The Revolution of Multimodal Large Language Models: A Survey

Davide Caffagni[1]*, **Federico Cocchi**[1,2]*, **Luca Barsellotti**[1]*, **Nicholas Moratelli**[1]*, **Sara Sarto**[1]*, **Lorenzo Baraldi**[2]*, **Lorenzo Baraldi**[1], **Marcella Cornia**[1], and **Rita Cucchiara**[1,3]
[1]University of Modena and Reggio Emilia, Italy
[2]University of Pisa, Italy
[3]IIT-CNR, Italy
[1]{name.surname}@unimore.it [2]{name.surname}@phd.unipi.it

Abstract

Connecting text and visual modalities plays an essential role in generative intelligence. For this reason, inspired by the success of large language models, significant research efforts are being devoted to the development of Multimodal Large Language Models (MLLMs). These models can seamlessly integrate visual and textual modalities, while providing a dialogue-based interface and instruction-following capabilities. In this paper, we provide a comprehensive review of recent visual-based MLLMs, analyzing their architectural choices, multimodal alignment strategies, and training techniques. We also conduct a detailed analysis of these models across a wide range of tasks, including visual grounding, image generation and editing, visual understanding, and domain-specific applications. Additionally, we compile and describe training datasets and evaluation benchmarks, conducting comparisons among existing models in terms of performance and computational requirements. Overall, this survey offers a comprehensive overview of the current state of the art, laying the groundwork for future MLLMs.

1 Introduction

The introduction of the attention operation and the Transformer architecture (Vaswani et al., 2017) has enabled the creation of models capable of handling various modalities on an increasingly large scale. This advancement is largely attributed to the versatility of the operator and the adaptability of the architecture. Initially, this breakthrough was leveraged for language-specific models (Devlin et al., 2018; Brown et al., 2020) but quickly extended to support diverse modalities (Li et al., 2019; Lu et al., 2019) and facilitate their integration within unified embedding spaces (Radford et al., 2021).

The surge in sophisticated Large Language Models (LLMs), particularly their capacity for in-context learning, has encouraged researchers to broaden the scope of these models to encompass multiple modalities, both as inputs and outputs. This expansion has led to the development of cutting-edge models such as GPT-4V (Achiam et al., 2023) and Gemini (Anil et al., 2023), showcasing state-of-the-art performance.

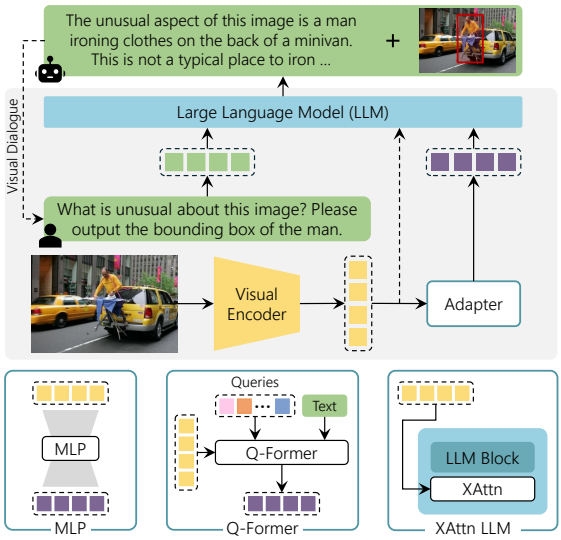

Figure 1: General architecture of Multimodal Large Language Models (MLLMs), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.

The development of Multimodal Large Language Models (MLLMs) entails merging single-modality architectures for vision and language, establishing effective connections between them through vision-to-language adapters, and devising innovative training approaches. These methodologies are crucial for ensuring modality alignment and the ability to follow instructions accurately.

In a context marked by the rapid release of new models, our goal is to offer an exhaustive overview of the MLLM landscape, with a focus on models exploiting the visual modality. This overview serves as both an update on the current state and a source of inspiration for future developments. We

*Equal contribution.

identify three core aspects that define these models: their architecture, training methodologies, and the tasks they are designed to perform. We begin by detailing the prevalent choices for vision encoders and adapter modules that equip LLMs with cross-modal capabilities. Following this, we delve into the training processes and data utilized. We then explore the range of tasks addressed by MLLMs. The review concludes with a discussion of the persisting challenges in the field and the promising directions for future research. Further details on training data, evaluation datasets, performance and computational requirements are reported in the supplementary material.

The motivation behind this survey stems from an emerging scientific interest in the field of MLLMs, as evidenced by the constant increase in published works. In comparison with existing surveys on the topic (Yin et al., 2023a; Wu et al., 2023b; Huang et al., 2023a), our paper exhibits substantial differences. Notably, it addresses several critical areas that were overlooked in prior works, including visual grounding, image generation, and editing. Furthermore, our survey details the main components utilized by each discussed MLLM, such as the visual encoders and the specific LLM employed. Additionally, our analysis offers a comparative perspective on the performance and hardware requirements of the discussed papers, incorporating both quantitative results and detailed information on benchmarks. Through this comprehensive approach, our survey aims to fill the existing gaps and provide a more nuanced understanding of the current landscape in the field.

2 Empowering LLMs with Multimodal Capabilities

2.1 Preliminaries

Large Language Models. Brown et al. (2020) discovered that in-context learning, *i.e.*, prepending the prompt with a few examples demonstrating the desired output of an LLM (Chowdhery et al., 2023; Hoffmann et al., 2022; Tay et al., 2022), improves its performance, especially over unseen tasks. Generalization can be further enhanced by providing the LLM with the natural language description of the desired task for each training sample. This technique, called instruction-tuning (Chung et al., 2022; Wang et al., 2022b,a; Jiang et al., 2024), turns out to be critical for aligning the behavior of an LLM with that of humans and currently empowers the most advanced LLMs, eventually boosted via reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Achiam et al., 2023; Chen et al., 2023l; Bai et al., 2023a).

PEFT. When a pre-trained LLM needs to be adapted to a specific domain or application, parameter-efficient fine-tuning (PEFT) schemes represent an important alternative to train the entire LLM, since these strategies only introduce a few new parameters. Among these, prompt-tuning (Hambardzumyan et al., 2021; Lester et al., 2021; Li and Liang, 2021; Liu et al., 2023j) learns a small set of vectors to be fed to the model as soft prompts before the input text. Differently, LoRA (Hu et al., 2021) constrains the number of new weights by learning low-rank matrices. This technique is orthogonal to quantization methods such as QLoRA (Dettmers et al., 2024), which further decreases the memory footprint of the LLM compared to the usual half-precision weights.

Towards Multimodal LLMs. The development of MLLMs follows a similar path to that of LLMs, with Flamingo (Alayrac et al., 2022) being the first to explore in-context learning at scale in the vision-language field. Then, visual instruction-tuning (Liu et al., 2023e) quickly became the most prominent training paradigm also in the multimodal domain, as well as the use of PEFT techniques to fine-tune the LLM. Any MLLM contains at least three components (Fig. 1): an LLM backbone serving as an interface with the user, one (or more) visual encoders, and one or more vision-to-language adapter modules. Popular choices for the LLM backbone often fall into the LLaMA family (Touvron et al., 2023a,b), given that their weights are freely accessible, they have been trained on public data solely, and they boast different sizes to accommodate various use cases. In addition, their derivative versions are popular as well, such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023). The former fine-tunes LLaMA on instructions written using GPT-3, while the latter exploits user-shared conversations with ChatGPT (OpenAI, 2022). Alternatives are OPT (Zhang et al., 2022b), Magneto (Wang et al., 2023b), MPT (MosaicML, 2023), and the instruction-tuned (Chung et al., 2022) or multilingual (Xue et al., 2020) flavors of T5 (Raffel et al., 2020), an encoder-decoder language model pre-trained for multiple tasks.

Pre-Training of Model Components. The main components of MLLMs are the visual encoder and

the language model. The visual encoder is designed to provide LLMs with visual information and the most used ones are CLIP-based architectures (Radford et al., 2021; Wortsman et al., 2022) whose pre-training objective is the alignment between CLIP embeddings, obtained thanks to a contrastive loss that aligns the correct image-text pairs. An exception is the EVA-CLIP models family (Fang et al., 2023), which exploits a MAE pre-training strategy (He et al., 2022) to reconstruct the masked-out image-text aligned visual features, conditioned on visible image patches. On the other hand, LLMs primarily rely on the widely employed Transformer model, although the Mamba architecture (Gu and Dao, 2023) has also emerged in recent times. This proposes to make a State-Space Model (SSM) time-dependent, effectively creating a selective SSM with favorable properties: (i) inference costs and memory requirements that scale linearly with the sequence length, and (ii) efficient parallel training thanks to a smart GPU implementation of the algorithm. Similar to Transformers, Mamba models for language modeling are pre-trained using the next token prediction task. Very recent studies propose MLLMs featuring Mamba as the language backbone (Qiao et al., 2024; Zhao et al., 2024).

A summary of the MLLMs covered in this survey is reported in Table 1, indicating for each model the LLM on which it is based, the visual encoder, the adapter used to connect visual and language components, whether the MLLM is trained with visual instruction tuning or not, and a short list of the main tasks and capabilities.

2.2 Visual Encoder

In MLLMs, one of the key components is a visual encoder, which is specifically designed to provide the LLM with the visual extracted features. It is common to employ a frozen pre-trained visual encoder while training only a learnable interface that connects visual features with the underlying LLM. While this is usually done using low-resolution images with fixed aspect ratios, some attempts (Xu et al., 2024; Li et al., 2023l) involve adapting pre-trained visual backbones to handle images of different resolutions and aspect ratios. Further details on how to handle higher-resolution images are provided in the supplementary.

The most often employed visual encoders are based on pre-trained Vision Transformer (ViT) models with a CLIP-based objective to exploit the inherent alignment of CLIP embeddings. Popular choices are the ViT-L model from CLIP (Radford et al., 2021), the ViT-H backbone from OpenCLIP (Wortsman et al., 2022), and the ViT-g version from EVA-CLIP (Fang et al., 2023).

As shown in (Li et al., 2023g), a stronger image encoder leads to better performance. Building on this insight, Lin et al. (2023b) and Gao et al. (2024) propose an ensemble of frozen visual backbones to capture robust visual representations and different levels of information granularity. Concurrently, PaLI models (Chen et al., 2023j,h), noticing an imbalance between language and visual parameters, propose scaling the visual backbone respectively to a 4- and 22-billion parameter ViT.

The utilization of such large and powerful models is made feasible by the common practice of maintaining the visual encoder frozen during training, as observed in (Li et al., 2023g; Huang et al., 2023b; Gao et al., 2023; Chen et al., 2023f). However, employing a frozen visual encoder has some limitations, primarily due to the constrained number of parameters, resulting in an inadequate alignment between the visual and language modalities. Specifically, the dense features, extracted from the visual model, may fragment the fine-grained image information and bring large computation due to the lengthy sequence when fed into the language model. To mitigate this issue, other approaches (Ye et al., 2023c,d) employ a two-stage training paradigm. In the first stage, they incorporate a trainable visual backbone while maintaining the pre-trained LLM frozen. According to their findings, enabling the vision encoder to be trainable enhances performance on tasks such as visual question answering or visual description. However, it may lead to performance degradation in other tasks, indicating a degree of forgetting and damage to the general visual representation.

