^*Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009). "ImageNet: A large-scale hierarchical image database". In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248-255. doi:10.1109/CVPR.2009.5206848.*
88. ^*Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016). "Context Encoders: Feature Learning by Inpainting". arXiv. arXiv:1604.07379 [cs.CV].*
89. a, b*Liu H, Li C, Wu Q, Lee YJ (2023). "Visual Instruction Tuning". arXiv. Available from: https://arxiv.org/abs/2304.08485.*
90. ^*Zamir A, Sax A, Shen W, Guibas L, Malik J, Savarese S (2018). "Taskonomy: Disentangling Task Transfer Learning". arXiv. arXiv:1804.08328 [cs.CV].*
91. ^*Holmberg OG, Köhler ND, Martins T, et al. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. Nature Machine Intelligence. 2:719–726, 2020. doi:10.1038/s42256-020-00247-1.*
92. ^*Peng B, Li C, He P, Galley M, Gao J (2023). "Instruction Tuning with GPT-4". arXiv. arXiv:2304.03277 [cs.CL].*
93. a, b*Lin H, Cheng X, Wu X, Yang F, Shen D, Wang Z, Song Q, Yuan W (2021). "CAT: Cross Attention in Vision Transformer". arXiv. Available from: https://arxiv.org/abs/2106.05786.*
94. ^*Kim W, Son B, Kim I (2021). "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision". arXiv. Available from: https://arxiv.org/abs/2102.03334.*
95. ^*Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T, Gong Z, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A, Simonyan K (2022). "Flamingo: a Visual Language Model for Few-Shot Learning". arXiv. arXiv:2204.14198.*
96. ^*Peng Z, Wang W, Dong L, Hao Y, Huang S, Ma S, Wei F (2023). "Kosmos-2: Grounding Multimodal Large Language Models to the World". arXiv. arXiv:2306.14824 [cs.CL].*

97. ^a, b^Lu J, Batra D, Parikh D, Lee S (2019). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". arXiv. arXiv:1908.02265 [cs.CV].

98. ^He K, Fan H, Wu Y, Xie S, Girshick R (2020). "Momentum Contrast for Unsupervised Visual Representation Learning". arXiv. arXiv:1911.05722 [cs.CV].

99. ^He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2021). "Masked Autoencoders Are Scalable Vision Learners". arXiv. arXiv:2111.06377 [cs.CV].

100. ^van den Oord A, Li Y, Vinyals O (2019). "Representation Learning with Contrastive Predictive Coding". arXiv. arXiv:1807.03748 [cs.LG].

101. ^Misra I, van der Maaten L. Self-Supervised Learning of Pretext-Invariant Representations. 2019. Available from: https://arxiv.org/abs/1912.01991.

102. ^O'Shea K, Nash R (2015). "An Introduction to Convolutional Neural Networks". arXiv. Available from: https://arxiv.org/abs/1511.08458.

103. ^Dosovitskiy A. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". arXiv preprint arXiv:2010.11929.

104. ^Li J, Li D, Xiong C, Hoi S. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In: International conference on machine learning. PMLR; 2022. p. 12888–12900.

105. ^Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020). "Language Models are Few-Shot Learners". arXiv. arXiv:2005.14165 [cs.CL].

106. ^Murtagh F (1991). "Multilayer perceptrons for classification and regression". Neurocomputing. 2 (5): 183–197. doi:10.1016/0925-2312(91)90023-5. Link to article.

107. ^Howard J, Ruder S (2018). "Universal Language Model Fine-tuning for Text Classification". arXiv. arXiv:1801.06146 [cs.CL].

108. ^Peters ME, Ruder S, Smith NA (2019). "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". arXiv. arXiv:1903.05987 [cs.CL].

109. ^Tsimpoukelli M, Menick J, Cabi S, Eslami SMA, Vinyals O, Hill F (2021). "Multimodal Few-Shot Learning with Frozen Language Models". arXiv. arXiv:2106.13884 [cs.CV].

110. ^Wang X, Zhang X, Luo Z, Sun Q, Cui Y, Wang J, Zhang F, Wang Y, Li Z, Yu Q, et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869. 2024.

111. ^*Zhang B, Sennrich R (2019). "Root Mean Square Layer Normalization". arXiv.* arXiv:1910.07467 [cs.LG].

112. ^*Ainslie J, Lee-Thorp J, de Jong M, Zemlyanskiy Y, Lebrón F, Sanghai S (2023). "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints". arXiv.* arXiv:2305.13245 [cs.CL].

113. ^*Yew Ken Chia, Hong P, Bing L, Poria S (2023). "INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models". arXiv.* arXiv:2306.04757 [cs.CL].

114. ^*Zhang J, Huang J, Jin S, Lu S. "Vision-Language Models for Vision Tasks: A Survey". 2024. Available from:* https://arxiv.org/abs/2304.00685.

115. a, b*Lu P, Bansal H, Xia T, Liu J, Li C, Hajishirzi H, Cheng H, Chang KW, Galley M, Gao J (2024). "MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts". arXiv.* arXiv:2310.02255 [cs.CV].

116. ^*Fu C, Zhang YF, Yin S, Li B, Fang X, Zhao S, Duan H, Sun X, Liu Z, Wang L, Shan C, He R (2024). "MME-Survey: A Comprehensive Survey on Evaluation of Multimodal LLMs". arXiv. Available from:* https://arxiv.org/abs/2411.15296.

117. ^*Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, Parikh D, Rohrbach M (2019). "Towards VQA Models That Can Read". arXiv. Available from:* https://arxiv.org/abs/1904.08920.

118. ^*Mathew M, Karatzas D, Jawahar CV (2021). "DocVQA: A Dataset for VQA on Document Images". arXiv.* arXiv:2007.00398 [cs.CV].

119. a, b, c*Liu Y, Duan H, Zhang Y, Li B, Zhang S, Zhao W, Yuan Y, Wang J, He C, Liu Z, Chen K, Lin D (2024). "MMBench: Is Your Multi-modal Model an All-around Player?" arXiv.* arXiv:2307.06281 [cs.CV].

120. ^*Li H, Zhang Y, Koto F, Yang Y, Zhao H, Gong Y, Duan N, Baldwin T (2024). "CMMLU: Measuring massive multitask language understanding in Chinese". arXiv.* arXiv:2306.09212 [cs.CL].

121. ^*Huang Y, Bai Y, Zhu Z, Zhang J, Zhang J, Su T, Liu J, Lv C, Zhang Y, Lei J, Fu Y, Sun M, He J (2023). "C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models". arXiv.* 2305.08322.

