A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges

Zongxia Li* **Xiyang Wu*** **Hongyang Du**
Fuxiao Liu **Huy Nghiem** **Guangyao Shi**

{zli12321, wuxiyang, hdu1, fl3es, nghiemh}@umd.edu
shig@usc.edu

https://github.com/zli12321/Vision-Language-Models-Overview.git

Abstract

Multimodal Vision Language Models (VLMs) have emerged as a transformative topic 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 [194], Claude [11], and GPT-4V [246] demonstrate strong reasoning and understanding abilities on visual and textual data and beat classical single modality vision models on zero-shot classification [94]. With their rapid advancements in research and growing popularity in various applications, we provide a comprehensive survey of VLMs. Specifically, we provide a systematic overview of VLMs in the following aspects: [1] model information of the major VLMs developed up to 2025; [2] the transition of VLM architectures and the newest VLM alignment methods; [3] summary and categorization of the popular benchmarks and evaluation metrics of VLMs; [4] the challenges and issues faced by current VLMs such as hallucination, alignment, and safety.

1. Introduction

Pretrained large language models (LLMs), such as LLaMA [212], GPT-4 [182] have achieved remarkable success across a wide range of NLP tasks [160, 171]. However, as these models continue to scale [177], they face two challenges: (1) The finite supply of high-quality text data [131, 217]; (2) The inherent limitations of single-modality architectures in capturing and processing real-world information that requires understanding the complex relationships between different modalities [66, 83]. 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 [23, 74]. VLMs expand the representational boundaries that have previous confined single-modality approaches, supporting a richer and more contextually informed view of the world [53, 158, 218], such as visual question answering (VQA) [4], autonomous driving [211]. 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 [70, 146]. There are already several reviews on single-modality models [29, 176] 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, and challenges faced by VLMs.

2. State-of-the-Art VLMs

In recent years, leading Artificial Intelligence (AI) organizations are consistently releasing new VLMs [145]. From OpenAI's CLIP [195], Salesforce's BLIP [122], DeepMind's Flamingo [9] to GPT-4V [246] and Gemini [10], these models are becoming larger and more interactive and illustrate the integration of chatbot functionality within VLM frameworks to support multimodality user interaction. The SoTA VLMs from 2019 to the end of 2024 are listed in Table 1 according to the following principal research directions.

Vision-Language correlation considers how training objectives or architectural design facilitate multimodal integration [262]. Training objectives such as contrastive learning are exemplified by approaches like SimCLR [33], 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 [262]. Older architectural approaches primarily train models from scratch (CLIP [247]), whereas more recent methods (LLaMA 3.2-vision [56]) 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).

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 [58]. It is trained to extract rich visual features from image or video data, enabling integration with language representations [159, 267].

Specifically, vision encoders used in many VLMs [36, 44, 143, 222], 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 [194], which aligns images and text embeddings via contrastive learning, and BLIP [123], which leverages bootstrapped pretraining for robust language-image alignment. Pretrained on large scale ImageNet [48] or Similar Datasets: These encoders are trained on vast amounts of labeled visual data or through self-supervised training [185], enabling them to capture domain-specific visual features. While initially unimodal, these encoders, such as ResNet [75] or Vision Transformers (ViTs) [52], 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 [222] and LLaVA [142], commonly incorporate pre-trained vision encoders. These encoders not only provide robust and meaningful visual representations but are also highly effective for transfer learning [256]. They outperform randomly initialized encoders [82] by leveraging learned vision knowledge from their training domains.

Text Encoder projects tokenized text sequences into an embedding space, similar to how vision encoders process images. Models such as CLIP [194], BLIP [123], and ALIGN [97] 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 [142], often do not include a dedicated text encoder. Instead, they rely on large language models (LLMs) (e.g., LLaMA [212], Vicuna [186]) for text understanding, integrating visual inputs through projection layers or cross-attention mechanisms [137]. 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 [106]. GPT-4V [182], Flamingo [8], and Kosmos-2 [188] 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 [127, 153] 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).

Cross-Attention Mechanisms enable visual-text interactions by allowing tokens from one modality (vision) to influence tokens from the other modality (text) [137]. These layers compute attention scores across modalities, but not all models use them. VisualBERT [153] and Flamingo [9] employ cross-attention, while CLIP [194] does not.

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 [76]. Variants of SSL techniques include masked image modeling [77], contrastive learning [215], and image transformation prediction [168]. In this section, we delve into contrastive learning, a common pre-training process to scale up VLM training from scratch.

Contrastive Learning employs separate encoders for visual and textual inputs, mapping them into a shared embedding space. The visual encoder extracts features us-