2.3 Vision-to-Language Adapters

The simultaneous presence of inputs from different modalities emphasizes the need to incorporate a module capable of delineating latent correspondences within these unimodal domains. These modules, termed as "adapters", are intended to facilitate interoperability between the visual and textual domains. A spectrum of different adapters are used in common MLLMs, ranging from elementary architectures such as linear layers or MLP to advanced methodologies such as Transformer-based solutions, exemplified by the Q-Former model, and con-

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning	Main Tasks & Capabilities
BLIP-2 (Li et al., 2023g)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗	Visual Dialogue, VQA, Captioning, Retrieval
FROMAGe (Koh et al., 2023b)	OPT-6.7B★	CLIP ViT-L	Linear	✗	Visual Dialogue, Captioning, Retrieval
Kosmos-1 (Huang et al., 2023b)	Magneto-1.3B♢	CLIP ViT-L	Q-Former*	✗	Visual Dialogue, VQA, Captioning
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B▲	CLIP ViT-L	Linear	✗	VQA, Captioning
OpenFlamingo (Awadalla et al., 2023)	MPT-7B★	CLIP ViT-L	XAttn LLM	✗	VQA, Captioning
Flamingo (Alayrac et al., 2022)	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗	Visual Dialogue, VQA, Captioning
PaLI (Chen et al., 2023j)	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗	Multilingual, VQA, Captioning, Retrieval
PaLI-X (Chen et al., 2023h)	UL2-32B♦	ViT-22B	XAttn LLM	✗	Multilingual, VQA, Captioning
LLaVA (Liu et al., 2023e)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
MiniGPT-4 (Zhu et al., 2023a)	Vicuna-13B★	EVA ViT-g	Linear	✓	VQA, Captioning
mPLUG-Owl (Ye et al., 2023c)	LLaMA-7B▲	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA
InstructBLIP (Dai et al., 2023)	Vicuna-13B★	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning
MultiModal-GPT (Gong et al., 2023)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
LaVIN (Luo et al., 2023)	LLaMA-13B▲	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
Otter (Li et al., 2023b)	LLaMA-7B★	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Kosmos-2 (Peng et al., 2023)	Magneto-1.3B♢	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning, Referring, REC
Shikra (Chen et al., 2023f)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Clever Flamingo (Chen et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
SVIT (Zhao et al., 2023a)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
BLIVA (Hu et al., 2024)	Vicuna-7B★	EVA ViT-g	Q-Former+Linear	✓	Visual Dialogue, VQA, Captioning
IDEFICS (Laurençon et al., 2024)	LLaMA-65B★	OpenCLIP ViT-H	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
Qwen-VL (Bai et al., 2023b)	Qwen-7B♦	OpenCLIP ViT-bigG	Q-Former*	✓	Visual Dialogue, Multilingual, VQA, Captioning, REC
StableLLaVA (Li et al., 2023i)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
Ferret (You et al., 2023)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Referring, REC, GroundCap
LLaVA-1.5 (Liu et al., 2023d)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
MiniGPT-v2 (Chen et al., 2023e)	LLaMA-2-7B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Pink (Xuan et al., 2023)	Vicuna-7B▲	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
CogVLM (Wang et al., 2023c)	Vicuna-7B♦	EVA ViT-E	MLP	✓	Visual Dialogue, VQA, Captioning, REC
DRESS (Chen et al., 2023l)	Vicuna-13B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning
LION (Chen et al., 2023d)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former+MLP	✓	Visual Dialogue, VQA, Captioning, REC
mPLUG-Owl2 (Ye et al., 2023d)	LLaMA-2-7B♦	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning
SPHINX (Lin et al., 2023b)	LLaMA-2-13B♦	Mixture	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Honeybee (Cha et al., 2023)	Vicuna-13B♦	CLIP ViT-L	ResNet blocks	✓	Visual Dialogue, VQA, Captioning
VILA (Lin et al., 2023a)	LLaMA-2-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
SPHINX-X (Gao et al., 2024)	Mixtral-8×7B♦	Mixture	Linear	✓	Visual Dialogue, Multilingual, VQA, Captioning, Referring, REC

Table 1: Summary of generalist MLLMs for vision-to-language tasks. For each model, we indicate the LLM used in its best configuration as shown in the original paper (♢: LLM training from scratch; ♦: LLM fine-tuning; ▲: LLM fine-tuning with PEFT techniques; ★: frozen LLM). The * marker indicates variants to the reported vision-to-language adapter, while gray color indicates models not publicly available.

ditioned cross-attention layers added to the LLM.

Linear and MLP Projections. The most straightforward approach for projecting visual inputs into textual embeddings involves learning a linear mapping, which translates visual features to the same dimensionality as the textual counterpart. Some approaches like LLaMA-Adapter (Gao et al., 2023) and FROMAGe (Koh et al., 2023b) only employ a single linear layer to perform the multimodal connection, while LLaVA-1.5 (Liu et al., 2023d) adopts a two-layer MLP, showing improved multimodal capabilities. Despite its widespread adoption in early MLLMs, the use of linear projections has proven highly effective even in recent methods with a more advanced understanding of the visual input (Chen et al., 2023f; Lin et al., 2023a; Wang et al., 2023c; You et al., 2023; Zhao et al., 2023a). It is, therefore, a simple yet effective technique for aligning visual features with textual counterparts. A different approach (Cha et al., 2023) proposes to replace linear layers with convolutional ones, demonstrating moderate improvements.

Q-Former. It is a Transformer-based model proposed in BLIP-2 (Li et al., 2023g) and then used in several other approaches (Chen et al., 2023d; Dai et al., 2023; Hu et al., 2024). It is characterized by its adaptable architecture, which consists of two Transformer blocks that share mutual self-attention layers, facilitating the alignment process between visual and textual representations. It involves a set of learnable queries that interact within the self-attention layers and interface with visual features via a cross-attention mechanism. Textual and visual elements communicate via shared self-attention within the modules.

Drawing inspiration from the Q-Former, various modified versions have been introduced. In this regard, mPLUG-Owl models (Ye et al., 2023c,d) simplify the Q-Former architecture and propose a visual abstractor component that operates by condensing visual information into distinct learnable tokens to derive more semantically enriched visual representations. On the same line, Qwen-VL (Bai et al., 2023b) compresses visual features using a single-layer cross-attention module with learnable queries also incorporating 2D positional encodings.

Additional Cross-Attention Layers. This approach has been proposed in Flamingo (Alayrac

et al., 2022) with the integration of dense cross-attention blocks among the existing pre-trained layers of the LLM. The newly added layers are often combined with a zero-initialized tanh-gating mechanism to ensure that, upon initialization, the conditioned model acts as its original version. The use of additional cross-attention layers imposes the need to train them from scratch, increasing the number of trainable parameters compared to other alternatives. To reduce computational complexity, this strategy is usually paired with a Perceiver-based component (Jaegle et al., 2021) that reduces the number of visual tokens before they are fed to the LLM. Since its introduction, several models (Awadalla et al., 2023; Chen et al., 2023b; Laurençon et al., 2024; Li et al., 2023b) employ this technique to connect the visual modality with the underlying LLM, demonstrating enhanced training stability and improved performance.

2.4 Multimodal Training

Starting from a pre-trained LLM, the training of an MLLM undergoes a single-stage or a two-stage process. In both cases, a standard cross-entropy loss is utilized for predicting the next token, serving as an auto-regressive objective.

Single-Stage Training. This possibility is explored by LLaMA-Adapter (Gao et al., 2023) which introduces additional trainable parameters to encapsulate the visual knowledge and manage text-only instruction learning at the same time. To achieve this, the model undergoes joint training using image-text pairs and instructions, operating on separate parameters. Concurrently, the model proposed in (Koh et al., 2023b) adapts the final loss function by incorporating two contrastive losses for image-text retrieval. During the training, only three linear layers are updated. On a different line, Kosmos-1 (Huang et al., 2023b) considers a frozen visual backbone and trains the language model of 1.3B parameters from scratch.

Flamingo (Alayrac et al., 2022) and its open source variants (Awadalla et al., 2023; Laurençon et al., 2024), instead, train the cross-attention layers and the Perceiver-based component to connect the visual features with the frozen LLM blocks. Additionally, Otter (Li et al., 2023b) extends Flamingo's training to increment its in-context capabilities. Given the amount of training data currently available, approaches like SPHINX-X (Gao et al., 2024) opt to perform a single all-in-one training stage in which to update all model components, possibly also using text-only data to preserve the conversational capabilities of the LLM.

Two-Stage Training. In the first of the two training stages, the objective is to align the image features with the text embedding space. After this stage, the outputs tend to be fragmented and not coherent. Therefore, a second step is done to improve multimodal conversational capabilities. LLaVA (Liu et al., 2023e,d) is among the first to introduce a visual instruction-following training scheme, which is performed as a second training stage updating the parameters of both the multimodal adapter and LLM. During the first stage, instead, only the multimodal adapter is trainable. Differently, MiniGPT-4 (Zhu et al., 2023a) is notable for training solely the linear layer responsible for multimodal alignment across both stages. In the second stage, it uses filtered data, collected and refined through the model itself after the first stage.

Another approach, as demonstrated in InstructBLIP (Dai et al., 2023), involves the freezing of the visual encoder and LLM. In both training stages, only the Q-Former and the connection module are trainable. In contrast to previous approaches where the visual backbone remains frozen, mPLUG-Owl (Ye et al., 2023c,d) updates it in the initial stage, facilitating the capture of both low- and high-level visual information. Also, in the second stage text-only and multimodal data are used jointly to increase alignment. Differently, Shikra (Chen et al., 2023f) updates all weights in both stages, with the only exception of the visual backbone which is kept frozen.

Training Data. During the first (or single) training stage, the datasets predominantly consist of large-scale, publicly available, and uncurated data. For instance, the Conceptual Captions 3M (CC3M) dataset (Sharma et al., 2018) is composed of 3M images paired with textual captions specifically designed for image captioning systems. Unlike the widely-used and curated MS-COCO (Lin et al., 2014) dataset, which serves similar purposes, images and captions in CC3M are gathered from the web, showcasing a broader spectrum of styles and content. Similarly, the LAION family (Schuhmann et al., 2021, 2022) represents an extended collection of non-curated image-text pairs sourced from web pages, providing a rich resource for pre-training multimodal language models. Additionally, the COYO-700M (Byeon et al., 2022) dataset

stands out as a significant resource, containing 747M image-text pairs. Notably, each alt-text in COYO-700M is linked to an image within HTML documents. Furthermore, DataComp (Gadre et al., 2023) presents an extensive pool of 12.8B filtered image-text pairs sourced from common crawl.

It is important to highlight the distinction between datasets used in the initial phase of training, which typically comprise large-scale, uncurated data, and those selected for refinement in subsequent stages. While the former emphasizes diversity and scale, the latter focuses on specificity and task relevance, facilitating a more tailored approach to model optimization. Especially in single-training stage approaches, certain methods (Alayrac et al., 2022; Laurençon et al., 2024) also leverage interleaved datasets, which contain images interleaved with text coming from the web, aiming to augment the dataset size for large models (Hoffmann et al., 2022). Images within these datasets can be positioned at the beginning or in the middle of a sentence, allowing models to support arbitrarily interleaved sequences of images and text as input, thereby enhancing flexibility in input formats by blending textual and visual elements. Among these datasets, the most used are WebLI (Chen et al., 2023j), composed of 10B images and image-text pairs, and MMC4 (Zhu et al., 2023d), an extension of the text-only C4 (Raffel et al., 2020) dataset composed of 365M documents and 156B tokens relatives to different concepts, and OBELICS (Laurençon et al., 2024), an open and curated collection of interleaved image-text web documents, containing 141M documents, 115B text tokens, and 353M images.

In the context of visual instruction tuning, which constitutes the second training stage for MLLMs, the available amount of data is limited. This limitation is mainly due to the creation process which is time-consuming and less well-defined. In this phase, different datasets are used to improve performances on a series of downstream tasks. Among them, LLaVA-Instruct (Liu et al., 2023e) is a collection of GPT-4 generated multimodal instruction-following data. It comprises 158k unique language-image descriptions, spanning various types of tasks including 58k conversations, 23k detailed descriptions, and 77k complex reasoning. Similarly, LRV-Instruction (Liu et al., 2023c) initially consisted of 400k visual instructions generated by GPT-4, and more recently, it has been updated with an additional set of 300k visual instructions. To enhance robustness in instruction tuning, LRV-Instruct also includes negative instructions organized across three semantic levels, showing that instruct-tuned MLLMs on this dataset suffer less from hallucination compared to the original versions. Moreover, LLaVAR (Zhang et al., 2023i) considers publicly available OCR tools to collect results on 422k text-rich images from the LAION dataset. The pipeline first collects 422k noisy text-rich images and then extracts the text through OCR models. With the help of GPT-4, the results and captions are used to create 16k conversations, also including specific questions to create complex instructions which can be helpful to boost performance on new tasks.

3 Tackling Visual Tasks with MLLMs

Standard MLLMs can tackle visual understanding tasks, such as VQA, captioning and multi-turn conversation. However, recently there has been an interest in addressing more fine-grained visual tasks, such as visual grounding and image generation.

3.1 Visual Grounding

The visual grounding capabilities of an MLLM correspond to the ability to carry a dialogue with the user that includes the positioning of the content, also referred to as a referential dialogue (Chen et al., 2023f). In particular, You et al. (2023) introduce *referring* as the ability to understand the content of an input region and can be evaluated on tasks such as region captioning and referring expression generation. Conversely, *grounding* is associated with localizing regions of a given textual description and corresponds to tasks such as referring expression comprehension (REC), referring expression segmentation (RES), phrase grounding, and grounded captioning. Two main components are required to equip MLLMs with these capabilities: a region-to-sequence method to process input regions and a sequence-to-region method to ground nouns and phrases. A summary of the MLLMs with visual grounding capabilities is reported in Table 2.