122. ^*Tang J, Liu Q, Ye Y, Lu J, Wei S, Lin C, Li W, Bin Mahmood MFF, Feng H, Zhao Z, Wang Y, Liu Y, Liu H, Bai X, Huang C (2024). "MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering". arXiv. Available from:* https://arxiv.org/abs/2405.11985.

123. ^*Wang K, Pan J, Shi W, Lu Z, Zhan M, Li H (2024). "Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset". arXiv. Available from:* https://arxiv.org/abs/2402.14804.

124. a, b, c, d, eYu W, Yang Z, Li L, Wang J, Lin K, Liu Z, Wang X, Wang L (2024). "MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities". arXiv. Available from: https://arxiv.org/abs/2308.02490.

125. ^Liu Y, Li Z, Huang M, Yang B, Yu W, Li C, Yin XC, Liu CL, Jin L, Bai X (2024). "Ocrbench: on the hidden mystery of ocr in large multimodal models". Science China Information Sciences. 67 (12).

126. ^Fu C, Chen P, Shen Y, Qin Y, Zhang M, Lin X, Yang J, Zheng X, Li K, Sun X, Wu Y, Ji R (2024). "MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models". arXiv. Available from: https://arxiv.org/abs/2306.13394.

127. a, b, cYing K, Meng F, Wang J, Li Z, Lin H, Yang Y, Zhang H, Zhang W, Lin Y, Liu S, Lei J, Lu Q, Chen R, Xu P, Zhang R, Zhang H, Gao P, Wang Y, Qiao Y, Luo P, Zhang K, Shao W (2024). "MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI". arXiv. Available from: https://arxiv.org/abs/2404.16006.

128. ^Mathew M, Bagal V, Pérez Tito R, Karatzas D, Valveny E, Jawahar CV. InfographicVQA. 2021. Available from: https://arxiv.org/abs/2104.12756.

129. ^Kembhavi A, Salvato M, Kolve E, Seo M, Hajishirzi H, Farhadi A (2016). "A Diagram Is Worth A Dozen Images". arXiv. arXiv:1603.07396 [cs.CV].

130. a, b, cMasry A, Long DX, Tan JQ, Joty S, Hoque E (2022). "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning". arXiv. arXiv:2203.10244 [cs.CL].

131. a, b, cXiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, Wenhu Chen (2024). "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI". arXiv. Available from: https://arxiv.org/abs/2311.16502.

132. a, bGhosh D, Hajishirzi H, Schmidt L (2023). "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment". arXiv. arXiv:2310.11513 [cs.CV].

133. a, bHuang K, Sun K, Xie E, Li Z, Liu X (2023). "T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation". arXiv. arXiv:2307.06350 [cs.CV].

134. a, bHu X, Wang R, Fang Y, Fu B, Cheng P, Yu G (2024). "ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment". arXiv. arXiv:2403.05135 [cs.CV].

135. ^Lin Z, Pathak D, Li B, Li J, Xia X, Neubig G, Zhang P, Ramanan D (2024). "Evaluating Text-to-Visual Generation with Image-to-Text Generation". arXiv. Available from: https://arxiv.org/abs/2404.01291.

136. ^Li B, Lin Z, Pathak D, Li J, Fei Y, Wu K, Ling T, Xia X, Zhang P, Neubig G, Ramanan D. "GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation". 2024. Available from: https://arxiv.org/abs/2406.13743.
137. ^Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen JR (2023). "Evaluating Object Hallucination in Large Vision-Language Models". arXiv. arXiv:2305.10355 [cs.CV].
138. a, bHendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021). "Measuring Massive Multitask Language Understanding". arXiv. arXiv:2009.03300.
139. ^Chen L, Li J, Dong X, Zhang P, Zang Y, Chen Z, Duan H, Wang J, Qiao Y, Lin D, Zhao F (2024). "Are We on the Right Way for Evaluating Large Vision-Language Models?" arXiv. arXiv:2403.20330 [cs.CV].
140. ^Song W, Li Y, Xu J, Wu G, Ming L, Yi K, Luo W, Li H, Du Y, Guo F, Yu K (2024). "M3GIA: A Cognition Inspired Multilingual and Multimodal General Intelligence Ability Benchmark". arXiv. arXiv:2406.05343 [cs.AI].
141. ^Zhong W, Cui R, Guo Y, Liang Y, Lu S, Wang Y, Saied A, Chen W, Duan N. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. 2023. Available from: https://arxiv.org/abs/2304.06364.
142. a, b, c, dMangalam K, Akshulakov R, Malik J (2023). "EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding". arXiv. Available from: https://arxiv.org/abs/2308.09126.
143. ^Zhou J, Shu Y, Zhao B, Wu B, Xiao S, Yang X, Xiong Y, Zhang B, Huang T, Liu Z (2024). "MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding". arXiv. Available from: https://arxiv.org/abs/2406.04264.
144. a, bLi K, Wang Y, He Y, Li Y, Wang Y, Liu Y, Wang Z, Xu J, Chen G, Luo P, Wang L, Qiao Y (2024). "MVBench: A Comprehensive Multi-modal Video Understanding Benchmark". arXiv. Available from: https://arxiv.org/abs/2311.17005.
145. ^Fu C, Dai Y, Luo Y, Li L, Ren S, Zhang R, Wang Z, Zhou C, Shen Y, Zhang M, Chen P, Li Y, Lin S, Zhao S, Li K, Xu T, Zheng X, Chen E, Ji R, Sun X (2024). "Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis". arXiv. arXiv:2405.21075 [cs.CV].
146. ^Pătrăucean V, Smaira L, Gupta A, Recasens Continente A, Markeeva L, Banarse D, Koppula S, Heyward J, Malinowski M, Yang Y, Doersch C, Matejovicova T, Sulsky Y, Miech A, Frechette A, Klimczak H, Koster R, Zhang J, Winkler S, Aytar Y, Osindero S, Damen D, Zisserman A, Carreira J (2023). "Perception Test: A Diagnostic Benchmark for Multimodal Video Models". arXiv. Available from: https://arxiv.org/abs/2305.13786.

147. ^Hudson DA, Manning CD (2019). "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering". arXiv. arXiv:1902.09506 [cs.CL].

148. a, b, cZellers R, Bisk Y, Farhadi A, Choi Y (2019). "From Recognition to Cognition: Visual Commonsense Reasoning". arXiv. arXiv:1811.10830 [cs.CV].