Model	Year	Architecture	Training Data	Parameters	Vision Encoder / Tokenizer	Pretrained Backbone Model
CLIP [194]	2021	Encoder-decoder	400M image-text pairs	63M-355M	ViT[52] / ResNet[75]	Pretrained from scratch
Flamingo [9]	2022	Decoder-only	M3W [9], ALIGN [97]	80B	Custom	Chinchilla [81]
BLIP [122]/2 [125]	2022/23	Encoder-decoder	COCO [139], Visual Genome [112]	223M-400M	ViT-B/L/g [52]	Pretrained from scratch
GPT-4V [246]	2023	Decoder-only	Undisclosed	Undisclosed	Undisclosed	Undisclosed
Gemini [10]	2023	Decoder-only	Undisclosed	Undisclosed	Undisclosed	Undisclosed
LLaVA-1.5 [144]	2023	Decoder-only	COCO [139]	13B	CLIP ViT-L/14 [52]	Vicuna [1]
PaLM-E [54]	2023	Decoder-only	All robots, WebLI [34]	562B	ViT [52]	PaLM [39]
CogVLM [223]	2023	Encoder-decoder	LAION-2B [231], COYO-700M [27]	18B	CLIP ViT-L/14 [52]	Vicuna [1]
InstructBLIP [43]	2023	Encoder-decoder	CoCo [139], VQAv2 [67]	13B	ViT [52]	Flan-T5 [40], Vicuna [1]
InternVL [36]	2023	Encoder-decoder	LAION-en [200], LAION-multi [200]	7B/20B	Eva CLIP ViT-g [52]	QLLaMA [41]
Claude 3 [11]	2024	Decoder-only	Undisclosed	Undisclosed	Undisclosed	Undisclosed
Emu3 [226]	2024	Decoder-only	Aquila [261]	7B	MoVQGAN[269]	LLaMA-2 [212]
NVLM [44]	2024	Encoder-decoder	LAION-115M [123]	8B-24B	Custom ViT	Qwen-2-Instruct [244]
Qwen2-VL [222]	2024	Decoder-only	Undisclosed	7B-14B	EVA-CLIP ViT-L [52]	Qwen-2 [244]
Pixtral [5]	2024	Decoder-only	Undisclosed	12B	CLIP ViT-L/14 [52]	Mistral Large 2 [169]
LLaMA 3.2 vision [56]	2024	Decoder-only	Undisclosed	11B-90B	CLIP[194]	LLaMA-3.1 [56]
Baichuan Ocean Mini [132]	2024	Decoder-only	Image / Video / Audio / Text	7B	CLIP ViT-L/14 [52]	Baichuan [243]
TransFusion [272]	2024	Encoder-decoder	Undisclosed	7B	VAE Encoder [107]	Pretrained from scratch on transformer architecture
DeepSeek-VL2 [237]	2024	Decoder-only	WiT [207], WikiHow [111]	4.5B x 74	SigLIP [108] / SAMB [259]	DeepSeekMoE [42, 221]
Molmo [47]	2024	Decoder-only	PixMo [47]	1B-72B	CLIP ViT-L/14 [52]	OLMoE [174] / OLMo [68] / Qwen-2 [244]
BLIP-3 [242]	2024	Decoder	OBELICS [113], MINT-1T [13]	4B	ViT [52]	Phi-3-mini [2]
OLMo-2 [181]	2024	Decoder-only	OLMo-mix-1124 [181]	7B-13B	GPT-NeoX-20B [22]	Pretrained from scratch
DeepSeek-Janus-Pro [35]	2025	Decoder-only	Undisclosed	7B	SigLIP-Large-Patch16-384 [259]	Pretrained from scratch
QWen2.5-VL [15]	2025	Decoder-only	VQA/long video	3B/7B/72B	Redesigned ViT [52]	Qwen2.5 [193]
LLaMA 4 [6]	2025	Decoder-only	Undisclosed	17B	-	LLaMA 4 MoE [6]

Table 1. There is a growing number of VLMs released in recent years, which 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.

ing convolutional neural networks (CNN) [183] or vision transformers (ViTs) [51]. 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, while maximizing the distance between embeddings of unrelated pairs. Pioneering models like CLIP [194], BLIP [124], and ALIGN [97] 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 [25, 182, 212] 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 maps visual features extracted by the vision en-

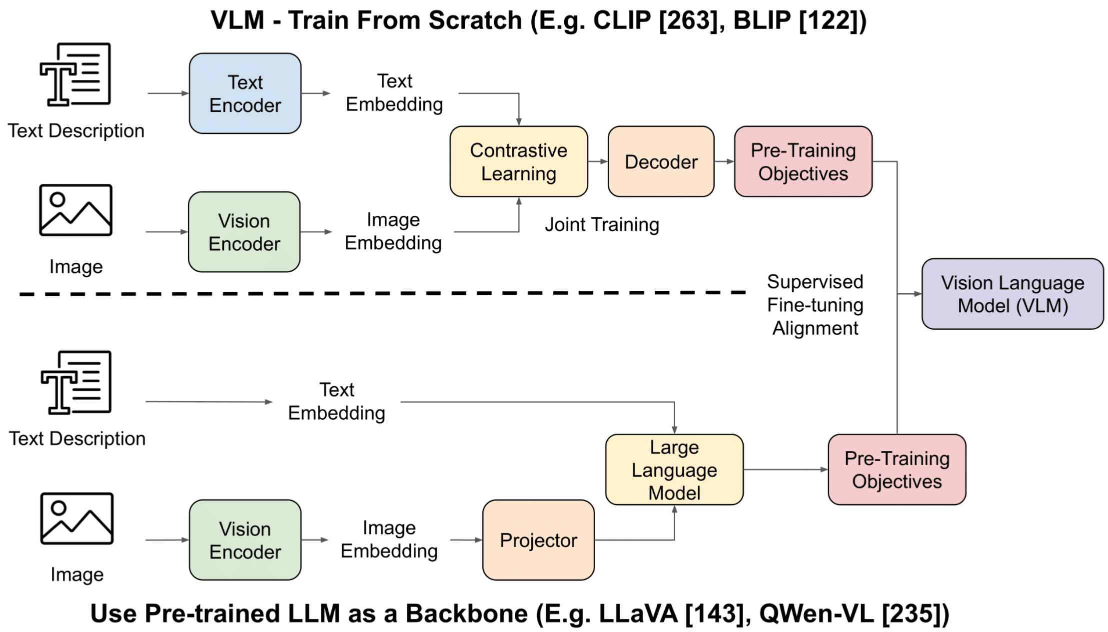


Figure 1. The basic components of common SoTA VLMs are transitioning from joint training from scratch to using a pretrained LLM as the backbone to fully leverage the knowledge of LLMs.

coder into a shared embedding space aligned with the text embeddings from the LLM. It typically consists of multi-layer perceptron (MLP) layers [175], 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 [143], QWen-2-VL [222], Nvidia VLM [44], Baichuan Ocean-mini [132], Emu3 [226], and Pixtral (multimodal decoder) [5] .