Region-as-Text. The most common way to output regions is to directly insert them into generated text as a series of coordinates, represented as numbers or as special tokens dedicated to location bins. Shikra (Chen et al., 2023f), Kosmos-2 (Peng et al., 2023), MiniGPT-v2 (Chen et al., 2023e), Ferret (You et al., 2023), CogVLM (Wang et al., 2023c), SPHINX (Lin et al., 2023b), Qwen-VL (Bai et al., 2023b), and Griffon (Zhan et al.,

Model	LLM	Visual Encoder	Supporting Model	Main Tasks & Capabilities
ContextDET (Zang et al., 2023)	OPT-6.7B★	Swin-B	-	Visual Dialogue, VQA, Captioning, Detection, REC, RES
DetGPT (Pi et al., 2023)	Vicuna-13B★	EVA ViT-g	G-DINO★	Visual Dialogue, Detection
VisionLLM (Wang et al., 2023e)	Alpaca-7B▲	Intern-H	Deformable-DETR▲	VQA, Captioning, Detection, Segmentation, REC
BuboGPT (Zhao et al., 2023c)	Vicuna-7B★	EVA ViT-g	RAM, G-DINO, SAM★	Visual Dialogue, Audio Understanding, Captioning, GroundCap
ChatSpot (Zhao et al., 2023b)	Vicuna-7B♦	CLIP ViT-L	-	Visual Dialogue, VQA, Captioning, Referring
GPT4RoI (Zhang et al., 2023g)	LLaVA-7B♦	OpenCLIP ViT-H	-	Visual Dialogue, VQA, Captioning, Referring
ASM (Wang et al., 2023d)	Husky-7B▲	EVA ViT-g	-	VQA, Captioning, Referring
LISA (Lai et al., 2023)	LLaVA-13B▲	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, RES
PVIT (Chen et al., 2023a)	LLaVA-7B♦	CLIP ViT-L	RegionCLIP★	Visual Dialogue, VQA, Captioning, Referring
GLaMM (Rasheed et al., 2023)	Vicuna-7B▲	OpenCLIP ViT-H	SAM♦	Visual Dialogue, Captioning, Referring, REC, RES, GroundCap
Griffon (Zhan et al., 2023)	LLaVA-13B♦	CLIP ViT-L	-	REC, Detection, Phrase Grounding
LLaFS (Zhu et al., 2023c)	CodeLLaMA-7B▲	CLIP RN50	-	Few-Shot Segmentation
NExT-Chat (Zhang et al., 2023a)	Vicuna-7B♦	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, Referring, REC, RES, GroundCap
GSVA (Xia et al., 2023b)	LLaVA-13B▲	CLIP ViT-L	SAM♦	VQA, Segmentation, REC, RES
Lenna (Wei et al., 2023)	LLaVA-7B▲	CLIP ViT-L	G-DINO♦	VQA, Captioning, REC
LISA++ (Yang et al., 2023b)	LLaVA-13B▲	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, RES
LLaVA-G (Zhang et al., 2023d)	Vicuna-13B♦	CLIP ViT-L	OpenSeeD, S-SAM♦	Visual Dialogue, REC, RES, Grounding
PixelLLM (Xu et al., 2023a)	FlanT5-XL-3B▲	EVA ViT-L	SAM★	Referring, REC, RES, GroundCap
PixelLM (Ren et al., 2023b)	LLaVA-7B▲	CLIP ViT-L	-	Visual Dialogue, RES
VistaLLM (Pramanick et al., 2023)	Vicuna-13B♦	EVA	-	Visual Dialogue, VQA, Referring, REC, RES, GroundCap
ChatterBox (Tian et al., 2024b)	LLaVA-13B▲	CLIP ViT-L	iTPN-B★, DINO♦	Visual Dialogue, Referring, REC, GroundCap
GELLA (Qi et al., 2024)	LLaVA-13B▲	CLIP ViT-L	Mask2Former♦	Segmentation, RES, GroundCap
PaLI-3 (Chen et al., 2023i)	UL2-3B♦	SigLIP ViT-g	VQ-VAE♦	VQA, Captioning, Retrieval, RES

Table 2: Summary of MLLMs with components specifically designed for visual grounding and region-level understanding. For each model, we indicate the LLM used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM, and any supporting models used to perform the task (♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ★: frozen). Gray color indicates models not publicly available.

2023) convert bounding boxes into text by indicating two points. VisionLLM (Wang et al., 2023e), VistaLLM (Pramanick et al., 2023), LLaFS (Zhu et al., 2023c), and ChatSpot (Zhao et al., 2023b) allow the MLLM to handle polygons by representing them as a series of points.

Embedding-as-Region. Another solution is to read input regions through region encoders and provide the output regions as embeddings extracted from the last layer of the MLLM to a decoder. For input regions, GLaMM (Rasheed et al., 2023), GPT4RoI (Zhang et al., 2023g), ASM (Wang et al., 2023d) and ChatterBox (Tian et al., 2024b) leverage features of the image encoder to perform ROI align on the bounding box, whereas PVIT (Chen et al., 2023a) exploits RegionCLIP (Zhong et al., 2022). PixelLLM (Xu et al., 2023a) and LLaVA-G (Zhang et al., 2023d) use the prompt encoder of SAM (Kirillov et al., 2023) and Semantic-SAM (Li et al., 2023e) respectively. For output regions, LISA (Lai et al., 2023), GLaMM, GSVA (Xia et al., 2023b), NeXt-Chat (Zhang et al., 2023a), and LISA++ (Yang et al., 2023b) send the embedding corresponding to special tokens to the mask decoder of SAM, LLaVA-G to OpenSeeD (Zhang et al., 2023c), Lenna (Wei et al., 2023) to Grounding-DINO (Liu et al., 2023i), and PixelLM (Ren et al., 2023b) to a custom lightweight pixel decoder.

Differently, ContextDET (Zang et al., 2023) introduces a decoder that receives the latent embedding of the noun with learnable queries, performs a cross-attention with image features, and then uses a segmentation head. ChatterBox (Tian et al., 2024b) combines features from the iTPN-B encoder (Tian et al., 2023) and the MLLM and provides them to the DINO detector (Zhang et al., 2022a). GELLA (Qi et al., 2024) presents a fusion module in Mask2Former (Cheng et al., 2022) to propose masks based on multi-modal image features and an association module to assign latent embeddings to them. PaLI-3 (Chen et al., 2023i) converts embeddings into segmentation masks through a VQ-VAE (Van Den Oord et al., 2017) decoder.

Text-to-Grounding. Other approaches are based on open-vocabulary models that accept textual categories as input. DetGPT (Pi et al., 2023) generates a list of categories for Grounding-DINO. BuboGPT (Zhao et al., 2023c) leverages a combination of RAM, Grounding-DINO, and SAM and matches tags with nouns in the output sequence.

3.2 Image Generation and Editing

While initial MLLMs excelled in extracting information from visual data, recent research included the generation of visual outputs. This advancement is realized through integrating MLLMs with image generation mechanisms, predominantly embodied by the Stable Diffusion (SD) (Rombach et al., 2022) models. These models feature a denoising U-Net (Ronneberger et al., 2015) architecture conditioned on textual or visual embeddings, through cross-attention layers. A complete list of the analyzed models is presented in Table 3.

Model	LLM	Visual Encoder	Supporting Model	Main Tasks & Capabilities
GILL (Koh et al., 2023a)	OPT-6.7B★	CLIP ViT-L	SD v1.5★	Visual Dialogue, Retrieval, Image Generation
Emu (Sun et al., 2023b)	LLaMA-13B♦	EVA ViT-g	SD v1.5♦	Visual Dialogue, VQA, Captioning, Image Generation
SEED (Ge et al., 2023a)	OPT-2.7B▲	EVA ViT-g	SD v1.4★	VQA, Captioning, Image Generation
DreamLLM (Dong et al., 2023)	Vicuna-7B♦	CLIP ViT-L	SD v2.1★	Visual Dialogue, VQA, Captioning, Image Generation, Interleaved Generation
LaVIT (Jin et al., 2023)	LLaMA-7B♦	EVA ViT-g	SD v1.5♦	VQA, Captioning, Image Generation
MGIE (Fu et al., 2024)	LLaVA-7B★	CLIP ViT-L	SD v1.5♦	Image Editing
TextBind (Li et al., 2023f)	LLaMA-2-7B♦	EVA ViT-g	SD XL★	Visual Dialogue, VQA, Captioning, Image Generation
Kosmos-G (Pan et al., 2023)	Magneto-1.3B◇	CLIP ViT-L	SD v1.5★	Image Generation, Compositional Image Generation
MiniGPT-5 (Zheng et al., 2023)	Vicuna-7B▲	EVA ViT-g	SD v2.1★	Visual Dialogue, Image Generation, Interleaved Generation
SEED-LLaMA (Ge et al., 2023b)	LLaMA-2-13B♦	EVA ViT-g	SD unCLIP★	Visual Dialogue, VQA, Captioning, Image Generation, Interleaved Generation
CoDi-2 (Tang et al., 2023)	LLaMA-2-7B▲	ImageBind	SD unCLIP★	Visual Dialogue, Audio Understanding, Image Generation, Image Editing
Emu2 (Sun et al., 2023a)	LLaMA-33B♦	EVA ViT-E	SD XL♦	Visual Dialogue, VQA, Captioning, Image Generation, Image Editing
LLMGA (Xia et al., 2023a)	LLaVA-13B♦	CLIP ViT-L	SD XL♦	Visual Dialogue, VQA, Image Generation, Image Editing
SmartEdit (Huang et al., 2023c)	LLaVA-13B▲	CLIP ViT-L	SD♦	Image Editing
VL-GPT (Zhu et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	SD v1.5★	Visual Dialogue, VQA, Captioning, Image Generation, Image Editing
MM-Interleaved (Tian et al., 2024a)	Vicuna-13B♦	CLIP ViT-L	SD v2.1♦	VQA, Captioning, REC, Image Generation, Interleaved Generation
JAM (Aiello et al., 2024)	LLaMA*-7B♦	-	CM3Leon♦	Image Generation, Interleaved Generation

Table 3: Summary of MLLMs with components specifically designed for image generation and editing. For each model, we indicate the LLM (*: LLM variants) used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM, and any supporting models used to perform the task (◇: training from scratch; ♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ★: frozen). Gray color indicates models not publicly available.

Connecting MLLMs with Diffusion Models. GILL (Koh et al., 2023a) is the pioneer in mapping the output embedding space of a frozen LLM to that of a frozen diffusion model. Specifically, inspired by Q-Former, a mapper component is trained by minimizing the ℓ_2 distance between the image output representation of the language model and the expected conditioning embedding of SD.

While GILL refrains from fine-tuning both the LLM and the diffusion U-Net, alternative methodologies fine-tune the language model to expand its multimodal generation capabilities. In this vein, Kosmos-G (Pan et al., 2023) is developed through a training regime that integrates the output of the LLM with an encoder-decoder structure, leveraging a reconstruction loss and the minimization of the distance within a CLIP-text embedding. Similarly, MiniGPT-5 (Zheng et al., 2023) includes the reconstruction loss of diffusion models in addition to the alignment loss of GILL. Moreover, it divides the overall training process into two distinct phases: the initial phase concentrates on text-to-image generation, while the subsequent is focused on interleaved vision-and-language generation. Distinctly, researchers have studied the alignment of discrete (Jin et al., 2023; Ge et al., 2023a,b) and continuous visual tokens (Zhu et al., 2023b) extracted from input images with the SD conditioning embedding. This is usually achieved by fine-tuning the textual model (Zhu et al., 2023b; Ge et al., 2023a,b). Conversely, Jin et al. (2023) fine-tune both the LLM and the SD U-Net.

A different approach has been studied by Li et al. (2023f) which proposes to fine-tune the LLM by adding two special tokens (*i.e.*, <start> and <end>), and directly encode the generated text between these two tokens using the text encoder in the SD model. Similarly, in (Xia et al., 2023a) the LLM is trained to output detailed language-based generation prompts which are employed for generation or editing tasks. The U-Net is fine-tuned with longer and more detailed textual captions. Furthermore, in DreamLLM (Dong et al., 2023) an alignment loss is eschewed in favor of a score distillation loss while keeping the U-Net frozen. Additional research endeavors have been conducted to introduce MLLMs in the field of image editing (Fu et al., 2024; Huang et al., 2023c; Tang et al., 2023).

End-to-End Pipelines. A different direction is the development of end-to-end training strategies. Specifically, in (Sun et al., 2023b,a) the SD U-Net is directly fine-tuned with the continuous visual embeddings generated by the LLM. Tian et al. (2024a) employ a feature synchronizer, that intervenes in intermediate layers of the LLM and diffusion decoder to cross-attend multi-scale high-resolution image features. Furthermore, end-to-end training approaches have been employed for non-diffusion-based generators, such as VQ-GAN (Esser et al., 2021), as demonstrated in the study by Lu et al. (2023a). Differently, Aiello et al. (2024) propose a methodology to mix an LLM architecture with an autoregressive generator, CM3Leon (Yu et al., 2023a), via bi-directional cross-attention across the architectures of both models.