149. a, b, c, dLi B, Wang R, Wang G, Ge Y, Ge Y, Shan Y (2023). "SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension". arXiv. arXiv:2307.16125 [cs.CL].

150. a, b{X.AI}. "Grok-1.5v: The Multimodal Version of our AI". Blog post. May 2024. Available from: https://x.ai/blog/grok-1.5v. Accessed on [Insert Date of Access].

151. ^Yue X, Zheng T, Ni Y, Wang Y, Zhang K, Tong S, Sun Y, Yu B, Zhang G, Sun H, Su Y, Chen W, Neubig G (2024). "MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark". arXiv. arXiv:2409.02813 [cs.CL].

152. ^Biten AF, Tito R, Mafla A, Gomez L, Rusiñol M, Valveny E, Jawahar CV, Karatzas D (2019). "Scene Text Visual Question Answering". arXiv. Available from: https://arxiv.org/abs/1905.13648.

153. ^Li B, Lin Z, Peng W, Nyandwi JD, Jiang D, Ma Z, Khanuja S, Krishna R, Neubig G, Ramanan D (2024). "NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples". arXiv. arXiv:2410.14669 [cs.CV].

154. a, b, c, dSavva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, Straub J, Liu J, Koltun V, Malik J, Parikh D, Batra D (2019). "Habitat: A Platform for Embodied AI Research". In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

155. a, b, c, dXia F, Zamir AR, He Z, Sax A, Malik J, Savarese S (2018). "Gibson env: Real-world perception for embodied agents". In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079.

156. a, b, c, dLi C, Xia F, Martín-Martín R, Lingelbach M, Srivastava S, Shen B, Vainio K, Gokmen C, Dharan G, Jain T, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. arXiv preprint arXiv:2108.03272. 2021.

157. a, bMittal M, Yu C, Yu Q, Liu J, Rudin N, Hoeller D, Yuan JL, Singh R, Guo Y, Mazhar H, Mandlekar A, Babich B, State G, Hutter M, Garg A (2023). "Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments". IEEE Robotics and Automation Letters. 8(6): 3740-3747. doi:10.1109/LRA.2023.3270034.

158. a, bZhou S, Xu FF, Zhu H, Zhou X, Lo R, Sridhar A, Cheng X, Bisk Y, Fried D, Alon U, et al. (2023). "WebArena: A Realistic Web Environment for Building Autonomous Agents". arXiv preprint arXiv:2307.13854. A

vailable from: https://webarena.dev.

159. a, b Mees O, Hermann L, Rosete-Beas E, Burgard W (2022). "CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks". arXiv. arXiv:2112.03227 [cs.RO].

160. a, b Zheng K, Chen X, Jenkins O, Wang XE (2022). "VLMbench: A Compositional Benchmark for Vision-and-Language Manipulation". Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. Available from: https://openreview.net/forum?id=NAYoSV3tk9.

161. a, b Garcia R, Chen S, Schmid C (2024). "Towards Generalizable Vision-Language Robotic Manipulation: A Benchmark and LLM-guided 3D Policy". arXiv. arXiv:2410.01345 [cs.RO].

162. a, b, c Jiang Y, Gupta A, Zhang Z, Wang G, Dou Y, Chen Y, Fei-Fei L, Anandkumar A, Zhu Y, Fan L (2022). "Vima: General robot manipulation with multimodal prompts". arXiv preprint arXiv:2210.03094. 2 (3): 6.

163. a, b Hu A, Russell L, Yeo H, Murez Z, Fedoseev G, Kendall A, Shotton J, Corrado G (2023). "Gaia-1: A generative world model for autonomous driving". arXiv preprint arXiv:2309.17080. Available from: https://arxiv.org/abs/2309.17080.

164. a, b Yang M, Du Y, Ghasemipour K, Tompson J, Schuurmans D, Abbeel P (2023). "Learning Interactive Real-World Simulators". arXiv preprint arXiv:2310.06114. arXiv:2310.06114.

165. a, b Liu H, Yan W, Zaharia M, Abbeel P (2024). "World Model on Million-Length Video And Language With Blockwise RingAttention". arXiv. Available from: https://arxiv.org/abs/2402.08268.

166. a, b Genesis Authors. Genesis: A Universal and Generative Physics Engine for Robotics and Beyond [software]. December 2024. Available from: https://github.com/Genesis-Embodied-AI/Genesis.

167. ^ Xu Z, Jain S, Kankanhalli M (2024). "Hallucination is Inevitable: An Innate Limitation of Large Language Models". arXiv. Available from: https://arxiv.org/abs/2401.11817.

168. a, b, c Li Z, Mondal I, Nghiem H, Liang Y, Boyd-Graber JL. "PEDANTS: Cheap but Effective and Interpretable Answer Equivalence." In: Al-Onaizan Y, Bansal M, Chen Y-N, editors. Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 9373-9398. doi:10.18653/v1/2024.findings-emnlp.548. Available from: https://aclanthology.org/2024.findings-emnlp.548.

169. ^ Izacard G, Grave E (2021). "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". arXiv. Available from: https://arxiv.org/abs/2007.01282.

170. ^ Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2021). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". arXiv.

arXiv:2005.11401 [cs.CL].

171. ^Chen D, Fisch A, Weston J, Bordes A (2017). "Reading Wikipedia to Answer Open-Domain Questions". arXiv. arXiv:1704.00051.

172. a, b Bulian J, Buck C, Gajewski W, Boerschinger B, Schuster T (2022). "Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation". arXiv preprint arXiv:2202.07654.

173. a, b Chen A, Stanovsky G, Singh S, Gardner M. Evaluating question answering evaluation. In: Fisch A, Talmor A, Jia R, Seo M, Choi E, Chen D, editors. Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics; 2019. p. 119-124. doi: 10.18653/v1/D19-5817. https://aclanthology.org/D19-5817.

174. a, b Xu F, Song Y, Iyyer M, Choi E (2023). "A Critical Evaluation of Evaluations for Long-form Question Answering". ArXiv. **abs/2305.18201**. S2CID 258960565.

175. ^Lin CY. "ROUGE: A Package for Automatic Evaluation of Summaries." In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74-81. Available from: https://aclanthology.org/W04-1013.

176. ^Papineni K, Roukos S, Ward T, Zhu WJ. "BLEU: a method for automatic evaluation of machine translation." In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics; 2002. p. 311–318. doi:10.3115/1073083.1073135.

177. ^Mañas O, Krojer B, Agrawal A (2023). "Improving Automatic VQA Evaluation Using Large Language Models". arXiv preprint arXiv:2310.02567. arXiv:2310.02567.