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 [9].

Freeze Training Stages involves selectively freezing model components during training, preserving pre-trained knowledge while adapting to new tasks [84]. Common strategies include freezing pre-trained vision encoders while fine-tuning projector layers, and implementing gradual unfreezing of components [189] or freezing LLM layers while only updating vision encoder weights [213].

3.4. Newer Architectures

Recent works have focused on enhancing the fusion of visual and textual features, which we will 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 [227] 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 [212], but comes with an expansion of the embedding layer to accommodate discrete vision tokens (Root Mean Square Layer Normalizatio layer [260] and Multi-query attention [7]). 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 [272]. 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.

3.5. VLM Alignments

Alignment can improve the downstream task performance, safety, and reliability of VLMs. Alignment has been a success in the LLM domain, as demonstrated by examples such as GPT-4 [182] and DeepSeek R1 [45]. The general alignment algorithm is Reinforcement Learning from Hu-

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

man Feedback (RLHF) that uses human annotations to train models to generate responses that align with human preferences and values. Specifically, Direct Preference Optimization (DPO) [196], Proxy Policy Optimization (PPO) [201] align LLMs with human preferred responses to generate outputs that better align with human preferences, where GRPO uses rule-based reward to leverage models' chain-of-thought abilities to solve a problem step by step to improve the model's reasoning ability and final task performance.

Title	Year	Model Size	RL
MM-Eureka [167]	2025	8/38B	RLOO [110]
MM-RLHF [266]	2025	8B	DPO
LMM-R1 [187]	2025	3B	PPO
Vision-R1 [89]	2025	72B	GRPO
R1-VL [264]	2025	2/7B	GRPO
Video-R1 [57]	2025	7B	GRPO

Table 2. A line of recent works show that RL can also improve VLMs downstream reasoning performance on visual math reasoning, video understanding and image understanding.

While RLHF succeeds on LLMs, VLMs' multimodal nature adds additional layers of complexity for alignment. For instance, when a model handles image inputs alongside text, it can reveal or even infer sensitive details about a person in an image or misinterpret visual context. The alignment problems of VLMs are still less considered than those of their text counterparts. RLHF is adopted to VLM and has become one of the most popular and effective ways to align VLMs [167, 266]. The key to RLHF is to collect human feedback and design reward models. In [266], authors introduces a high-quality, human-annotated dataset with 120k preference comparison pairs to enhance the alignment of VLMs. It proposes a Critique-Based Reward Model that improves interpretability by generating critiques before assigning scores. By contrast, [167, 273] extends large-scale rule-based reinforcement learning to multimodal scenarios and reproduces key characteristics of text-based RL (like DeepSeek-R1 [72]) in visual contexts. Despite using only simple, sparse rewards (format and accuracy) and minimal data filtering, the authors achieve stable improvements in accuracy and response length.

In addition to RLHF, Reinforcement Learning with Verifiable Rewards (RLVR) is also getting attention [152]. RLVR bypasses the need for training a reward model by utilizing a direct verification function to evaluate correctness. This method streamlines the reward process while ensuring strong alignment with the task's correctness criteria.

4. Benchmarks and Evaluation

The number of VLM benchmarks has grown rapidly with the quick development of new VLMs since 2022 [37, 263]. Comprehensive benchmarking is important for evaluating model performance and ensuring robust training across different capabilities various aspects such as math reasoning, scene recognition, etc [67, 154]. 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 [61]. In this section, we summarize and categorize existing 54 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 3).

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[254] 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[257], 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.

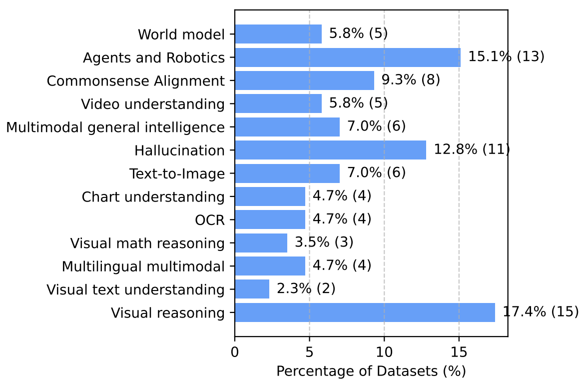
Synthetic question generation has become a more popular part of benchmark generation pipeline on various disciplines such as chart understanding [163], video understanding [162] to quickly scale up dataset sizes. Common practices include using human written examples as seed examples, giving a powerful LLM to generate more adversarial example questions and answers [116]. 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 [126, 162]. However, LLMs are not always accurate and may produce unfaithful content