3.3 Other Modalities and Applications

Video Understanding. Although much of the research focuses on images, some works propose MLLMs specifically designed to handle video sequences. These models process video frames independently, using CLIP-based backbones to extract

frame-level features which are then combined using pooling mechanisms (Li et al., 2023j; Maaz et al., 2023) or Q-Former based solutions (Li et al., 2023h; Ren et al., 2023a). The connection between visual features and the language model mainly follows the same trend as image-based MLLMs, with linear projections being the most common choice. However, there are also some attempts to develop video-specific adapters (Liu et al., 2023g; Ma et al., 2023a) that can capture fine-grained temporal information. In addition to encoding video frames, some works (Munasinghe et al., 2023; Zhang et al., 2023b) also employ audio features to enrich the representation of input video sequences. Furthermore, effective strategies for visual instruction tuning are also designed in the video domain (Song et al., 2024), enabling more effective understanding of long video sequences.

Any-Modality Models. Almost all models described so far treat a single modality as input to the LLM. However, a significant body of work focuses on designing effective solutions that can handle multiple modalities. This is usually achieved by aligning multimodal features through Transformer blocks such as Q-Former (Chen et al., 2023c; Panagopoulou et al., 2023) and Perceiver (Zhao et al., 2023d), or by utilizing ImageBind to effectively extract features that are inherently multimodal (Su et al., 2023). Images, videos, and audio are the most commonly treated modalities. Additionally, some works also effectively encode 3D data (Yin et al., 2023d) and IMU sensor signals (Moon et al., 2023). While all these solutions can manage multimodal inputs, approaches like NExT-GPT (Wu et al., 2023c) and Unified-IO 2 (Lu et al., 2023a) are also capable of generating outputs of different modalities.

Domain-Specific MLLMs. In addition to dealing with generic visual inputs, some research efforts are dedicated to developing MLLMs for specific domains and applications, either training the model starting from a pre-trained LLM or fine-tuning an existing MLLM with domain-specific data. Some examples are MLLMs designed for document analysis and text-intensive visual inputs (Lv et al., 2023; Ye et al., 2023a), those proposed for embodied AI and robotics (Driess et al., 2023; Mu et al., 2023), and those tailored for specific domains such as medicine (Li et al., 2023d) and autonomous driving (Xu et al., 2023c). A complete list of domain-specific MLLMs is reported in the supplementary.

4 Conclusion and Future Directions

In this survey, we have provided a comprehensive overview of the recent evolution of MLLMs, first focusing on how to equip LLMs with multimodal capabilities and then exploring the main tasks addressed by these models. Based on the analysis presented, in the following, we outline important open challenges and promising future research directions to further empower MLLMs.

Multimodal Retrieval-Augmented Generation. While retrieval-augmented generation (RAG) is a consolidated technique in LLMs (Lewis et al., 2020; Asai et al., 2023), its application in MLLMs is still under-explored. We believe that the emergence of VQA datasets that require external retrieved knowledge (Chen et al., 2023k; Mensink et al., 2023) may enable the development of MLLMs with RAG capabilities (Hu et al., 2023b; Caffagni et al., 2024).

Correction of Hallucinations. Several studies (Liu et al., 2023b; Zhu et al., 2023a) show that MLLMs tend to exhibit high hallucination rates, especially when generating longer captions. While some solutions are emerging to mitigate this problem (Liu et al., 2023b; Wang et al., 2023a; Wu et al., 2023d; Yin et al., 2023b; Jing et al., 2023), understanding and correcting the underlying causes of hallucinations remains an important open challenge that is worth addressing to allow the application of these models in more critical contexts (*e.g.*, medicine) and guarantee their accuracy and trustworthiness.

Prevent Harmful and Biased Generation. Ensuring the safety and fairness of large-scale models is of fundamental interest in the community. Recent works show that models trained on web-crawled data are prone to generate inappropriate and biased content. Although recent efforts are being made to reduce this phenomenon in text-to-image generative models (Schramowski et al., 2023; Friedrich et al., 2023; Poppi et al., 2024), further exploration is needed to prevent the same behavior in MLLMs (Pi et al., 2024).

Reduce Computational Load. As shown in the supplementary, MLLMs are highly computationally demanding. Effective strategies (Chu et al., 2024) are needed to reduce computational requirements and enable more accessible development of MLLMs. Possible directions entail reducing training requirements both in terms of model scale and data quantity and optimizing the inference stage.

Limitations

This survey provides a comprehensive review of visual-based MLLMs. Although we have made a significant effort to include all relevant works available to the date of submission, the review might have missed some minor works, and might not have a complete coverage of MLLMs treating modalities that are different from the visual one. Additionally, given the space constraints required by the submission venue, we have restricted our explanations of existing approaches so as to include only the most relevant novelty points. We encourage the reader to refer to the original papers for further technical details and implementation notes.

Acknowledgments

This work has been partially supported by the projects: PNRR-M4C2 (PE00000013) "FAIR - Future Artificial Intelligence Research" funded by the European Commission, the PNRR project "Italian Strengthening of Esfri RI Resilience" (ITSERR) funded by the European Union - NextGenerationEU (CUP B53C22001770006), and the PRIN project "CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content" cofunded by the Italian Ministry of University and Research (CUP B87G22000460001).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *ICCV*.

Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. 2024. Jointly Training Large Autoregressive Multimodal Models. In *ICLR*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023c. TouchStone: Evaluating Vision-Language Models by Language Models. *arXiv preprint arXiv:2308.16890*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. COYO-700M: Image-Text Pair Dataset.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *CVPR Workshops*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced Projector for Multimodal LLM. *arXiv preprint arXiv:2312.06742*.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. 2023a. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. *arXiv preprint arXiv:2308.13437*.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023b. Visual Instruction Tuning with Polite Flamingo. *arXiv preprint arXiv:2307.01003*.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023c. X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. *arXiv preprint arXiv:2305.04160*.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2023d. LION: Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge. *arXiv preprint arXiv:2311.11860*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023e. MiniGPT-v2: Large Language Model As a Unified Interface for Vision-Language Multi-task Learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023f. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2023g. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. *arXiv preprint arXiv:2311.18651*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023h. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023i. PaLI-3 Vision Language Models: Smaller, Faster, Stronger. *arXiv preprint arXiv:2310.09199*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2023j. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *ICLR*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023k. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *EMNLP*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023l. DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback. *arXiv preprint arXiv:2311.10081*.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-Attention Mask Transformer for Universal Image Segmentation. In *CVPR*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *JMLR*, 24(240):1–113.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *arXiv preprint arXiv:2402.03766*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. In *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun,

Hongyu Zhou, Haoran Wei, et al. 2023. DreamLLM: Synergistic Multimodal Comprehension and Creation. *arXiv preprint arXiv:2309.11499*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*.

Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2023. DocPedia: Unleashing the Power of Large Multimodal Model in the Frequency Domain for Versatile Document Understanding. *arXiv preprint arXiv:2311.11810*.

Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. *arXiv preprint arXiv:2302.10893*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2024. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *ICLR*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Hongsheng Li, and Yu Qiao. 2024. SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.05935*.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023a. Planting a SEED of Vision in Large Language Model. *arXiv preprint arXiv:2307.08041*.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023b. Making LLaMA SEE and Draw with SEED Tokenizer. *arXiv preprint arXiv:2310.01218*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *CVPR*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *arXiv preprint arXiv:2305.04790*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. 2023. Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following. *arXiv preprint arXiv:2309.00615*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions From Blind People. In *CVPR*.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. OneLLM: One Framework to Align All Modalities with Language. In *CVPR*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. In *NeurIPS*.

Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. 2023. SCITUNE: Aligning Large Language Models with Scientific Multimodal Instructions. *arXiv preprint arXiv:2307.01139*.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2023a. mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model. *arXiv preprint arXiv:2311.18248*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Wenbo Hu, Yifan Xu, Y Li, W Li, Z Chen, and Z Tu. 2024. BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions. In *AAAI*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023b. REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory. In *CVPR*.

Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, and Shijian Lu. 2023a. Visual Instruction Tuning towards General-Purpose Multimodal Model: A Survey. *arXiv preprint arXiv:2312.16602*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023b. Language Is Not All You Need: Aligning Perception with Language Models. *arXiv preprint arXiv:2302.14045*.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual Storytelling. In *NAACL*.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. 2023c. SmartEdit: Exploring Complex Instruction-based Image Editing with Multimodal Large Language Models. *arXiv preprint arXiv:2312.06739*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M^2UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *ICML*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. 2023. Unified Language-Vision Pretraining with Dynamic Discrete Visual Tokenization. *arXiv preprint arXiv:2309.04669*.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models. *arXiv preprint arXiv:2311.01477*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating Images with Multimodal Language Models. In *NeurIPS*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding Language Models to Images for Multimodal Inputs and Outputs. In *ICML*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV*, 123:32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *IJCV*, 128:1956–1981.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. LISA: Reasoning Segmentation via Large Language Model. *arXiv preprint arXiv:2308.00692*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *NeurIPS*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. 2023a. OtterHD: A High-Resolution Multi-modality Model. *arXiv preprint arXiv:2311.04219*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023c. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023d. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890*.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. 2023e. Semantic-SAM: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*.

Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. 2023f. TextBind: Multi-turn Interleaved Multimodal Instruction-following in the Wild. *arXiv preprint arXiv:2309.08637*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023g. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023h. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *CVPR*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023i. StableLLaVA: Enhanced Visual Instruction Tuning with Synthesized Image-Dialogue Data. *arXiv preprint arXiv:2308.10253*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023j. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. *arXiv preprint arXiv:2311.17043*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023k. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023l. Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models. *arXiv preprint arXiv:2311.06607*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023a. VILA: On Pre-training for Visual Language Models. *arXiv preprint arXiv:2312.07533*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023b. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *arXiv preprint arXiv:2311.07575*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual Spatial Reasoning. *TACL*, 11:635–651.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning Large Multi-Modal Model with Robust Instruction Tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023c. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. *arXiv preprint arXiv:2306.14565*.

Haogeng Liu, Quanzeng You, Xiaotian Han, Yiqi Wang, Bohan Zhai, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. 2024a. InfiMM-HD: A Leap Forward in High-Resolution Multimodal Understanding. *arXiv preprint arXiv:2403.01487*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023d. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual Instruction Tuning. In *NeurIPS*.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023f. Qilin-Med-VL: Towards Chinese Large Vision-Language Model for General Healthcare. *arXiv preprint arXiv:2310.17956*.

Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. 2023g. One For All: Video Conversation is Feasible Without Video Instruction Tuning. *arXiv preprint arXiv:2309.15785*.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023h. LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents. *arXiv preprint arXiv:2311.05437*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023i. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023j. GPT understands, too. *AI Open*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023k. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. 2023l. InternGPT: Solving Vision-Centric Tasks by Interacting with ChatGPT Beyond Language. *arXiv preprint arXiv:2305.05662*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023a. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. *arXiv preprint arXiv:2312.17172*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023b. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *NeurIPS*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. In *NeurIPS*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models. *arXiv preprint arXiv:2305.15023*.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2024. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. *arXiv preprint arXiv:2403.03003*.

Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. 2023. Kosmos-2.5: A Multimodal Literate Model. *arXiv preprint arXiv:2309.11419*.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2023a. Vista-LLaMA: Reliable Video Narrator via Equal Distance to Visual Tokens. *arXiv preprint arXiv:2312.08870*.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2023b. Dolphins: Multimodal Language Model for Driving. *arXiv preprint arXiv:2312.00438*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *ICCV*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model. *arXiv preprint arXiv:2309.16058*.

MosaicML. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. EmbodiedGPT: Vision-Language Pre-Training via Embodied Chain of Thought. *arXiv preprint arXiv:2305.15021*.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. 2023. PG-Video-LLaVA: Pixel Grounding Large Video-Language Models. *arXiv preprint arXiv:2311.13435*.

OpenAI. 2022. Introducing ChatGPT.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2023. Kosmos-G: Generating Images in Context with Multimodal Large Language Models. *arXiv preprint arXiv:2310.02992*.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-InstructBLIP: A Framework for aligning X-Modal instruction-aware representations to LLMs and Emergent Cross-modal Reasoning. *arXiv preprint arXiv:2311.18799*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. DetGPT: Detect What You Need via Reasoning. *arXiv preprint arXiv:2305.14167*.

Renjie Pi, Tianyang Han, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance. *arXiv preprint arXiv:2401.02906*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. *arXiv preprint arXiv:2311.16254*.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. 2023. Jack of All Tasks, Master of Many: Designing General-purpose Coarse-to-Fine Vision-Language Model. *arXiv preprint arXiv:2312.12423*.