178. ^Zhao Y, Zhang H, Si S, Nan L, Tang X, Cohan A (2023). "Large Language Models are Effective Table-to-Text Generators, Evaluators, and Feedback Providers". arXiv preprint arXiv:2305.14987. Available from: https://arxiv.org/abs/2305.14987.

179. ^Kamalloo E, Dziri N, Clarke C, Rafiei D. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023 Jul; Toronto, Canada. Association for Computational Linguistics. p. 5591-5606. doi:10.18653/v1/2023.acl-long.307. Available from: https://aclanthology.org/2023.acl-long.307.

180. ^Balepur N, Ravichander A, Rudinger R. "Artifacts or abduction: How do LLMs answer multiple-choice questions without the question?" In: Ku LW, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thail

and: Association for Computational Linguistics; 2024. p. 10308-10330. doi:10.18653/v1/2024.acl-long.555. Available from: https://aclanthology.org/2024.acl-long.555.

181. ^*Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y (2022). "CLIPScore: A Reference-free Evaluation Metric for Image Captioning". arXiv. arXiv:2104.08718 [cs.CV].*
182. ^*Turing AM. "Computing Machinery and Intelligence". Mind. 59 (236): 433–460, 1950. Available from: http://www.jstor.org/stable/2251299.*
183. ^*Liu Y, Chen W, Bai Y, Liang X, Li G, Gao W, Lin L (2024). "Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI". arXiv. arXiv:2407.06886 [cs.CV].*
184. a, b*Han Y, Zhang C, Chen X, Yang X, Wang Z, Yu G, Fu B, Zhang H (2023). "ChartLlama: A Multimodal LLM for Chart Understanding and Generation". arXiv. arXiv:2311.16483 [cs.CV].*
185. ^*Mondal I, Li Z, Hou Y, Natarajan A, Garimella A, Boyd-Graber JL. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In: Al-Onaizan Y, Bansal M, Chen Y-N, editors. Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 13342-13375. doi:10.18653/v1/2024.findings-emnlp.780. Available from: https://aclanthology.org/2024.findings-emnlp.780.*
186. ^*Terzidou T, Tsiatsos T. Embodied agent. In: Khosrow-Pour M, editor. Encyclopedia of Information Science and Technology, Third Edition. IGI Global; 2015. Chapter 250, p. 10. doi:10.4018/978-1-4666-5888-2.ch250.*
187. ^*Gunter T, Wang Z, Wang C, Pang R, Narayanan A, Zhang A, Zhang B, Chen C, Chiu C-C, Qiu D, Gopinath D, Yap DA, Yin D, Nan F, Weers F, Yin G, Huang H, Wang J, Lu J, Peebles J, Ye K, Lee M, Du N, Chen Q, Keunebroek Q, Wiseman S, Evans S, Lei T, Rathod V, Kong X, Du X, Li Y, Wang Y, Gao Y, Ahmed Z, Xu Z, Lu Z, Rashid A, Jose AM, Doane A, Bencomo A, Vanderby A, Hansen A, Jain A, Anupama AM, Kamal A, Wu B, Brum C, Maalouf C, Erdenebileg C, Dulhanty C, Moritz D, Kang D, Jimenez E, Ladd E, Shi F, Bai F, Chu F, Hohman F, Kotek H, Coleman HG, Li J, Bigham J, Cao J, Lai J, Cheung J, Shan J, Zhou J, Li J, Qin J, Singh K, Vega K, Zou K, Heckman L, Gardiner L, Bowler M, Cordell M, Cao M, Hay N, Shahdadpuri N, Godwin O, Dighe P, Rachapudi P, Tantawi R, Frigg R, Davarnia S, Shah S, Guha S, Sirovica S, Ma S, Ma S, Wang S, Kim S, Jayaram S, Shankar V, Paidi V, Kumar V, Wang X, Zheng X, Cheng W, Shrager Y, Ye Y, Tanaka Y, Guo Y, Meng Y, Luo ZT, Ouyang Z, Aygar A, Wan A, Walkingshaw A, Narayanan A, Lin A, Farooq A, Ramerth B, Reed C, Bartels C, Chaney C, Riazati D, Yang EL, Feldman E, Hochstrasser G, Seguin G, Belousova I, Pelemans J, Yang K, Vahid KA, Cao L, Najibi M, Zuliani M, Horton M, Cho M, Bhendawade N, Dong*

P, Maj P, Agrawal P, Shan Q, Fu Q, Poston R, Xu S, Liu S, Rao S, Heeramun T, Merth T, Rayala U, Cui V, Sridhar VR, Zhang W, Zhang W, Wu W, Zhou X, Liu X, Zhao Y, Xia Y, Ren Z, Ren Z. *Apple Intelligence Foundation Language Models*. 2024. Available from: https://arxiv.org/abs/2407.21075.

188. ^Baechler G, Sunkara S, Wang M, Zubach F, Mansoor H, Etter V, Cărbune V, Lin J, Chen J, Sharma A (2024). "ScreenAI: A Vision-Language Model for UI and Infographics Understanding". arXiv. Available from: https://arxiv.org/abs/2402.04615.

189. ^Dorka N, Marecki J, Anwar A (2024). "Training a Vision Language Model as Smartphone Assistant". arXiv. arXiv:2404.08755.

190. ^Niu R, Li J, Wang S, Fu Y, Hu X, Leng X, Kong H, Chang Y, Wang Q (2024). "ScreenAgent: A Vision Language Model-driven Computer Control Agent". arXiv. arXiv:2402.07945 [cs.HC].

191. ^Yang J, Dong Y, Liu S, Li B, Wang Z, Jiang C, Tan H, Kang J, Zhang Y, Zhou K, Liu Z (2024). "Octopus: Embodied Vision-Language Programmer from Environmental Feedback". arXiv. arXiv:2310.08588 [cs.CV].

192. ^Zhang G, Zhang Y, Zhang K, Tresp V (2023). "Can Vision-Language Models be a Good Guesser? Exploring VLMs for Times and Location Reasoning". arXiv. arXiv:2307.06166 [cs.CV].

193. ^Huang J, Zhang J (2024). "A survey on evaluation of multimodal large language models". arXiv preprint arXiv:2408.15769.

194. ^Ding H, Seenivasan L, Shu H, Byrd G, Zhang H, Xiao P, Barragan JA, Taylor RH, Kazanzides P, Unberath M (2024). "Towards Robust Automation of Surgical Systems via Digital Twin-based Scene Representations from Foundation Models". arXiv. Available from: https://arxiv.org/abs/2409.13107.