Category	Description	Datasets
Visual text understanding	Evaluates models' ability to extract and understand texts within visual components	TextVQA [204], DocVQA [165]
Multilingual multimodal understanding	Evaluates VLMs on different languages on different tasks such as question answering and reasoning	MM-En/CN [150], CMMLU [121], C-Eval [90], MTVQA [210]
Visual math reasoning	Tests models' ability to solve math problems in image forms	MathVista [154], MathVision [220], MM-Vet [252]
Optical Character Recognition (OCR)	Test models' ability to extract objects from visual inputs	MM-Vet [252], OCRBench [151], MME [59], MMT-Bench [250]
Chart graphic understanding	Evaluates models' ability to interpret graphic-related data	infographic VQA [164], AI2D [105], ChartQA [163], MMMU [254]
Text-to-Image generation	Evaluates models' ability to generate images	MSCOCO [139], GenEval [65], T2I-CompBench [88], DPG-Bench [87], VQAScore [140], GenAI-Bench [117]
Hallucination	Evaluates whether models are likely to hallucinate on certain visual and textual inputs	HallusionBench [70], POPE [129], CHAIR [198], M-HalDetect [71], Hallu-Pi [50], Halle-Switch [258], BEAF [249], AutoHallusion [236], GAIVE [141], Hal-Eval [98], AMBER [219]
Multimodal general intelligence	Evaluates models' ability on diverse domains of tasks	MMLU [79], MMMU [254], MMStar [32], M3GIA [206], AGIEval [271]
Video understanding	Evaluates models' ability to understand videos (sequences of images)	EgoSchema [162], MLVU [275], MVBench [126], VideoMME [60], MovieChat [205], Perception-Test [191],
Visual reasoning, understanding, recognition, and question answering	Evaluate VLMs' ability to recognize objects, answer questions, and reason through both visual and textual information	MMTBench [250], GQA [92], MM-En/CN [150], VCR [257], VQAv2 [67], MM-Vet [252], MMU [150], SEEDBench [116], Real World QA [238], MMMU-Pro [255], DPG [87] , MSCOCO-30K [139], MM-Vet [252], ST-VQA [21], NaturalBench [118]
Alignment with common sense and physics	Evaluate the alignment between the AIGC images and videos generated by VLMs and the real world	VBench [91], PhysBench [38], VideoPhy [20], WISE [179], VideoScore [78], CRAVE [208], World-SimBench [192], WorldModelBench [120]
Robot benchmark, web agent benchmark	Evaluate the embodied VLMs' abilities online in rule-based simulators or offline datasets recording collected interactions	Habitat [199], Gibson [239], iGibson [119], Isaac Lab [170], WebArena [276], CALVIN [166], VLM-Bench [270], GemBench [64], VIMA-Bench [99], VirtualHome [190], AI2-THOR [109], ProcTHOR [46], ThreeDWorld [63]
Generative model, world model	Evaluate the embodied AI models' abilities with interactive models representing the environments	GAIA-1 [85], UniSim [245], LWM [147], Genesis [12], RoboGen [228]

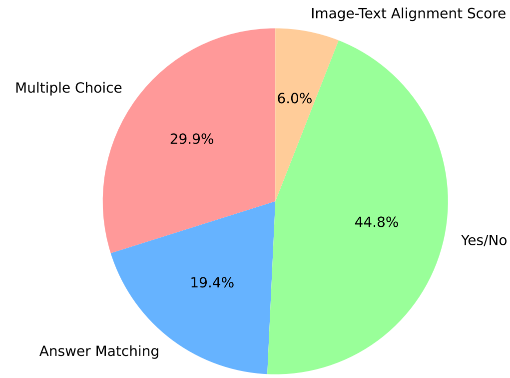
Table 3. We collect 95 benchmarks covering 13 basic categories to evaluate VLMs. However, most of these categories test VLMs' general abilities to understand visual contents, and many of them are still far from practical evaluations in real-world applications, such as scene understanding in autonomous driving [148, 232].

or hallucinations without human supervision [69, 134, 241]. 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 [116, 162, 163]. 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 [70, 172], 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 sense. 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 [64, 119, 166, 170, 199, 239, 270] and web agents [276] to simulate human agents or robots' interactions with the physical world. Nonetheless, benchmarks [119, 199, 239] 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 [245] or even world model [12, 85, 147] 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



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



(b) Majority of the benchmarks are designed in multiple choice or yes/no format for ease of evaluations.

Figure 2. Our surveyed (a) benchmark dataset categories and (b) common evaluation practices. Most existing benchmarks focus on Yes/No and multiple choice format for the ease of evaluation. However, multiple choice and Yes/No questions also have their limitations that VLMs/LLMs can blindly answer above random guessing probability without giving them the questions [19]. Current scope of VLM benchmark and evaluation is broad but not comprehensive due to the challenges of reliability of answer matching.

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, 3).

Answer matching is widely used for open-ended and closed-ended question types, where 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 [133], containment exact match [95] 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 [31, 115]. 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 [26, 30, 133] and becomes impossible for benchmarks that seek long-form answers [240]. Prior to the instruction following success of LLM period, standard token overlapping socres such as F_1, ROUGE [136], BLEU [184] 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 [26, 30, 133, 240].

Thus, some of the benchmarks like MM-Vet [252] adopt 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 [102, 161, 268]. 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 [116, 238, 250, 257]. 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 [19]. 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 [65, 88] to evaluate the alignment between generated images and their corresponding textual descriptions. They often rely on measures such as CLIPScore [80] 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 [79] alone covering 57 distinct tasks. However,