Lu Qi, Yi-Wen Chen, Lehan Yang, Tiancheng Shen, Xiangtai Li, Weidong Guo, Yu Xu, and Ming-Hsuan Yang. 2024. Generalizable Entity Grounding via Assistance of Large Language Model. *arXiv preprint arXiv:2402.02555*.

Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. 2024. VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv preprint arXiv:2403.13600*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. 2023. GLaMM : Pixel Grounding Large Multimodal Model. *arXiv preprint arXiv:2311.03356*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023a. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. *arXiv preprint arXiv:2312.02051*.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2023b. PixelLM: Pixel Reasoning with Large Multimodal Model. *arXiv preprint arXiv:2312.02228*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *CVPR*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshops*.

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. 2023. Tiny LVLM-eHub: Early Multimodal Experiments with Bard. *arXiv preprint arXiv:2308.03729*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. 2023. UnIVAL: Unified Model for Image, Video, Audio and Language Tasks. *TMLR*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *CVPR*.

Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. 2024. MovieLLM: Enhancing Long Video Understanding with AI-Generated Movies. *arXiv preprint arXiv:2403.01422*.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One Model To Instruction-Follow Them All. *arXiv preprint arXiv:2305.16355*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023a. Generative Multimodal Models are In-Context Learners. *arXiv preprint arXiv:2312.13286*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative Pretraining in Multimodality. *arXiv preprint arXiv:2307.05222*.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023. CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation. *arXiv preprint arXiv:2311.18775*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford Alpaca: An Instruction-Following LLaMA Model.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying Language Learning Paradigms. *arXiv preprint arXiv:2205.05131*.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. 2024a. MM-Interleaved: Interleaved Image-Text Generative Modeling via Multi-modal Feature Synchronizer. *arXiv preprint arXiv:2401.10208*.

Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. 2024b. ChatterBox: Multi-round Multimodal Referring and Grounding. *arXiv preprint arXiv:2401.13307*.

Yunjie Tian, Lingxi Xie, Zhaozhi Wang, Longhui Wei, Xiaopeng Zhang, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. 2023. Integrally Pre-Trained Transformer Pyramid Networks. In *CVPR*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural Discrete Representation Learning. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *CVPR*.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. 2023a. VIGC: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*.

Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, et al. 2023b. Magneto: A Foundation Transformer. In *ICML*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023c. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*.

Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. 2023d. The All-Seeing Project: Towards Panoptic Visual Recognition and Understanding of the Open World. *arXiv preprint arXiv:2308.01907*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023e. VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. *arXiv preprint arXiv:2305.11175*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. *arXiv preprint arXiv:2204.07705*.

Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xiangxiang Chu. 2023. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2022. Robust Fine-Tuning of Zero-Shot Models. In *CVPR*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. *arXiv preprint arXiv:2303.04671*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023b. Multimodal Large Language Models: A Survey. In *BigData*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023c. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05519*.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. 2023d. See, Say, and Segment: Teaching LMMs to Overcome False Premises. *arXiv preprint arXiv:2312.08366*.

Bin Xia, Shiyin Wang, Yingfan Tao, Yitong Wang, and Jiaya Jia. 2023a. LLMGA: Multimodal Large Language Model based Generation Assistant. *arXiv preprint arXiv:2311.16500*.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. 2023b. GSVA: Generalized Segmentation via Multimodal Large Language Models. *arXiv preprint arXiv:2312.10103*.

Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. 2023a. Pixel Aligned Language Models. *arXiv preprint arXiv:2312.09237*.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2023b. PointLLM: Empowering Large Language Models to Understand Point Clouds. *arXiv preprint arXiv:2308.16911*.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. *arXiv preprint arXiv:2403.11703*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. 2023c. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model. *arXiv preprint arXiv:2310.01412*.

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. 2023. Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs. *arXiv preprint arXiv:2310.00582*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic scene graph generation. In *ECCV*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023a. GPT4Tools: Teaching Large Language Model to Use Tools via Self-instruction. *arXiv preprint arXiv:2305.18752*.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. 2023b. LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2312.17240*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023c. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action. *arXiv preprint arXiv:2303.11381*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023a. mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding. *arXiv preprint arXiv:2307.02499*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023b. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. In *EMNLP*.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2024. CAT: Enhancing Multimodal Large Language Model to Answer Questions in Dynamic Audio-Visual Scenarios. *arXiv preprint arXiv:2403.04640*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023c. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023d. mPLUG-Owl2: Revolutionizing Multimodal Large Language Model with Modality Collaboration. *arXiv preprint arXiv:2311.04257*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023a. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023b. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2023c. FoodLMM: A Versatile Food Assistant using Large Multi-modal Model. *arXiv preprint arXiv:2312.14991*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023d. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. In *NeurIPS*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv:2310.07704*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling Context in Referring Expressions. In *ECCV*.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023a. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning. *arXiv preprint arXiv:2309.02591*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv preprint arXiv:2308.02490*.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 2022. Point-BERT: Pre-Training 3D Point Cloud Transformers With Masked Point Modeling. In *CVPR*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2023. Contextual Object Detection with Multimodal Large Language Models. *arXiv preprint arXiv:2305.18279*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. *arXiv preprint arXiv:2402.12226*.

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. 2023. Griffon: Spelling out All Object Locations at Any Granularity with Large Language Models. *arXiv preprint arXiv:2311.14552*.

Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. 2023a. NExT-Chat: An LMM for Chat, Detection and Segmentation. *arXiv preprint arXiv:2311.04498*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *EMNLP*.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022a. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv preprint arXiv:2203.03605*.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. 2023c. A simple framework for open-vocabulary segmentation and detection. In *CVPR*.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. 2023d. LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models. *arXiv preprint arXiv:2312.02949*.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023e. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *NeurIPS*.

Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. 2023f. Learning 3D Representations From 2D Pre-Trained Models via Image-to-Point Masked Autoencoders. In *CVPR*.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023g. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. *arXiv preprint arXiv:2307.03601*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023h. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *arXiv preprint arXiv:2305.10415*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023i. LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding. *arXiv preprint arXiv:2306.17107*.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. SVIT: Scaling up Visual Instruction Tuning. *arXiv preprint arXiv:2307.04087*.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference. *arXiv preprint arXiv:2403.14520*.

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. 2023b. ChatSpot: Bootstrapping Multimodal LLMs via Precise Referring Instruction Tuning. *arXiv preprint arXiv:2307.09474*.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023c. BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs. *arXiv preprint arXiv:2307.08581*.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023d. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *arXiv preprint arXiv:2305.16103*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens. *arXiv preprint arXiv:2310.02239*.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. RegionCLIP: Region-based Language-Image Pretraining. In *CVPR*.

Bin Zhu, Peng Jin, Munan Ning, Bin Lin, Jinfa Huang, Qi Song, Mingjun Pan, and Li Yuan. 2024. LLMBind: A Unified Modality-Task Integration Framework. *arXiv preprint arXiv:2402.14891*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.

Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. 2023b. VL-GPT: A Generative Pre-trained Transformer for Vision and Language Understanding and Generation. *arXiv preprint arXiv:2312.09251*.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. 2023c. LLaFS: When Large-Language Models Meet Few-Shot Segmentation. *arXiv preprint arXiv:2311.16926*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023d. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *CVPR*.

A Handling Images of Different Resolutions and Aspect Ratios

Most existing MLLMs perceive images in a low resolution and a fixed squared aspect ratio. Some works (Liu et al., 2023d; You et al., 2023; Chen et al., 2023j) have demonstrated that adopting visual backbones trained on higher resolutions leads to fewer hallucinations and improved multimodal understanding abilities, translating into better performance over tasks that require fine-grained details. However, scaling an MLLM to arbitrary input resolutions and aspect ratios raises two important concerns: (i) the adaptation issue of switching from small images seen during training to larger ones at inference time and (ii) computational costs provided by the increased number of tokens in both the visual encoder and the LLM, given by the quadratic complexity of the attention-based architectures. In the following, we distinguish three different approaches to address these problems.

Positional-Encoding Interpolation. These models interpolate the positional encoding of their visual backbones, trained at low resolutions, to handle high-resolution images. While being simple, these methods are prone to adaptation issues. As a consequence, they partially mitigate this issue by performing at least one high-resolution training stage. To reduce the input sequence length to the LLM, and thus the computational cost, MiniGPT-v2 (Chen et al., 2023e) and VILA (Lin et al., 2023a) propose to project multiple visual tokens together into the same token within the embedding space of the LLM. For the same reason, mPLUG-Owl2 (Ye et al., 2023d) and Qwen-VL (Bai et al., 2023b) compress the visual features into fixed-length sequences, independent of the resolution, using learnable queries. The latter further saves computation in most layers of the ViT backbone due to a window attention mechanism.

Sub-Images Slicing. To avoid the adaptation issue, some methods propose to slice a high-resolution image into multiple sub-images of fixed size according to the native resolution of their visual encoder. Then, each sub-image is processed independently by the visual backbone, along with the whole image downsized at the same resolution, and the features are concatenated to obtain the global representation. SPHINX (Lin et al., 2023b) divides the input image in a squared grid of sub-images (*i.e.*, 2×2 or 3×3) at the training resolution of the visual backbone. Moreover, to handle rectangular aspect ratios, SPHINX pads the image to reach the desired square size. For extreme aspect ratios, the padding leads to sub-images which are only composed of padding. Hence, in SPHINX-X (Gao et al., 2024) a skip token is introduced to replace noisy tokens associated with only padding sub-images and reduce the sequence length provided to the LLM, increasing efficiency. Similarly, LLaVA-NeXT (Liu et al., 2024b) ignores sub-images composed only of padding and handles different shapes of grids, by introducing a special token that indicates when a row of sub-images ends. Monkey (Li et al., 2023l) uses a Perceiver-like resampler to extract fixed-length sequences from each sub-image and trains on an image-text dataset curated by several vision expert models integrated by ChatGPT. InfiMM-HD (Liu et al., 2024a) proposes a dynamic resolution adaptation training stage to increase the image size up to 1,344 pixels. It employs gated cross-attention layers (as in Flamingo) to inject the visual features into the LLM, without increasing the input sequence length. LLaVA-UHD (Xu et al., 2024) finds the optimal partitioning scheme leading to sub-images that most resemble the native resolution and aspect ratio of the visual encoder. The number of visual tokens is compressed through a Perceiver-like adapter and employs a spatial schema in such a way that the LLM can understand the grid of sub-images.

Others. Another solution, namely OtterHD (Li et al., 2023a), can seamlessly deal with any resolution or aspect ratio, as it directly feeds large image patches of 30×30 pixels to the LLM, without the need for a visual encoder. LLaVA-HR (Luo et al., 2024), instead, introduces a mixture-of-resolution adaptation to fuse into the ViT layers high-resolution features extracted with a CNN and low-resolution ones produced by the ViT itself.

B Additional Training Data

Specific training datasets are required to empower MLLMs with visual grounding and image gener-

ation capabilities. Here we briefly describe the common choices in this domain.

Visual Grounding. To enable visual grounding, MLLMs can be trained directly on task-specific data using predetermined instruction templates. For instance, CoinIt (Pramanick et al., 2023) is a unified set of 14 benchmarks converted into an instruction-tuning format, spanning from single-image coarse-level to multi-image region-level tasks. An additional training step is usually performed on an instruction-tuning dataset, such as LLaVa-Instruct (Liu et al., 2023h), to preserve the conversational capabilities of the MLLM. However, some methods create their custom datasets to simultaneously improve the grounding and conversational capabilities. Specifically, Shikra (Chen et al., 2023f), DetGPT (Pi et al., 2023), ChatSpot (Zhao et al., 2023b), and PVIT (Chen et al., 2023a) leverage LLMs (Achiam et al., 2023; OpenAI, 2022) to combine regions and captions from datasets that present both annotations (*e.g.*, COCO). Differently, Kosmos-2 (Peng et al., 2023) and Ferret (You et al., 2023) exploit an open-vocabulary detector (Li et al., 2022) to ground noun chunks parsed from captions and then reconstruct referring expressions. ASM (Wang et al., 2023d), GLaMM (Rasheed et al., 2023), and LLaVA-G (Zhang et al., 2023d) propose automated pipelines comprising multiple steps based on off-the-shelf models for generating large corpora of conversations grounded in their corresponding images.

Image Generation and Editing. To perform image generation, datasets containing both textual captions and images are required, as the one mentioned in Sec. 2.4 (*e.g.*, LAION-400M, COYO-700M, and COCO). To enable interleaved text-image generation, MMC4, OBELICS, and VIST (Huang et al., 2016) are popular choices. Instead, for image editing tasks, additional datasets like the one introduced in InstructPix2Pix (Brooks et al., 2023) and MagicBrush (Zhang et al., 2023e) are typically used.