195. ^Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014). "Generative Adversarial Networks". arXiv. arXiv:1406.2661 [stat.ML].

196. ^Ho J, Jain A, Abbeel P (2020). "Denoising Diffusion Probabilistic Models". arXiv. arXiv:2006.11239 [cs.LG].

197. ^Supermeme.ai: Turn Text into Memes Using AI. 2024. Accessed: 2024-12-24.

198. ^Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, Vyas A, Shi B, Ma C, Chuang C, et al. (2024). "Movie gen: A cast of media foundation models". arXiv preprint arXiv:2410.13720.

199. ^Liu Y, Chen W, Bai Y, Liang X, Li G, Gao W, Lin L (2024). "Aligning cyber space with physical world: A comprehensive survey on embodied ai". arXiv preprint arXiv:2407.06886.

200. ^a, bDuan J, Pumacay W, Kumar N, Wang YR, Tian S, Yuan W, Krishna R, Fox D, Mandlekar A, Guo Y (2024). "AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulatio

n". arXiv preprint arXiv:2410.00371. arXiv:2410.00371.

201. ^Chen B, Xu Z, Kirmani S, Ichter B, Sadigh D, Guibas L, Xia F (2024). "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14455–14465.

202. a, bShek CL, Wu X, Suttle WA, Busart C, Zaroukian E, Manocha D, Tokekar P, Bedi AS (2023). "Lancar: Leveraging language for context-aware robot locomotion in unstructured environments". arXiv preprint arXiv:2310.00481.

203. ^Liu S, Zhang J, Gao RX, Wang XV, Wang L. "Vision-language model-driven scene understanding and robotic object manipulation." In: 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE). IEEE; 2024. p. 21–26.

204. ^Yang Z, Garrett C, Fox D, Lozano-Pérez T, Kaelbling LP (2024). "Guiding Long-Horizon Task and Motion Planning with Vision Language Models". arXiv preprint arXiv:2410.02193.

205. a, bChen Y, Arkin J, Dawson C, Zhang Y, Roy N, Fan C. "Autotamp: Autoregressive task and motion planning with llms as translators and checkers." In: 2024 IEEE International conference on robotics and automation (ICRA). IEEE; 2024. p. 6695-6702.

206. a, b, c, dHuang S, Jiang Z, Dong H, Qiao Y, Gao P, Li H (2023). "Instruct2act: Mapping multi-modality instructions to robotic actions with large language model". arXiv preprint arXiv:2305.11176.

207. a, bMajumdar A, Aggarwal G, Devnani B, Hoffman J, Batra D (2022). "Zson: Zero-shot object-goal navigation using multimodal goal embeddings". Advances in Neural Information Processing Systems. **35**: 32340–32352.

208. a, bGuan T, Yang Y, Cheng H, Lin M, Kim R, Madhivanan R, Sen A, Manocha D (2024). "LOC-ZSON: Language-driven Object-Centric Zero-Shot Object Retrieval and Navigation". arXiv preprint arXiv:2405.05363. Available from: https://arxiv.org/abs/2405.05363.

209. a, bYokoyama N, Ha S, Batra D, Wang J, Bucher B. "Vlfm: Vision-language frontier maps for zero-shot semantic navigation." In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2024. p. 42–48.

210. a, bSong D, Liang J, Payandeh A, Xiao X, Manocha D. "Socially aware robot navigation through scoring using vision-language models". arXiv preprint arXiv:2404.00210. 2024.

211. a, bWang B, Zhang J, Dong S, Fang I, Feng C (2024). "Vlm see, robot do: Human demo video to robot action plan via vision language model". arXiv preprint arXiv:2410.08792.

212. ^Chen Y, Arkin J, Zhang Y, Roy N, Fan C. "Scalable multi-robot collaboration with large language mode ls: Centralized or decentralized systems?" In: 2024 IEEE International Conference on Robotics and Auto mation (ICRA). IEEE; 2024. p. 4311–4317.

213. ^Wang Y, Xiao R, Kasahara JYL, Yajima R, Nagatani K, Yamashita A, Asama H (2024). "DART-LLM: De pendency-Aware Multi-Robot Task Decomposition and Execution using Large Language Models". arXi v preprint arXiv:2411.09022.

214. a, b, cSeff A, Cera B, Chen D, Ng M, Zhou A, Nayakanti N, Refaat KS, Al-Rfou R, Sapp B (2023). "Motionl m: Multi-agent motion forecasting as language modeling". Proceedings of the IEEE/CVF International Conference on Computer Vision. 8579--8590.

215. a, bJiang B, Chen X, Liu W, Yu J, Yu G, Chen T (2023). "Motiongpt: Human motion as a foreign languag e". Advances in Neural Information Processing Systems. 36: 20067–20079.

216. ^Zeng Y, Mu Y, Shao L (2024). "Learning Reward for Robot Skills Using Large Language Models via Self -Alignment". arXiv preprint arXiv:2405.07162.

217. a, bYu W, Gileadi N, Fu C, Kirmani S, Lee KH, Gonzalez Arenas M, Chiang HTL, Erez T, Hasenclever L, Hu mplik J, et al. Language to rewards for robotic skill synthesis. arXiv preprint arXiv:2306.08647. 2023.

218. ^Ma YJ, Liang W, Wang G, Huang DA, Bastani O, Jayaraman D, Zhu Y, Fan L, Anandkumar A (2023). "Eu reka: Human-level reward design via coding large language models". arXiv preprint arXiv:2310.12931.

219. ^Sermanet P, Ding T, Zhao J, Xia F, Dwibedi D, Gopalakrishnan K, Chan C, Dulac-Arnold G, Maddineni S, Joshi NJ, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2024. p. 645-652.

220. ^Xu M, Huang P, Yu W, Liu S, Zhang X, Niu Y, Zhang T, Xia F, Tan J, Zhao D (2023). "Creative Robot Too l Use with Large Language Models". arXiv. arXiv:2310.13065.

221. ^O'Neill A, Rehman A, Gupta A, Maddukuri A, Gupta A, Padalkar A, Lee A, Pooley A, Gupta A, Mandlekar A, et al. Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint arXiv:2310.088 64. 2023.

222. ^Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, Gopalakrishnan K, Hausman K, Herzog A, Hsu J, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817. 2022.

223. ^Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, Ding T, Driess D, Dubey A, Finn C, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint ar Xiv:2307.15818. 2023.