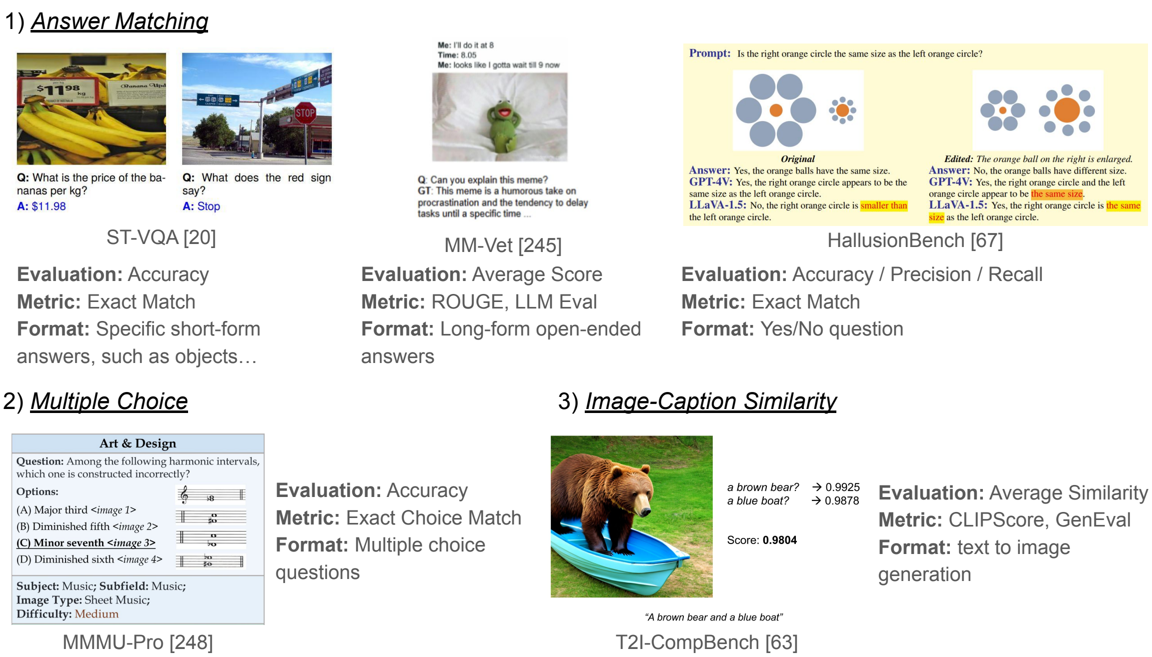


Figure 3. Common benchmark evaluation metrics restrict the formats of most benchmarks, which mostly evaluate whether a VLM can generate a short-form answer that matches the correct answers.

popular evaluations remain largely confined to simple answer matching or multiple choice formats, far from the broader general intelligence of the Turing test [214].

5. Challenges of VLMS

This section examines key challenges in VLM research, including hallucination, safety, fairness, alignment, efficiency in training and fine-tuning, and data scarcity. Despite recent advancements, understanding these limitations is crucial to mitigating risks and ensuring ethical, reliable deployment, particularly for marginalized users.

5.1. Hallucination

Hallucination in VLMs refers to referencing nonexistent objects in images [198]. Despite benchmark-setting performance, hallucination is still a pervasive issue in VLMs, especially in visual-text tasks. Researchers have proposed datasets and metrics to quantify hallucination, with early efforts tending to require human annotation. CHAIR [198] quantifies hallucination in image captioning using per-instance and per-sentence metrics. POPE [128] assesses hallucination with Yes-No questions on object existence. M-HalDetect [71] provides 16K fine-grained VQA samples for training VLMs to detect and prevent hallucinations.

Subsequent research investigated hallucination in finer detail. Halle-Switch [258] evaluates hallucination based on data amount, quality, and granularity, balancing contextual and parametric knowledge. Hallu-Pi [50] provides 1,260 images with detailed annotations to detect perturbation-induced hallucinations. BEAF [249] examines before-and-after image changes, introducing new metrics: true understanding, ignorance, stubbornness, and indecision. HallusionBench [70] tests VLM visual reasoning with dependent, unanswerable questions across diverse topics and formats. AutoHallusion [236] automates hallucination benchmark generation, probing VLM language modules for context-based hallucination examples.

The advent of more sophisticated LLMs has also assisted the development of larger benchmark datasets in this area. GAIVE [141] uses GPT-4 to generate 400K samples across 16 vision-language tasks, covering hallucinations like nonexistent object and knowledge manipulation. Hal-Eval [98] constructs 2M image-caption pairs, leveraging GPT-4 for fine-grained hallucination induction. AMBER [219] is an LLM-free multi-dimensional benchmark designed for generative and discriminative tasks, annotating four types of hallucination.

5.2. Safety

Given VLMs' versatility, safeguarding against unethical use is crucial. Jailbreaking [100] allows malicious circumvention of ethical boundaries, posing risks in robotics and other downstream tasks [55, 103, 197, 235, 253]. SafeBench [251] introduces harmful queries across 23 risk scenarios using an LLM-based jury deliberation framework. MM-

SafetyBench [149] evaluates VLM safety with image-text query pairs in unsafe contexts. JailbreakV [155] introduces 28K malicious image-based queries, testing attack transferability across models. SHIELD [203] evaluates face spoofing and forgery detection using True-False queries in zero- and few-shot settings. HADES [130] exploits gradient updates and adversarial methods to conceal and amplify harmful content, breaking multimodal alignment. imgJP [180] bypasses refusal guardrails using images instead of prompts, demonstrating high transferability across VLMs.

5.3. Fairness

Extensive literature has explored inequities in LLMs and VLMs [17, 62]. Like unimodal LLMs, VLMs show disparate performance, particularly affecting marginalized groups [3, 14, 93]. MMBias [96] presents a human-annotated image dataset targeting bias in religion, nationality, disability, and sexual orientation. FMBench [234] proposes a benchmark using medical images to assess gender, skin tone, and age bias. Harvard-FairVL [156] shows CLIP and BLIP2 favor Asian, Male, and Non-Hispanic groups. FairmedFM [101] integrates 17 datasets to evaluate fairness in medical tasks. CulturalVQA [178] (2,378 image-question pairs) shows better performance for North American cultures and worse performance for African and Islamic ones.

5.4. Alignment

Multi-modality Alignment. The alignment issue in multimodal models arises from contextual deviation between modalities, leading to hallucinations [225]. Many efforts to mitigate this include leveraging VLM reasoning for self-reflection [224] or designing projectors to bridge modalities. SIMA [224] improves L-VLM alignment via self-critique and vision metrics. SAIL [265] aligns pretrained unimodal models for better multimodal learning. Ex-MCR [230] enables paired-data-free semantic alignment using contrastive representation. OneLLM [230] unifies eight modalities to language through a unified encoder and progressive multimodal alignment. SeeTRUE [248] benchmarks text-image alignment, proposing VQA-based and end-to-end classification methods for better misalignment detection and ranking.