C Evaluation

MLLMs are evaluated across different benchmarks, taking into account both more classic visual comprehension and recognition skills and advanced multimodal conversation capabilities. Table 4 shows the performance of the most common MLLMs on both standard VQA and captioning datasets and benchmarks specifically designed for evaluating MLLMs. In the following, we detail the datasets reported in the table and other benchmarks typically used for the evaluation of MLLMs.

C.1 Standard Benchmarks

One of the most important skills of MLLMs is their ability to effectively answer questions based on the given input image. This ability is quantitatively evaluated across several visual question-answering datasets, measuring the accuracy (Antol et al., 2015) of the answers provided by the MLLM.

VQAv2 (Goyal et al., 2017) is an extended and balanced version of VQA (Antol et al., 2015) built by collecting similar images for the same question, but whose answer is different compared to the original one. This makes it difficult to perform favorably for those models that ignore visual information and only rely on language priors while answering questions. The reported results are related to the test-dev split.

GQA (Hudson and Manning, 2019) is based on Visual Genome scene graph annotations (Krishna et al., 2017) and comprises 113k images and 22M questions focusing on scene understanding and compositionality. We report the results over the test split, which contains 10% of the total images.

OKVQA (Marino et al., 2019) is a benchmark to study how vision-and-language models can address visual questions whose answers cannot be completely found in the image, encouraging systems that also rely on external knowledge. The test set has 14,055 open-ended questions.

VizWiz (Gurari et al., 2018) originates from authentic situations involving individuals with visual impairments who have taken images and articulated accompanying inquiries about them, together with 10 responses. The validation split consists of 4,319 images paired with their corresponding questions, while the test split encompasses roughly 8,000 instances.

ScienceQA (SQA) (Lu et al., 2022) evaluates models over challenging multimodal multiple-choice questions about 3 subjects (*i.e.*, natural science, language science, and social science), 26 topics, 127 categories, and 379 skills. Each question is annotated with explanations linked to relevant lectures. The test set includes 4,241 examples.

Visual Spatial Reasoning (VSR) (Liu et al., 2023a) contains images from COCO, each paired with a caption mentioning two concepts and the spatial relation between them. Models have to choose

Model	VQA					Captioning		MLLM Evaluation						
	VQA^{v2}	GQA	VizWiz	SQA	VQA^{T}	COCO	Flickr	POPE	MME	MMB	SEED	$LLaVA^{W}$	MM-Vet	$Math^{V}$
Flamingo (Alayrac et al., 2022)	82.0	-	**65.7**	-	57.1	138.1	75.4	-	-	-	-	-	-	-
BLIP-2 (Li et al., 2023g)	65.0	41.0	19.6	61.0	42.5	144.5	-	85.3	1293.8	-	46.4	38.1	22.4	-
OpenFlamingo (Awadalla et al., 2023)	52.7	-	27.5	-	24.2	75.9	59.5	-	-	-	-	-	-	-
MiniGPT-4 (Zhu et al., 2023a)	53.7	32.2	-	-	-	-	-	-	581.7	23.0	42.8	45.1	22.1	23.1
mPLUG-Owl (Ye et al., 2023c)	59.5	40.9	-	-	-	-	-	-	967.3	46.6	34.0	-	-	-
ChatBridge (Zhao et al., 2023d)	-	41.8	-	-	-	-	82.5	-	-	-	-	-	-	-
InstructBLIP (Dai et al., 2023)	69.4	49.5	33.4	63.1	50.7	102.2	82.8	78.9	1212.8	36.0	53.4	58.2	25.6	25.3
Shikra (Chen et al., 2023f)	77.4	-	-	-	-	117.5	-	-	-	58.8	-	**79.9**	-	-
Emu (Sun et al., 2023b)	62.0	46.0	38.3	-	-	117.7	-	-	-	-	-	-	36.3	-
SVIT (Zhao et al., 2023a)	80.3	64.1	56.4	70.0	60.8	-	-	-	1565.8	69.1	61.9	-	-	-
BLIVA (Hu et al., 2024)	-	-	42.9	-	58.0	-	87.1	-	**1669.2**	-	-	-	-	-
IDEFICS (Laurençon et al., 2024)	60.0	45.2	36.0	-	30.9	-	-	-	-	54.5	-	-	-	
Qwen-VL (Bai et al., 2023b)	78.2	57.5	38.9	68.2	61.5	120.2	81.0	-	1487.6	60.6	58.2	56.7	-	-
DreamLLM (Dong et al., 2023)	56.6	-	38.1	-	34.9	115.4	-	-	-	49.9	-	-	35.9	-
LLaVA-1.5 (Liu et al., 2023d)	80.0	63.3	53.6	71.6	61.3	-	-	85.9	1531.3	67.7	61.6	70.7	35.4	23.6
CogVLM (Wang et al., 2023c)	82.3	-	-	-	-	**148.7**	**94.9**	87.9	-	**77.6**	72.5	77.8	**51.1**	34.5
LION (Chen et al., 2023d)	-	51.6	-	-	-	139.3	87.1	88.9	-	-	-	-	-	-
mPLUG-Owl2 (Ye et al., 2023d)	79.4	56.1	54.5	-	-	137.3	-	86.2	1450.2	64.5	57.8	25.0	36.2	25.3
SPHINX (Lin et al., 2023b)	80.2	62.9	46.8	69.1	-	-	-	**90.8**	1560.2	67.1	71.6	74.3	36.6	27.5
Emu2 (Sun et al., 2023a)	**84.9**	**65.1**	54.9	-	**66.6**	-	-	-	-	-	62.8	-	48.5	-
Honeybee (Cha et al., 2023)	-	-	-	-	-	-	-	-	1632.0	73.6	68.6	77.5	-	-
Unified-IO 2 (Lu et al., 2023a)	79.4	-	-	**88.7**	-	125.4	-	87.7	-	71.5	61.8	-	-	-
VILA (Lin et al., 2023a)	80.8	63.3	60.6	73.7	**66.6**	115.7	74.2	84.2	1570.1	70.3	62.8	73.0	38.8	
SPHINX-X (Gao et al., 2024)	81.1	63.8	61.9	74.5	-	-	-	89.6	1485.3	71.3	**73.0**	70.2	40.9	**42.7**

Table 4: Performance analysis on 14 evaluation benchmarks for VQA, image captioning, and MLLM evaluation. Best scores are in bold, second best are underlined.

if a given caption is true or false according to the picture. MLLMs are typically evaluated on the 616 samples from the zero-shot test split.

IconQA (Lu et al., 2021) tests the visual reasoning abilities of vision-and-language models on three types of questions: multiple-image-choice, multiple-text-choice, and fill-in-the-blank. The dataset stems from real-world problems found in math textbooks and focuses on abstract images (*i.e.*, icons). There are 107,439 questions, 20% of which makes up for the test split.

TextVQA (VQA^{T}) (Singh et al., 2019) is a dataset based on pictures from Open Images (Kuznetsova et al., 2020) and challenges OCR capabilities of vision-and-language models. The test set comprises 5,734 examples.

OCR-VQA (Mishra et al., 2019) presents a new task in visual question answering by interpreting text within images and involves a collection of 207,572 images of book covers, accompanied by more than 1M question-answer pairs.

Comprehensively describing the visual input is another important skill desired in MLLMs. To evaluate this, various image captioning datasets are commonly employed. As regards the evaluation metric, the CIDEr score (Vedantam et al., 2015), which is the reference metric for the task, is used to compare generated image descriptions with ground-truth captions.

COCO (Lin et al., 2014) contains more than 120k images, each associated with five human-generated captions. For captioning tasks, the splits defined by Karpathy and Fei-Fei (2015) are typically employed, with 113k, 5k, and 5k images respectively for train, validation and test.

Flickr30k (Young et al., 2014) comprises 31,783 images, depicting diverse everyday activities, events, and scenes. Complementing these images are 158,915 captions, obtained through crowd-sourcing techniques.

nocaps (Agrawal et al., 2019) represents a benchmark for novel object captioning, boasting an extensive collection of almost 400 novel object categories compared to the COCO dataset. The validation and test sets include approximately 4.5k and 10.6k images, obtained from Open Images (Kuznetsova et al., 2020). Each image is annotated with 11 human-generated captions. Both validation and test sets are further categorized into in-domain, near-domain, and out-of-domain, where images from the out-of-domain subset contain object categories that are never present in COCO.

TextCaps (Sidorov et al., 2020) includes 145k captions aligned with 28k images. The goal is to recognize and understand the text in images and provide an effective caption that describes the entire visual content. This requires the model to possess OCR capabilities along with image description skills.

C.2 MLLM-Specific Benchmarks

Thoroughly evaluating MLLMs is challenging and remains an open frontier. While evaluating on standard datasets represents a valid choice, many benchmarks designed for MLLMs have been recently proposed. They require very strong perception and cognitive skills to succeed, and often they query for deep domain-specific knowledge. To facilitate the evaluation, many works propose to leverage state-of-the-art proprietary models (*e.g.*, ChatGPT (OpenAI, 2022), GPT-4 (Achiam et al., 2023)) to automatically judge candidate answers. In Table 4, we report the performance of some models on a subset of these new benchmarks.

POPE (Li et al., 2023k) is a valuable benchmark for evaluating object hallucination challenges within MLLMs. This dataset encompasses several distinct subsets, namely random, popular, and adversarial, which are generated utilizing a variety of sampling methodologies. Cumulatively, it is a binary classification query dataset that comprises 8,910 entries, facilitating comprehensive investigations into the phenomenon of object hallucination within the context of MLLMs.

MME (Fu et al., 2023) is an evaluation benchmark that aims to assess proficiency in various communication modalities through 14 tasks covering comprehension and manipulation across modalities like quantification, spatial determination, color identification, and others.

MMBench (MMB) (Liu et al., 2023k) includes approximately 3,000 multiple-choice questions, distributed across 20 distinct domains. Questions are curated to evaluate the efficacy of MLLM across diverse task paradigms. These competencies are systematically arranged into a hierarchical taxonomy, delineating overarching categories such as perception and reasoning, while also outlining granular capabilities including object localization and attribute inference.

SEED-Bench (SEED) (Li et al., 2023c) is specifically designed to evaluate LLMs and MLLMs across 12 dimensions spanning from scene understanding to OCR and action recognition. The benchmark consists of 19k multiple-choice questions written by human annotators.

LLaVA-Bench (LLaVA^{W}) (Liu et al., 2023e) comprehends 24 images with 60 manually-curated questions, including indoor and outdoor scenes, memes, paintings, and sketches. GPT-4 is used to generate the reference solutions and score given answers.

MM-Vet (Yu et al., 2023b) evaluates MLLMs over 16 tasks covering six fundamental vision-and-language capabilities such as recognition, OCR, knowledge, language generation, spatial awareness, and math. The benchmark comprises 200 images and 218 questions. The evaluation scores are obtained from GPT-4 by few-shot prompting.

MathVista (Math^{V}) (Lu et al., 2023b) probes the mathematical reasoning skills of MLLMs for visual question answering. There are 6,141 questions, but only 5,141 are used for evaluation. Before computing the accuracy, the authors propose to parse the answers using an LLM such as GPT-4.

MMMU (Yue et al., 2023) is a challenging benchmark targeting domain-specific knowledge of multimodal models. It consists of 10.5k test samples drawn from university textbooks or online courses spanning six main disciplines. Questions may contain multiple images interleaved with text. Exact matching and word matching are used to assess the correctness of an answer for multiple-choice and open-ended questions respectively. Models are evaluated on zero or few-shot settings.

Tiny LVLM (Shao et al., 2023) focuses on six multimodal capabilities distributed among 2.1k image-question pairs. It introduces a new evaluation metric called ChatGPT ensemble evaluation (CEE). In practice, given the question and the ground-truth solution, ChatGPT is queried with five different prompts to assign the candidate answer either 0 or 1, and the scores are eventually ensembled.

TouchStone (Bai et al., 2023c) is a visual dialog benchmark with manually annotated open-world images, totaling 908 questions corresponding to five major categories of abilities and 27 sub-tasks. The evaluation score is computed by an LLM such as GPT-4, which is asked to compare a candidate answer with a reference one. The latter is computed by GPT-4 itself, with fine-grained annotations of the query image being part of the prompt.