224. ^*Shah D, Osiński B, Levine S, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on robot learning. PMLR; 2023. p. 492–504.*

225. ^*Cheng AC, Ji Y, Yang Z, Zou X, Kautz J, Bıyık E, Yin H, Liu S, Wang X (2024). "NaVILA: Legged Robot Vision-Language-Action Model for Navigation". arXiv preprint arXiv:2412.04453.*

226. ^*Shah D, Equi MR, Osiński B, Xia F, Ichter B, Levine S. Navigation with large language models: Semantic guesswork as a heuristic for planning. In: Conference on Robot Learning. PMLR; 2023. p. 2683–2699.*

227. a, b*Garrett CR, Chitnis R, Holladay R, Kim B, Silver T, Kaelbling LP, Lozano-Pérez T (2021). "Integrated task and motion planning". Annual review of control, robotics, and autonomous systems. 4 (1): 265–293.*

228. ^*Shah R, Mart{\'\i}n-Mart{\'\i}n R, Zhu Y (2023). "Mutex: Learning unified policies from multimodal task specifications". arXiv preprint arXiv:2309.14320. Available from: https://arxiv.org/abs/2309.14320.*

229. ^*Wang C, Hasler S, Tanneberg D, Ocker F, Joublin F, Ceravola A, Deigmoeller J, Gienger M (2024). "Large language models for multi-modal human-robot interaction". arXiv preprint arXiv:2401.15174.*

230. a, b*Liu M, Jiang J, Zhu C, Yin XC (2023). "Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 6662–6671.*

231. ^*Wu D, Han W, Wang T, Dong X, Zhang X, Shen J (2023). "Referring multi-object tracking". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14633–14642.*

232. ^*Tian X, Gu J, Li B, Liu Y, Wang Y, Zhao Z, Zhan K, Jia P, Lang X, Zhao H (2024). "Drivevlm: The convergence of autonomous driving and large vision-language models". arXiv preprint arXiv:2402.12289.*

233. ^*Mao J, Qian Y, Ye J, Zhao H, Wang Y (2023). "Gpt-driver: Learning to drive with gpt". arXiv preprint arXiv:2310.01415.*

234. ^*Sha H, Mu Y, Jiang Y, Chen L, Xu C, Luo P, Li SE, Tomizuka M, Zhan W, Ding M (2023). "Languagempc: Large language models as decision makers for autonomous driving". arXiv preprint arXiv:2310.03026. 2023.*

235. ^*Chen L, Sinavski O, Hünermann J, Karnsund A, Willmott AJ, Birch D, Maund D, Shotton J. "Driving with llms: Fusing object-level vector modality for explainable autonomous driving." In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2024. p. 14093–14100.*

236. ^*Wen L, Fu D, Li X, Cai X, Ma T, Cai P, Dou M, Shi B, He L, Qiao Y (2023). "Dilu: A knowledge-driven approach to autonomous driving with large language models". arXiv preprint arXiv:2309.16292. arXiv:2309.16292.*

237. ^*Pan C, Yaman B, Nesti T, Mallik A, Allievi AG, Velipasalar S, Ren L (2024). "VLP: Vision Language Planning for Autonomous Driving". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14760–14769.*

238. ^*Xu Z, Zhang Y, Xie E, Zhao Z, Guo Y, Wong KYK, Li Z, Zhao H (2024). "Drivegpt4: Interpretable end-to-end autonomous driving via large language model". IEEE Robotics and Automation Letters.* **2024**.

239. ^*Robey A, Ravichandran Z, Kumar V, Hassani H, Pappas GJ (2024). "Jailbreaking LLM-controlled robots". arXiv preprint arXiv:2410.13691.*

240. a, b*Hundt A, Agnew W, Zeng V, Kacianka S, Gombolay M (2022). "Robots enact malignant stereotypes". In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. p. 743–756.*

241. a, b*Azeem R, Hundt A, Mansouri M, Brandão M (2024). "LLM-Driven Robots Risk Enacting Discrimination, Violence, and Unlawful Actions". arXiv preprint arXiv:2406.08824.*

242. a, b*Wu X, Chakraborty S, Xian R, Liang J, Guan T, Liu F, Sadler BM, Manocha D, Bedi AS (2024). "Highlighting the Safety Concerns of Deploying LLMs/VLMs in Robotics". arXiv. arXiv:2402.10340 [cs.RO].*

243. ^*Wang P, Zhou Q, Wu Y, Chen T, Hu J (2024). "DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis". arXiv preprint arXiv:2412.12225.*

244. ^*Huang Z, Pohovey J, Yammanuru A, Driggs-Campbell K (2024). "LIT: Large Language Model Driven Intention Tracking for Proactive Human-Robot Collaboration--A Robot Sous-Chef Application". arXiv preprint arXiv:2406.13787.*

245. ^*Patel D, Eghbalzadeh H, Kamra N, Iuzzolino ML, Jain U, Desai R (2023). "Pretrained language models as visual planners for human assistance". Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15302–15314.*

246. ^*Cao Q, Chen Y, Lu L, Sun H, Zeng Z, Yang X, Zhang D (2024). "Promoting AI Equity in Science: Generalized Domain Prompt Learning for Accessible VLM Research". arXiv preprint arXiv:2405.08668.*

247. ^*Zheng Z, He Z, Khattab O, Rampal N, Zaharia MA, Borgs C, Chayes JT, Yaghi OM (2023). "Image and Data Mining in Reticular Chemistry Using GPT-4V". arXiv preprint arXiv:2312.05468. arXiv:2312.05468.*

248. a, b*Zhu H, Qin S, Su M, Lin C, Li A, Gao J (2024). "Harnessing large vision and language models in agriculture: A review". arXiv preprint arXiv:2407.19679. Available from: arXiv:2407.19679.*

249. a, b*Ma Y, Hu S, Li X, Wang Y, Liu S, Cheong KH (2024). "Students rather than experts: A new AI for education pipeline to model more human-like and personalised early adolescences". arXiv preprint arXiv:2410.15701. Available from: https://arxiv.org/abs/2410.15701.*

250. ^a, b^ *Wu D, Chen M, Chen X, Liu X (2024). "Analyzing K-12 AI education: A large language model study of classroom instruction on learning theories, pedagogy, tools, and AI literacy". Computers and Education: Artificial Intelligence. 7: 100295.*

251. ^a, b^ *Oliveira EAM de, Silva DFC, Filho ARG (2024). "Improving VR Accessibility Through Automatic 360 S cene Description Using Multimodal Large Language Models". In: Proceedings of the 26th Symposium o n Virtual and Augmented Reality. pp. 289–293.*