Commonsense and Physics Understanding. The LVLMs used for AI-generated content (AIGC) images and videos, sometimes known as World Models [73], like SORA [24] and Veo2 [216], attract much attention throughout the community. However, these LVLMs face challenges in commonsense alignment and physics adherence. Many recent benchmarks and evaluation models aim to address these issues. VBench [91] evaluates video generative models across structured quality dimensions. PhysBench [38] and VideoPhy [20] assess VLMs and text-to-video models on physical understanding and commonsense adherence. WISE [179] introduces WiScore for T2I knowledge-image alignment. CRAVE [208] focuses on AIGC video quality assessment, aligning textual prompts with video dynamics. VideoScore [78] tracks model progress using VideoFeedback, a large-scale human-annotated dataset. WorldSimBench [192] and WorldModelBench [120] evaluate World Simulators for video-action consistency and decision-making applications. GPT4Motion [157] integrates LLMs, physics engines, and diffusion models for physics-aware text-to-video synthesis. Despite these efforts, key challenges remain in advanced video evaluation and bridging the gap between AIGC and real-world fidelity.

Training Efficiency. Efficient training and alignment of VLMs remain a very heated research topic due to their high cost and difficulty in training. Recent studies explore the impact of different pre-training settings over modules [138] or supervision [229] on the ultimate performance of VLMs. However, many applications require specialized rather than multi-task capabilities. Low-Rank Adaptation (LoRa) [49, 86] enables efficient fine-tuning with fewer parameters. RLHF [16, 114] integrates human or model feedback for alignment. Rule-based RL, requiring multiple input generations, increases computational costs, limiting its use to small VLMs [187]. Alternative RL methods (PPO, DPO) reduce computation but demand extensive human annotation to trade for computation resources [167, 266].

5.5. Data Scarcity

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 scarcity 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) [173] that learns the representation automatically from the unlabelled dataset. Another major direction is to use the synthetic data generated by following some rules [18] or utilizing some third-party tools [202]. In VLM, specifically designed for physical world-related purposes, like robotics [209] or web agents [28], another option is to gather datasets from the interactions with the physical simulators or world model [12, 104, 228, 233], or learning from videos with human demonstrations [135, 274]. 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."

6. 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, trends, 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.

7. Limitation

Given the rapid growth of VLM research, our survey is not exhaustive. We focus on the most popular and representative models to provide a comprehensive overview. However, as the field evolves quickly, newer models, benchmarks, and techniques will emerge. While we cannot update this paper continuously, the latest VLM developments will be reflected on our website.

References

[1] Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. Accessed: 2024-12-23. 3

[2] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 3

[3] Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*, 2024. 9

[4] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 1

[5] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. 3, 4

[6] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. Accessed: 2025-04-05. 3

[7] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023. 4

[8] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 2

[9] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022. 1, 2, 3, 4

[10] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1, 2023. 1, 3

[11] Anthropic. Claude: An ai assistant by anthropic, 2024. Accessed: 2024-12-23. 1, 3

[12] Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, 2024. 6, 9

[13] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens, 2024. 3

[14] Rumaisa Azeem, Andrew Hundt, Masoumeh Mansouri, and Martim Brandão. Llm-driven robots risk enacting discrimination, violence, and unlawful actions. *arXiv preprint arXiv:2406.08824*, 2024. 9

[15] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 3

[16] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 9

[17] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 9

[18] Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. Synthetic vision: Training vision-language models to understand physics, 2024. 9

[19] Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7

[20] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 6, 9

[21] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering, 2019. 6

[22] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022. 3

[23] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024. 1

[24] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024. 9

[25] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 3

[26] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022. 7

[27] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 3

[28] Hyungjoo Chae, Namyoung Kim, Kai Tzu iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation, 2024. 9

[29] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024. 1

[30] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China, 2019. Association for Computational Linguistics. 7

[31] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions, 2017. 7

[32] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. 6

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[34] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. 3

[35] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. 3

[36] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 2, 3

[37] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. Instructeval: Towards holistic evaluation of instruction-tuned large language models, 2023. 5

[38] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. 6, 9

[39] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 3

[40] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. 3

[41] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca, 2024. 3

[42] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. 3

[43] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[44] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms, 2024. 2, 3, 4

[45] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 4

[46] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Ad-*

vances in Neural Information Processing Systems, 35: 5982–5994, 2022. 6

[47] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. 3

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[49] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. 9

[50] Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10707–10715, 2024. 6, 8

[51] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2, 3

[53] Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*, 2024. 1

[54] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 3

[55] Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024. 8

[56] Aaron Grattafiori et al. The llama 3 herd of models, 2024. 2, 3

[57] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. 5

[58] Enrico Fini*, Mustafa Shukor*, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Louis Béthune, Zhe Gan, Victor Turrisi, Alexander Toshev, Marcin Eichner, Yinfei Yang, Moin Nabi, Josh Susskind, and Alaaeldin El-Nouby*. Multimodal autoregressive pre-training of large vision encoders, 2024. 2

[59] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 6

[60] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024. 6

[61] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. Mme-survey: A comprehensive survey on evaluation of multimodal llms, 2024. 5

[62] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024. 9

[63] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020. 6

[64] Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy, 2024. 6

[65] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. 6, 7

[66] Charles Goodwin and Johanne Stege Bjørndahl. Why multimodality? why co-operative action?(transcribed by j. philipsen). *Social Interaction. Video-Based Studies of Human Sociality*, 1(2), 2018. 1