C.3 Visual Grounding Evaluation

The assessment of visual grounding capabilities of MLLMs comprises a variety of standard referring tasks, including region captioning, referring expression generation (REG), and region-level question answering, as well as grounding tasks like referring expression comprehension (REC), referring expression segmentation (RES) and grounded captioning. As regards evaluation metrics, for

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
Kosmos-2 (Peng et al., 2023)	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7
Shikra (Chen et al., 2023f)	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2
Qwen-VL (Bai et al., 2023b)	88.6	92.3	84.5	82.8	88.6	76.8	86.0	86.3
Ferret (You et al., 2023)	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
MiniGPT-v2 (Chen et al., 2023e)	88.7	91.7	85.3	80.0	85.1	74.5	84.4	84.7
CogVLM (Wang et al., 2023c)	**92.8**	**94.8**	**89.0**	**88.7**	**92.9**	83.4	89.8	**90.8**
Griffon (Zhan et al., 2023)	90.1	93.4	86.1	84.8	90.5	77.8	86.1	87.2
LION (Chen et al., 2023d)	89.8	93.0	85.6	84.0	89.2	78.1	85.5	85.7
NExT-Chat (Zhang et al., 2023a)	85.5	90.0	77.9	77.2	84.5	68.0	80.1	79.8
SPHINX (Lin et al., 2023b)	91.0	92.7	86.6	86.6	91.1	80.4	88.2	88.4
Lenna (Wei et al., 2023)	90.3	93.2	87.0	88.1	90.1	**84.0**	**90.3**	90.3
LLaVA-G (Zhang et al., 2023d)	89.2	-	-	81.7	-	-	84.8	-
Unified-IO 2 (Lu et al., 2023a)	90.7	-	-	83.1	-	-	86.6	-
MM-Interleaved (Tian et al., 2024a)	89.9	92.6	86.5	83.0	88.6	77.1	85.2	84.9
SPHINX-X (Gao et al., 2024)	90.6	93.7	86.9	85.5	90.5	79.9	88.3	88.5

Table 5: Performance analysis on the RefCOCO benchmarks for referring expression comprehension (REC). Best scores are in bold, second best are underlined.

REC the accuracy is computed by assuming as correct predictions the ones that correspond to an intersection over union with the ground-truth above 0.5 (Acc@0.5). For referring expression segmentation the cumulative intersection over union (cIoU) is considered, while for region captioning METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015) are commonly used. However, few methods introduce their own benchmarks to evaluate the performance in more realistic scenarios, with grounded conversations that may involve multiple rounds. Quantitative results on the REC, RES, and region captioning tasks are respectively reported in Table 5, Table 6, and Table 7.

RefCOCO and RefCOCO+ (Mao et al., 2016) are collections of referring expressions based on images from the COCO dataset. They were gathered through the ReferItGame (Kazemzadeh et al., 2014), a two-player game where the first player examines an image featuring a segmented target object and formulates a natural language description referring to that object. The second player, who has access only to the image and the referring expression, selects the corresponding object. Players swap roles if they perform correctly, otherwise they receive a new object and image for description. The RefCOCO dataset has no constraints on the natural language and consists of 142,209 expressions for 50,000 objects across 19,994 images. Instead, in the RefCOCO+ players are disallowed from using location words in their referring expressions and it has 141,564 expressions for 49,856 objects in 19,992 images. Evaluation is performed on 1,500, 750, and 750 images corresponding to the validation, testA, and testB splits for both datasets.

RefCOCOg (Yu et al., 2016) was collected by a set of annotators who wrote natural language refer-

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
LISA (Lai et al., 2023)	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GLaMM (Rasheed et al., 2023)	**79.5**	**83.2**	**76.9**	**72.6**	**78.7**	**64.6**	74.2	74.9
NExT-Chat (Zhang et al., 2023a)	74.7	78.9	69.5	65.1	71.9	56.7	67.0	67.0
GSVA (Xia et al., 2023b)	79.2	81.7	77.1	70.3	73.8	63.6	**75.7**	**77.0**
LLaVA-G (Zhang et al., 2023d)	77.1	-	-	68.8	-	-	71.5	-
PixelLLM (Xu et al., 2023a)	76.9	78.5	74.4	69.2	72.1	64.5	70.7	72.4
GELLA (Qi et al., 2024)	76.7	80.5	73.6	67.0	73.2	60.6	70.4	71.5

Table 6: Performance analysis on the RefCOCO benchmarks for referring expression segmentation (RES). Best scores are in bold, second best are underlined.

Model	RefCOCO		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
Kosmos-2 (Peng et al., 2023)	14.1	62.3	-	-
GPT4RoI (Zhang et al., 2023g)	-	-	17.4	145.2
ASM (Wang et al., 2023d)	**20.8**	103.0	18.0	145.1
GLaMM (Rasheed et al., 2023)	16.2	**106.0**	19.7	**180.5**
NExT-Chat (Zhang et al., 2023a)	13.6	79.6	-	-
PixelLLM (Xu et al., 2023a)	14.3	82.3	**19.9**	148.9

Table 7: Performance analysis on the RefCOCO and Visual Genome benchmarks for region captioning. Best scores are in bold, second best are underlined.

ring expressions for objects in COCO images, and another set of annotators who selected objects corresponding to given referring expressions. When a selected object was correct, the corresponding referring expression was inserted in the dataset. It consists of 85,474 referring expressions for 54,822 objects in 26,711 images. Evaluation is carried out on 1,300 and 2,600 images corresponding to the validation and test splits.

Visual Genome (Krishna et al., 2017) connects structured image concepts to language and comprises 108,077 images along with detailed descriptions of all objects present in them, providing 5.4M region descriptions and 1.7M visual question-answer pairs. This dataset is typically used for region-level captioning and question-answering.

Visual7W (Zhu et al., 2016) is a visual question-answering dataset that combines textual descriptions with image regions through object-level grounding. It comprises 328k question-answer pairs on 47k COCO images, together with 1.3M human-generated multiple-choice and more than 560k object groundings from 36,579 categories.

GRIT (Peng et al., 2023) is a large-scale dataset of grounded image-text pairs (*i.e.*, noun phrases or referring expressions associated with regions of the image) based on a subset of COYO-700M and LAION-2B. The construction pipeline consists of two steps: (i) extracting noun chunks from the captions and grounding them to bounding boxes with an open-vocabulary detector (*e.g.*, GLIP); (ii) expanding the noun chunks to referring expressions

by exploiting their dependency relations in the original caption. The resulting dataset comprises 91M images, 115M text spans, and 137M associated bounding boxes.

ReasonSeg (Lai et al., 2023) is a benchmark introduced for the reasoning segmentation task, which consists of providing segmentation masks for complex and implicit query texts. Images are from OpenImages (Kuznetsova et al., 2020) and ScanNetv2 (Dai et al., 2017) and are annotated with text instructions and corresponding segmentation masks. The resulting dataset comprises 1,218 image-instruction pairs. Evaluation metrics are the same as the RES standard benchmark. Two extended variants, ReasonDet (Wei et al., 2023) and ReasonSeg-Inst (Yang et al., 2023b), are respectively introduced for reasoning detection and reasoning instance segmentation tasks.

Grounding-anything Dataset (GranD) (Rasheed et al., 2023) is a dataset designed for the grounded conversation generation (GCG) task, which aims to construct image-level captions with phrases associated with segmentation masks in the image. This dataset was built with an automated annotation pipeline composed of four stages: (i) object localization with the corresponding semantic label, segmentation mask, attributes, and depth information, (ii) extracting relationships between detected objects, (iii) combining previously collected relations to produce dense captions, (iv) enriching captions with contextual information. It comprises annotations for 11M SAM (Kirillov et al., 2023) images. Another dataset, GranD$_f$, is introduced for further fine-tuning and evaluating over the GCG task. It was gathered by extending Flickr30k (Young et al., 2014), RefCOCOg, and PSG (Yang et al., 2022) through GPT-4 and by manually annotating a set of samples. It comprises 214k image-grounded text pairs with 2.5k validation and 5k test samples. Evaluation metrics include METEOR and CIDEr for captioning, class-agnostic mask AP for grounding, intersection over union for segmentation, and mask recall for grounded captioning.

Grounded-Bench (Zhang et al., 2023d) is a benchmark introduced to assess the capabilities of an MLLM in carrying a grounded visual chat. It is built on top of the LLaVA-Bench (Liu et al., 2023h), comprising conversational data generated with GPT-4 and instance annotations from COCO. It is expanded using 1,000 images with 7,000 entities from COCO annotated through an automated pipeline that involves GPT-4 to associate noun phrases from captions to ground-truth instances.

Model	COCO		
	FID	CLIP-I	CLIP-T
Stable Diffusion (Rombach et al., 2022)	9.22	0.667	0.302
Stable Diffusion XL (Podell et al., 2023)	-	0.674	0.310
GILL (Koh et al., 2023a)	12.20	0.684	-
Emu (Sun et al., 2023b)	11.66	0.656	0.286
SEED (Ge et al., 2023a)	-	0.682	-
DreamLLM (Dong et al., 2023)	8.46	-	-
LaVIT (Jin et al., 2023)	**7.40**	-	-
NExT-GPT (Wu et al., 2023c)	11.28	-	-
Kosmos-G (Pan et al., 2023)	10.99	-	-
SEED-LLaMa (Ge et al., 2023b)	-	**0.707**	-
Emu2 (Sun et al., 2023a)	-	0.686	**0.297**
VL-GPT (Zhu et al., 2023b)	11.53	-	-
Unified-IO 2 (Lu et al., 2023a)	13.39	-	-
MM-Interleaved (Tian et al., 2024a)	7.90	-	-

Table 8: Image generation results on the COCO dataset. Best scores are in bold, second best are underlined.

Model	MagicBrush		
	DINO	CLIP-I	CLIP-T
InstructPix2Pix (Brooks et al., 2023)	0.698	0.854	0.292
MagicBrush (Zhang et al., 2023e)	0.868	0.934	0.302
MGIE (Fu et al., 2024)	**0.903**	**0.943**	**0.317**
SmartEdit (Huang et al., 2023c)	0.815	0.914	0.305

Table 9: Image editing results on the MagicBrush benchmark.

MUSE (Ren et al., 2023b) is a multi-target reasoning segmentation dataset. It was created with an automated pipeline on top of 910k instance segmentation masks from the LVIS dataset (Gupta et al., 2019) by exploiting GPT-4V to combine instance categories with natural language descriptions. The resulting dataset comprises 246k question-answer pairs, averaging 3.7 targets per answer.

ChatterBox-300k (Tian et al., 2024b) is a benchmark established to evaluate models on multimodal dialogue systems in multi-round referring and grounding. The dataset is built on images from Visual Genome (Krishna et al., 2017) providing bounding boxes, object relationships, and object attributes information to GPT-4 to generate question-answer pairs.

C.4 Image Generation and Editing Evaluation

To evaluate image generation and editing results, a set of different benchmarks is usually utilized. In terms of evaluation metrics, Fréchet Inception Distance (FID) (Heusel et al., 2017) is the reference metric to evaluate generated images. It quantitatively assesses the congruence between the distribution of synthetically generated images and the distribution of real ones. A diminution in the FID score indicates an enhanced alignment between the

Model	DreamBench		
	DINO	CLIP-I	CLIP-T
DreamBooth (Ruiz et al., 2023)	0.668	0.803	0.305
Kosmos-G (Pan et al., 2023)	0.694	0.847	0.287
CoDi-2 (Tang et al., 2023)	0.703	**0.852**	**0.311**
Emu2 (Sun et al., 2023a)	**0.766**	0.850	0.287

Table 10: Subject-driven image generation results on the DreamBench dataset.

two distributions, denoting a superior visual quality and realism within the generated images.

Other metrics measure the coherence of the generated image with the input prompt and the real ground-truth image corresponding to it. Specifically, CLIP-I and DINO scores consist of computing the cosine similarity between generated and ground-truth images leveraging CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) as visual backbones. CLIP-T, instead, measures image-text alignment through cosine similarity between input captions and generated images, using CLIP to encode both images and textual prompts.

COCO is employed for evaluating text-to-image generation. The evaluation is conducted using either the original validation set comprising 41k samples or a subset of 30k samples randomly selected from the same set. Results on this dataset of MLLMs with image generation capabilities are reported in Table 8.

VIST (Huang et al., 2016) is specifically curated for the task of interleaved image-text generation. It includes 34k and 5k samples for training and evaluation. Each sample is a sequence consisting of 5 images accompanied by 5 textual narratives that collectively form a coherent story.

MagicBrush (Zhang et al., 2023e) is a benchmark in the area of image editing and contains a collection of 10,000 manually annotated triplets, each consisting of a source image, an editing instruction, and the corresponding target image. Performances on this benchmark are reported in Table 9.

DreamBench (Ruiz et al., 2023) is a benchmark that evaluates the generative capabilities of the models on subject-driven generation. Specifically, it contains 30 subjects, each illustrated with 4 to 6 images, and 25 template prompts enabling modification and accessorization of the given subjects. Results on this benchmark are shown in Table 10.