252. ^a, b^ *Zhang J, Zheng M, Boyd M, Ohn-Bar E (2021). "X-world: Accessibility, vision, and autonomy meet". Proceedings of the IEEE/CVF International Conference on Computer Vision. 9762–9771.*

253. ^a, b^ *Yildirim N, Richardson H, Wetscherek MT, Bajwa J, Jacob J, Pinnock MA, Harris S, Coelho De Castro D, Bannur S, Hyland S, et al. Multimodal healthcare AI: identifying and designing clinically relevant visi on-language applications for radiology. In: Proceedings of the CHI Conference on Human Factors in Co mputing Systems. 2024. p. 1-22.*

254. ^a, b^ *Li Z, Song D, Yang Z, Wang D, Li F, Zhang X, Kinahan PE, Qiao Y (2024). "VisionUnite: A Vision-Lang uage Foundation Model for Ophthalmology Enhanced with Clinical Knowledge". arXiv preprint arXiv:2 408.02865.*

255. ^a, b^ *Chen J, Zhou P, Hua Y, Chong D, Cao M, Li Y, Yuan Z, Zhu B, Liang J (2024). "Vision-Language Mode ls Meet Meteorology: Developing Models for Extreme Weather Events Detection with Heatmaps". arXiv preprint arXiv:2406.09838.*

256. ^ *Gur I, Furuta H, Huang A, Safdari M, Matsuo Y, Eck D, Faust A (2024). "A Real-World WebAgent with P lanning, Long Context Understanding, and Program Synthesis". arXiv. 2307.12856.*

257. ^ *Hong W, Wang W, Lv Q, Xu J, Yu W, Ji J, Wang Y, Wang Z, Dong Y, Ding M, et al. Cogagent: A visual lang uage model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 14281–14290.*

258. ^ *He H, Yao W, Ma K, Yu W, Dai Y, Zhang H, Lan Z, Yu D (2024). "WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models". arXiv preprint arXiv:2401.13919. Available from: https://arx iv.org/abs/2401.13919.*

259. ^ *Qinghong Lin K, Li L, Gao D, Yang Z, Wu S, Bai Z, Lei W, Wang L, Shou MZ (2024). "ShowUI: One Vision -Language-Action Model for GUI Visual Agent". arXiv e-prints. pages arXiv--2411.*

260. ^ *Niu R, Li J, Wang S, Fu Y, Hu X, Leng X, Kong H, Chang Y, Wang Q (2024). "Screenagent: A vision langu age model-driven computer control agent". arXiv preprint arXiv:2402.07945.*

261. ^*Gubbi Mohanbabu A, Pavel A. Context-Aware Image Descriptions for Web Accessibility. In: The 26th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS ’24. ACM; 2024. p. 1–17. doi:10.1145/3663548.3675658.*

262. ^*Liu C, Cheng S, Chen C, Qiao M, Zhang W, Shah A, Bai W, Arcucci R (2023). "M-FLAG: Medical Vision-Language Pre-training with Frozen Language Models and Latent Space Geometry Optimization". arXiv. arXiv:2307.08347 [cs.CV].*

263. ^*Wang Z, Wu Z, Agarwal D, Sun J (2022). "Medclip: Contrastive learning from unpaired medical images and text". arXiv preprint arXiv:2210.10163. Available from: https://arxiv.org/abs/2210.10163.*

264. ^*Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, Zakka C, Reis EP, Rajpurkar P. "Med-flamingo: a multimodal medical few-shot learner." In: Machine Learning for Health (ML4H). PMLR; 2023. p. 353–367.*

265. ^*Bei G. A vision-language model for predicting potential distribution land of soybean double cropping. Frontiers in Environmental Science. 12:1515752.*

266. ^*Lee N, Laghuvarapu S, Park C, Sun J. "Vision Language Model is NOT All You Need: Augmentation Strategies for Molecule Language Models." In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024. p. 1153–1162.*

267. ^*Baral S, Lucy L, Knight R, Ng A, Soldaini L, Heffernan N, Lo K. "DrawEduMath: Evaluating Vision Language Models with Expert-Annotated Students’ Hand-Drawn Math Images." In: The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24; 2024.*

268. ^*Peng S, Fu D, Gao L, Zhong X, Fu H, Tang Z (2024). "Multimath: Bridging visual and mathematical reasoning for large language models". arXiv preprint arXiv:2409.00147.*

269. ^*Garimella A, Amarnath A, Kumar K, Yalla AP, Anandhavelu N, Chhaya N, Srinivasan BV. "He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation." In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021. p. 4534-4545.*

270. ^*Hao X, Chen W, Yan Y, Zhong S, Wang K, Wen Q, Liang Y (2024). "UrbanVLP: A Multi-Granularity Vision-Language Pre-Trained Foundation Model for Urban Indicator Prediction". arXiv preprint arXiv:2403.16831.*

271. a, b*Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. p. 4035–4045.*

272. ^Li Y, Du Y, Zhou K, Wang J, Zhao X, Wen J-R. "Evaluating Object Hallucination in Large Vision-Language Models." In: The 2023 Conference on Empirical Methods in Natural Language Processing.

273. ^Gunjal A, Yin J, Bas E (2024). "Detecting and preventing hallucinations in large vision language models". Proceedings of the AAAI Conference on Artificial Intelligence. 38 (16): 18135–18143.

274. ^Zhai B, Yang S, Zhao X, Xu C, Shen S, Zhao D, Keutzer K, Li M, Yan T, Fan X (2023). "Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption". arXiv preprint arXiv:2310.01779. arXiv:2310.01779.

275. ^Ding P, Wu J, Kuang J, Ma D, Cao X, Cai X, Chen S, Chen J, Huang S (2024). "Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs". Proceedings of the 32nd ACM International Conference on Multimedia. pages 10707–10715.

276. ^Moon YB, Nam HW, Choi W, Oh TH. "Beaf: Observing before-after changes to evaluate hallucination in vision-language models." In: European Conference on Computer Vision. Springer; 2025. p. 232–248.

277. ^Wu X, Guan T, Li D, Huang S, Liu X, Wang X, Xian R, Shrivastava A, Huang F, Boyd-Graber JL, Zhou T, Manocha D. "AutoHallusion: Automatic Generation of Hallucination Benchmarks for Vision-Language Models". 2024. Available from: https://arxiv.org/abs/2406.10900.

278. a, bLiu F, Lin K, Li L, Wang J, Yacoob Y, Wang L (2023). "Mitigating hallucination in large multi-modal models via robust instruction tuning". In: The Twelfth International Conference on Learning Representations.