[67] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 3, 5, 6

[68] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David

Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hanna Hajishirzi. Olmo: Accelerating the science of language models. *arXiv preprint*, 2024. 3

[69] Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. Large language models are effective human annotation assistants, but not good independent annotators, 2025. 6

[70] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 1, 6, 8

[71] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 6, 8

[72] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[73] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 9

[74] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7: 1430984, 2024. 1

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2, 3

[76] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 2

[77] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2

[78] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 6, 9

[79] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. 6, 7

[80] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 7

[81] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. 3

[82] O.G. Holmberg, N.D. Köhler, T. Martins, 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. 2

[83] Haodong Hong, Sen Wang, Zi Huang, Qi Wu, and Jiajun Liu. Why only text: Empowering vision-and-language navigation with multi-modal prompts. *arXiv preprint arXiv:2406.02208*, 2024. 1

[84] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. 4

[85] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 6

[86] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 9

[87] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. 6

[88] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation, 2023. 6, 7

[89] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. 5

[90] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models, 2023. 6

[91] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6, 9

[92] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 6

[93] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 743–756, 2022. 9

[94] Ashhadul Islam, Md. Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. Pushing boundaries: Exploring zero shot object classification with large multimodal models, 2023. 1

[95] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering, 2021. 7

[96] Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, 2023. 9

[97] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 2, 3

[98] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534, 2024. 6, 8

[99] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2(3):6, 2022. 6

[100] Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*, 2024. 8

[101] Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: Fairness benchmarking for medical imaging foundation models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 9

[102] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 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)*, pages 5591–5606, Toronto, Canada, 2023. Association for Computational Linguistics. 7

[103] Sathwik Karnik, Zhang-Wei Hong, Nishant Abhangi, Yen-Chen Lin, Tsun-Hsuan Wang, Christophe Dupuy, Rahul Gupta, and Pulkit Agrawal. Embodied red teaming for auditing robotic foundation models. *arXiv preprint arXiv:2411.18676*, 2024. 8

[104] Pushkal Katara, Zhou Xian, and Katerina Fragkiadaki. Gen2sim: Scaling up robot learning in simulation with generative models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6672–6679. IEEE, 2024. 9

[105] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. 6

[106] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. 2

[107] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 3

[108] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3

[109] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 6

[110] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! In *DeepRLStructPred@ICLR*, 2019. 5

[111] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset, 2018. 3

[112] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 3

[113] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 3

[114] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023. 9

[115] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. 7

[116] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 5, 6, 7

[117] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. 6

[118] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples, 2024. 6

[119] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson

2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021. 6

[120] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025. 6, 9

[121] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024. 6

[122] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3

[123] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2, 3

[124] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[125] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3

[126] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. 5, 6

[127] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019. 2

[128] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. 8

[129] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 6

[130] Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *arXiv preprint arXiv:2403.09792*, 2024. 9

[131] Yijiang Li, Sucheng Ren, Weipeng Deng, Yuzhi Xu, Ying Gao, Edith Ngai, and Haohan Wang. Beyond finite data: Towards data-free out-of-distribution generalization via extrapola. *arXiv preprint arXiv:2403.05523*, 2024. 1

[132] Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Ocean-omni: To understand the world with omni-modality, 2024. 3, 4

[133] Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA, 2024. Association for Computational Linguistics. 7

[134] Zongxia Li, Lorena Calvo-Bartolomé, Alexander Hoyle, Paiheng Xu, Alden Dima, Juan Francisco Fung, and Jordan Boyd-Graber. Large language models struggle to describe the haystack without human help: Human-in-the-loop evaluation of llms, 2025. 6

[135] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024. 9

[136] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 7

[137] Hezheng Lin, Xing Cheng, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Qing Song, and Wei Yuan. Cat: Cross attention in vision transformer, 2021. 2

[138] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 9

[139] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3, 6

[140] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. 6

[141] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 6, 8

[142] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[143] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 4

[144] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 3

[145] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[146] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei

Peng. A survey on hallucination in large vision-language models, 2024. 1

[147] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2024. 6

[148] Mengyin Liu, Jie Jiang, Chao Zhu, and Xu-Cheng Yin. Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6662–6671, 2023. 6

[149] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024. 9

[150] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 6

[151] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 6

[152] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 5

[153] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. 2

[154] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. 5, 6

[155] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024. 9

[156] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024. 9

[157] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1430–1440, 2024. 9

[158] Maria Lymperaiou and Giorgos Stamou. A survey on knowledge-enhanced multimodal learning. *Artificial Intelligence Review*, 57(10):284, 2024. 1

[159] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding, 2024. 2

[160] Spyros Makridakis, Fotios Petropoulos, and Yanfei Kang. Large language models: Their success and impact. *Forecasting*, 5(3):536–549, 2023. 1

[161] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. *arXiv preprint arXiv:2310.02567*, 2023. 7

[162] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. 5, 6

[163] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. 5, 6

[164] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021. 6

[165] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. 6

[166] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks, 2022. 6

[167] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. Accessed: 2025-03-12. 5, 9

[168] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019. 2

[169] Mistral AI Team. Mistral large 2: A new generation of advanced language models. `https://www.mistral.ai/`, 2024. Accessed: 2024-12-23. 3

[170] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8 (6):3740–3747, 2023. 6

[171] Atif Farid Mohammad, Bryan Clark, and Ramya Hegde. Large language model (llm) & gpt, a monolithic study in generative ai. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, pages 383–388. IEEE, 2023. 1

[172] Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Lee Boyd-Graber. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13342–13375, Miami, Florida, USA, 2024. Association for Computational Linguistics. 6

[173] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 9

[174] Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. Olmoe: Open mixture-of-experts language models, 2024. 3