D Computational Requirements

To provide a quantification of the computational requirements necessary to train an MLLM, we compare some of the most common models in Table 11 and indicate for each of them the type and number of GPUs/TPUs employed during training. Except for Flamingo and PaLI, which are trained on a large amount of TPUs, all other models employ A100 or A6000 GPUs. As it can be seen, most MLLMs distribute training across 8 A100s.

Model	Hardware Type	#
Flamingo(Alayrac et al., 2022)	TPUv4	1,535
PaLI (Chen et al., 2023j)	TPUv4	1,024
IDEFICS (Laurençon et al., 2024)	A100	512
SPHINX (Lin et al., 2023b)	A100	32
Emu (Sun et al., 2023b)	A100	128
VILA (Lin et al., 2023a)	A100	128
BLIP-2 (Li et al., 2023g)	A100	16
SEED-LLaMA (Ge et al., 2023b)	A100	64
Shikra (Chen et al., 2023f)	A100	8
MiniGPT-v2 (Chen et al., 2023e)	A100	8
InstructBLIP (Dai et al., 2023)	A100	16
BLIVA (Hu et al., 2024)	A6000	8
CleverFlamingo (Chen et al., 2023b)	A100	8
LLaVA 1.5 (Liu et al., 2023d)	A100	8
LLaVA (Liu et al., 2023e)	A100	8
MiniGPT-4 (Zhu et al., 2023a)	A100	4
FROMAGe (Koh et al., 2023b)	A100	1
LaVIN (Luo et al., 2023)	A100	8

Table 11: Summary of the hardware required to train common MLLMs.

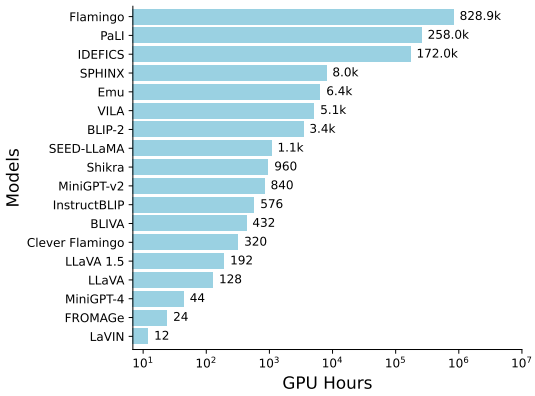

Figure 2: Number of GPU training hours for various MLLMs. Here 1 TPU hour is approximated as 1.5 GPU hours following public benchmarks.

Moreover, in Figure 2 we show for each MLLM the total amount of GPU training hours, approximating 1 TPU hour as 1.5 GPU hours. Notably, models like Flamingo, PaLI, and IDEFICS require a significant amount of GPU time (in the order of magnitude of a few hundred thousand GPU hours). Instead, lighter models like LLaVA only require a few hundred GPU hours to complete training.

E Additional Details on Other Modalities and Applications

Video Understanding. As a complement of Sec. 3.3, we report in Table 12 a summary of the

Model	LLM	Visual Encoder	Main Tasks & Capabilities
VideoChat (Li et al., 2023h)	StableVicuna-13B★	EVA ViT-g	Visual Dialogue, VQA, Captioning
Video-ChatGPT (Maaz et al., 2023)	Vicuna-7B★	CLIP ViT-L	Visual Dialogue, VQA, Captioning
Video-LLaMA (Zhang et al., 2023b)	Vicuna-7B★	EVA ViT-g	Visual Dialogue, Captioning, VQA, Audio Understanding
BT-Adapter (Liu et al., 2023g)	Vicuna-7B★	CLIP ViT-L	Visual Dialogue, Captioning, VQA, Retrieval
LLaMA-VID (Li et al., 2023j)	Vicuna-13B♦	EVA ViT-g	Visual Dialogue, VQA, Captioning
PG-Video-LLaVA (Munasinghe et al., 2023)	LLaVA-1.5-13B★	CLIP ViT-L	Visual Dialogue, Captioning, VQA, Grounding
TimeChat (Ren et al., 2023a)	LLaMA-2-7B▲	EVA ViT-g	Visual Dialogue, Captioning, Temporal Grounding, Highlight Detection
Vista-LLaMA (Ma et al., 2023a)	Vicuna-7B★	EVA ViT-g	Visual Dialogue, VQA, Captioning

Table 12: Summary of video-based MLLMs. For each model, we indicate the LLM used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM (★: frozen LLM; ♦: LLM fine-tuning; ▲: LLM fine-tuning with PEFT techniques).

main characteristics of video-based MLLMs. For each model, we indicate the LLM used as starting point, which in some cases is initialized with the parameters of a pre-trained MLLM, the visual encoder, and the main tasks and capabilities of the MLLM. Additionally, we specify whether the LLM is kept frozen, is entirely fine-tuned, or is fine-tuned with PEFT-based strategies.

3D Understanding. MLLMs are also applied to 3D data for solving complex tasks like 3D VQA, 3D conversation, and 3D dense captioning. Differently from standard visual encodings which exploit 2D pre-trained embeddings, in the context of 3D data, appropriate strategies are designed to project them to the LLM space. In 3D-LLM (Hong et al., 2023), 3D scenes are rendered in different views and 3D features are built using an EVA-CLIP backbone connected to a fine-tuned BLIP-2 model. Similarly, Xu et al. (2023b) employ a pre-trained Point-BERT (Yu et al., 2022) as 3D encoder and conducts a two-stage training that initially aligns the input features via an MLP projection layer, and then performs an instruction tuning phase of the model. Differently, in Point-Bind (Guo et al., 2023), 3D point-clouds are aligned with ImageBind (Girdhar et al., 2023) and by leveraging I2P-MAE (Zhang et al., 2023f) as 3D encoder. This alignment allows the introduction of new tasks such as any-to-3D generation and 3D embedding-space arithmetic. Recently, LL3DA (Chen et al., 2023g) introduces the Interactor3D module, which consists of a frozen 3D scene encoder, a visual prompt encoder, and a Q-Former to address 3D captioning and VQA.

Any-Modality Models. Several studies focus on extending the reasoning capabilities of the MLLMs by including multiple modalities, such as video, 3D, and audio. A line of research investigates the usage of dedicated pathways to input the different modalities to the LLM. UniVAL (Shukor et al., 2023) maps the features from each modality encoder into the shared representation space of the LLM through dedicated linear projections. X-LLM (Chen et al., 2023c) leverages Q-Former interfaces for the image and video modalities, interpreting the video as a sequence of independent frames, each one encoded as an image. For the speech modality, it uses a C-Former interface that compresses the feature sequence from the speech encoder into token-level embeddings. X-InstructBLIP (Panagopoulou et al., 2023) and ChatBridge (Zhao et al., 2023d) propose to freeze both the modality encoders and the LLM and to leverage, respectively, dedicated Q-Former or Perceiver adapters for each modality. To maximize feature compatibility, AnyMAL (Moon et al., 2023) uses an encoder that has already been aligned to a text embedding space for each modality, including also IMU signals, and a dedicated adapter, which is a Perceiver for the visual modality and linear layers for the others. On the other hand, PandaGPT (Su et al., 2023) and NExT-GPT (Wu et al., 2023c) exploit a single frozen multimodal encoder (*i.e.*, ImageBind) to extract features from different modalities. OneLLM (Han et al., 2024) builds a unified universal encoder and a universal projection module by mixing multiple image projection modules and a modality router to align input signals with language. CAT (Ye et al., 2024) adds a clue aggregator to aggregate question-aware audio-visual hidden features and produce clue tokens that are provided to the LLM.

In addition to handling different modalities in input to the LLM, some works investigate the generation of outputs of different modalities. For example, NExT-GPT (Wu et al., 2023c) introduces signal tokens in the LLM that indicate whether the diffusion-based decoder for a specific modality has to be activated. Moreover, the signal tokens are provided to a transformer-based output projector to condition the generation. Similarly, M2UGen (Hussain et al., 2023) handles the music modality by using the LLM output corresponding to signal tokens, along with unimodal music fea-

Model	LLM	Visual Encoder	Main Tasks & Capabilities
Document Analysis			
mPLUG-DocOwl (Ye et al., 2023a)	mPLUG-Owl-7B▲	CLIP ViT-L	Visual Dialogue, Captioning, VQA
Kosmos-2.5 (Lv et al., 2023)	Magneto-1.3B♦	Pix2Struct ViT-L	Text Recognition, Image-to-Markdown Generation
UReader (Ye et al., 2023b)	mPLUG-Owl-7B▲	CLIP VIT-L	Visual Dialogue, VQA, Captioning, Information Extraction
mPLUG-PaperOwl (Hu et al., 2023a)	mPLUG-Owl-7B▲	CLIP ViT-L	Visual Dialogue, VQA, Captioning, Diagram Analysis
LLaMA-SciTune (Horawalavithana et al., 2023)	LLaMA-13B♦	CLIP ViT-L	Visual Dialogue, VQA, Captioning, Diagram Analysis
DocPedia (Feng et al., 2023)	Vicuna-7B♦	Swin-B	Visual Dialogue, VQA, Information Extraction
Embodied AI			
EmbodiedGPT (Mu et al., 2023)	LLaMA-7B★	EVA ViT/g, RN50	Visual Dialogue, VQA, Captioning, Task Planning
PaLM-E (Driess et al., 2023)	PaLM-540B♦	ViT-22B	Visual Dialogue, VQA, Captioning, Task Planning, Manipulation
Medical Vision Learning			
PMC-VQA (Zhang et al., 2023h)	PMC-LLaMA-7B★	PMC-CLIP RN50	VQA
LLaVA-Med (Li et al., 2023d)	LLaVA-7B♦	CLIP ViT-L	Visual Dialogue, VQA
Qilin-Med-VL (Liu et al., 2023f)	CN-LLaMA2-13B♦	CLIP ViT-L	Visual Dialogue, VQA
Autonomous Driving			
Dolphins (Ma et al., 2023b)	OpenFlamingo-7B▲	CLIP ViT-L	Visual Dialogue, VQA, Captioning, Traffic Condition Understanding
DriveGPT4 (Xu et al., 2023c)	LLaMA-2-7B♦	CLIP ViT-L	Visual Dialogue, VQA, Captioning
Food Understanding			
FoodLLM (Yin et al., 2023c)	LISA-7B▲	CLIP ViT-L	Visual Dialogue, VQA, Nutrition Estimation, RES

Table 13: Summary of MLLMs designed for domain-specific applications. For each model, we indicate the LLM used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM (★: frozen LLM; ♦: LLM fine-tuning; ▲: LLM fine-tuning with PEFT techniques). Gray color indicates models not publicly available.

tures from a music encoder, to condition the generation of an audio encoder. LLMBind (Zhu et al., 2024) indicates the conditioning text to generate image, video, or audio by wrapping it in special tokens. Thus, this text is provided to the corresponding modality-specific diffusion model. Unified-IO2 (Lu et al., 2023a) uses VQ-GAN decoders for both image and audio modalities to decode output discrete tokens and can generate surface normals, depth, and segmentation masks for the input images. AnyGPT (Zhan et al., 2024) interprets all the continuous non-text modalities as discrete tokens in both input and output, using, respectively, multimodal tokenizers and de-tokenizers. To enable the 3D modality, LAMM (Yin et al., 2023d) introduces a novel instruction tuning dataset and benchmark that comprise both image-text and point cloud-text instruction-response pairs, covering a wide range of 2D and 3D tasks.

Interactive and Compositional Systems. A different trend is to build systems that can combine multiple tools (*i.e.*, existing vision-only or vision-and-language models), usually through ChatGPT or another LLM. In particular, these approaches aim to let the user interact with the LLM which is in charge of selecting the useful tools to carry out complex tasks. In this context, some solutions study how to prompt ChatGPT (Wu et al., 2023a; Yang et al., 2023c) to invoke visual foundation models. GPT4Tools (Yang et al., 2023a), instead, employs open-source LLMs such as LLaMA and OPT, that are fine-tuned with PEFT techniques to use tools for performing a wide range of visual tasks. Differently, Liu et al. (2023l) introduce more sophisticated user-chatbot interactions, through the incorporation of mouse-based pointing instructions on images or videos.

While in all these approaches the LLM does not directly handle the visual input which is instead processed by other external tools, in LLaVA-Plus (Liu et al., 2023h) the query image is directly input to the MLLM (*i.e.*, LLaVA) and is therefore involved during the selection and invocation of the most helpful tool according to the user needs. This is achieved also thanks to the introduction of a new instruction-following use tool dataset, which is employed to fine-tune the MLLM.

Domain-Specific MLLMs. Finally, in Table 13 we summarize the main characteristics of domain-specific MLLMs, also in this case indicating for each model the LLM used as starting point.