279. ^Jiang C, Jia H, Dong M, Ye W, Xu H, Yan M, Zhang J, Zhang S (2024). "Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models". In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 525–534.

280. ^Wang J, Wang Y, Xu G, Zhang J, Gu Y, Jia H, Yan M, Zhang J, Sang J (2023). "An llm-free multi-dimensional benchmark for mllms hallucination evaluation". arXiv preprint arXiv:2311.07397.

281. ^Jin H, Hu L, Li X, Zhang P, Chen C, Zhuang J, Wang H (2024). "Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models". arXiv preprint arXiv:2407.01599. Available from: https://arxiv.org/abs/2407.01599.

282. ^Yue L, Zhou D, Xie L, Zhang F, Yan Y, Yin E (2024). "Safe-VLN: Collision Avoidance for Vision-and-Language Navigation of Autonomous Robots Operating in Continuous Environments". IEEE Robotics and Automation Letters. IEEE.

283. ^Ying Z, Liu A, Liang S, Huang L, Guo J, Zhou W, Liu X, Tao D (2024). "SafeBench: A Safety Evaluation Framework for Multimodal Large Language Models". arXiv preprint arXiv:2410.18927.

284. ^Luo W, Ma S, Liu X, Guo X, Xiao C (2024). *"Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks". arXiv preprint arXiv:2404.03027.*

285. ^Shi Y, Gao Y, Lai Y, Wang H, Feng J, He L, Wan J, Chen C, Yu Z, Cao X (2024). *"Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models". arXiv preprint arXiv:2402.04178.*

286. ^Li Y, Guo H, Zhou K, Zhao WX, Wen JR (2024). *"Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models". arXiv preprint arXiv:2403.09792.*

287. a, bNiu Z, Ren H, Gao X, Hua G, Jin R (2024). *"Jailbreaking attack against multimodal large language model". arXiv preprint arXiv:2402.02309.*

288. ^Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ (2024). *"Hallucination of multimodal large language models: A survey". arXiv preprint arXiv:2404.18930.*

289. ^Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, Yu T, Zhang R, Ahmed NK (2024). *"Bias and fairness in large language models: A survey". Computational Linguistics. pages 1–79.*

290. ^Adewumi T, Alkhaled L, Gurung N, van Boven G, Pagliai I (2024). *"Fairness and bias in multimodal ai: A survey". arXiv preprint arXiv:2406.19097.*

291. ^Janghorbani S, De Melo G (2023). *"Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision--Language Models". In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 1725--1735.*

292. ^Wu P, Liu C, Chen C, Li J, Bercea CI, Arcucci R (2024). *"FMBench: Benchmarking Fairness in Multimodal Large Language Models on Medical Tasks". arXiv preprint arXiv:2410.01089.*

293. a, bLuo Y, Shi M, Khan MO, Afzal MM, Huang H, Yuan S, Tian Y, Song L, Kouhana A, Elze T, et al. *Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 12289–12301.*

294. ^Jin R, Xu Z, Zhong Y, Yao Q, Dou Q, Zhou SK, Li X. *"FairMedFM: Fairness Benchmarking for Medical Imaging Foundation Models". In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

295. ^Nayak S, Jain K, Awal R, Reddy S, Steenkiste S, Hendricks L, Stanczak K, Agrawal A (2024). *"Benchmarking Vision Language Models for Cultural Understanding." In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5769–5790.*

296. ^*Wang X, Pan J, Ding L, Biemann C (2024). "Mitigating hallucinations in large vision-language models with instruction contrastive decoding". arXiv preprint arXiv:2403.18715.*

297. a, b*Wang X, Chen J, Wang Z, Zhou Y, Zhou Y, Yao H, Zhou T, Goldstein T, Bhatia P, Huang F, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. arXiv preprint arXiv:2405.15973. 2024.*

298. ^*Zhang L, Yang Q, Agrawal A. "Assessing and Learning Alignment of Unimodal Vision and Language Models". 2024. Available from: https://arxiv.org/abs/2412.04616.*

299. ^*Wang Z, Zhang Z, Liu L, Zhao Y, Huang H, Jin T, Zhao Z (2023). "Extending multi-modal contrastive representations". arXiv preprint arXiv:2310.08884. Available from: https://arxiv.org/abs/2310.08884.*

300. ^*Han J, Gong K, Zhang Y, Wang J, Zhang K, Lin D, Qiao Y, Gao P, Yue X (2024). "Onellm: One framework to align all modalities with language". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 26584–26595.*

301. ^*Lin J, Yin H, Ping W, Molchanov P, Shoeybi M, Han S (2024). "Vila: On pre-training for visual language models". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 26689–26699.*

302. ^*Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021). "Lora: Low-rank adaptation of large language models". arXiv preprint arXiv:2106.09685.*

303. ^*Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2023). "QLoRA: Efficient Finetuning of Quantized LLMs". arXiv. arXiv:2305.14314 [cs.LG].*

304. ^*Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, et al. (2022). "Training a helpful and harmless assistant with reinforcement learning from human feedback". arXiv preprint arXiv:2204.05862.*

305. ^*Lee H, Phatale S, Mansoor H, Lu KR, Mesnard T, Ferret J, Bishop C, Hall E, Carbune V, Rastogi A (2023). "Rlaif: Scaling reinforcement learning from human feedback with ai feedback".*

306. ^*Mu N, Kirillov A, Wagner D, Xie S (2022). "Slip: Self-supervision meets language-image pre-training". In: European conference on computer vision. Springer. p. 529–544.*

307. ^*Balazadeh V, Ataei M, Cheong H, Khasahmadi AH, Krishnan RG (2024). "Synthetic Vision: Training Vision-Language Models to Understand Physics". arXiv. arXiv:2412.08619 [cs.CV].*

308. ^*Sharifzadeh S, Kaplanis C, Pathak S, Kumaran D, Ilic A, Mitrovic J, Blundell C, Banino A. "Synth^2: Boosting Visual-Language Models with Synthetic Captions and Image Embeddings". 2024. Available from: https://arxiv.org/abs/2403.07750.*

309. ^*Tang G, Rajkumar S, Zhou Y, Walke HR, Levine S, Fang K (2024). "Kalie: Fine-tuning vision-language models for open-world manipulation without robot data". arXiv preprint arXiv:2409.14066. Available from:* https://arxiv.org/abs/2409.14066.

310. ^*Chae H, Kim N, Ong KT, Gwak M, Song G, Kim J, Kim S, Lee D, Yeo J (2024). "Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation". arXiv.* arXiv:2410.13232 [cs.CL].

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.