[175] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5):183–197, 1991. 4

[176] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. 1

[177] Sania Nayab, Giulio Rossolini, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2024. 1

[178] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, 2024. 9

[179] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation, 2025. 6, 9

[180] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024. 9

[181] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2024. 3

[182] OpenAI. Gpt-4 technical report, 2024. 1, 2, 3, 4

[183] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. 3

[184] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318, USA, 2002. Association for Computational Linguistics. 7

[185] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016. 2

[186] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4, 2023. 2

[187] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025. 5, 9

[188] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. 2

[189] Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. To tune or not to tune? adapting pretrained representations to diverse tasks, 2019. 4

[190] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018. 6

[191] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023. 6

[192] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024. 6, 9

[193] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 3

[194] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3

[195] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[196] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct

preference optimization: Your language model is secretly a reward model, 2024. 5

[197] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J Pappas. Jailbreaking llm-controlled robots. *arXiv preprint arXiv:2410.13691*, 2024. 8

[198] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 6, 8

[199] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6

[200] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 3

[201] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 5

[202] Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. Synth^2: Boosting visual-language models with synthetic captions and image embeddings, 2024. 9

[203] Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024. 9

[204] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. 6

[205] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024. 6

[206] Wei Song, Yadong Li, Jianhua Xu, Guowei Wu, Lingfeng Ming, Kexin Yi, Weihua Luo, Houyi Li, Yi Du, Fangda Guo, and Kaicheng Yu. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark, 2024. 6

[207] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 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*, page 2443–2449. ACM, 2021. 3

[208] Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency. *arXiv preprint arXiv:2502.04076*, 2025. 6, 9

[209] Grace Tang, Swetha Rajkumar, Yifei Zhou, Homer Rich Walke, Sergey Levine, and Kuan Fang. Kalie: Fine-tuning vision-language models for open-world manipulation without robot data. *arXiv preprint arXiv:2409.14066*, 2024. 9

[210] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024. 6

[211] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models, 2024. 1

[212] Hugo Touvron, Louis Martin, and et al. Llama 2: Open foundation and fine-tuned chat models, 2023. 1, 2, 3, 4

[213] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models, 2021. 4

[214] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. 8

[215] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 2

[216] Veo-Team. Veo 2. 2024. 9

[217] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*. 1

[218] Fei Wang, Liang Ding, Jun Rao, Ye Liu, Li Shen, and Changxing Ding. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12):1–22, 2024. 1

[219] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 6, 8

[220] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. 6

[221] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts, 2024. 3

[222] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2, 3, 4

[223] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan

Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024. 3

[224] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024. 9

[225] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024. 9

[226] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. 3, 4

[227] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4

[228] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023. 6, 9

[229] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2021. 9

[230] Zehan Wang, Ziang Zhang, Luping Liu, Yang Zhao, Haifeng Huang, Tao Jin, and Zhou Zhao. Extending multi-modal contrastive representations. *arXiv preprint arXiv:2310.08884*, 2023. 9

[231] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b, 2023. 3

[232] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023. 6

[233] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023. 9

[234] Peiran Wu, Che Liu, Canyu Chen, Jun Li, Cosmin I Bercea, and Rossella Arcucci. Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. *arXiv preprint arXiv:2410.01089*, 2024. 9

[235] Xiyang Wu, Souradip Chakraborty, Ruiqi Xian, Jing Liang, Tianrui Guan, Fuxiao Liu, Brian M. Sadler, Dinesh Manocha, and Amrit Singh Bedi. Highlighting the safety concerns of deploying llms/vlms in robotics, 2024. 8

[236] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models, 2024. 6, 8

[237] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 3

[238] X.AI. Grok-1.5v: The multimodal version of our ai. Blog post, 2024. Accessed on [Insert Date of Access]. 6, 7

[239] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018. 6

[240] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. *ArXiv*, abs/2305.18201, 2023. 7

[241] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2024. 6

[242] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models, 2024. 3

[243] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. Baichuan 2: Open large-scale language models, 2023. 3

[244] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu

Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. 3

[245] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 6

[246] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 1, 3

[247] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 2

[248] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619, 2023. 9

[249] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pages 232–248. Springer, 2025. 6, 8

[250] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024. 6, 7

[251] Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*, 2024. 8

[252] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. 6, 7

[253] Lu Yue, Dongliang Zhou, Liang Xie, Feitian Zhang, Ye Yan, and Erwei Yin. Safe-vln: Collision avoidance for vision-and-language navigation of autonomous robots operating in continuous environments. *IEEE Robotics and Automation Letters*, 2024. 8

[254] Xiang 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, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. 5, 6

[255] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024. 6

[256] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning, 2018. 2

[257] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019. 5, 6, 7

[258] Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*, 2023. 6, 8

[259] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 3

[260] Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. 4

[261] Bo-Wen Zhang, Liangdong Wang, Jijie Li, Shuhao Gu, Xinya Wu, Zhengduo Zhang, Boyan Gao, Yulong Ao, and Guang Liu. Aquila2 technical report, 2024. 3

[262] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[263] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2024. 5

[264] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization, 2025. 5

[265] Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models, 2024. 9

[266] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025. 5, 9

[267] Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: A foundational visual encoder for video understanding, 2024. 2

[268] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Large language models are effective table-to-text generators, evaluators, and feedback providers. arXiv preprint arXiv:2305.14987, 2023. 7

[269] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation, 2022. 3

[270] Kaizhi Zheng, Xiaotong Chen, Odest Jenkins, and Xin Eric Wang. VLMbench: A compositional benchmark for vision-and-language manipulation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 6

[271] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. 6

[272] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. 3, 4

[273] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025. 5

[274] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. *arXiv preprint arXiv:2501.14208*, 2025. 9

[275] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding, 2024. 6

[276] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023. 6