MM-LLMs: Recent Advances in MultiModal Large Language Models

Anonymous ACL submission

Abstract

In the past year, MultiModal Large Language Models (MM-LLMs) have undergone substantial advancements, augmenting off-the-shelf LLMs to support MM inputs or outputs via cost-effective training strategies. The resulting models not only preserve the inherent reasoning and decision-making capabilities of LLMs but also empower a diverse range of MM tasks. In this paper, we provide a comprehensive survey aimed at facilitating further research of MM-LLMs. Initially, we outline general design formulations for model architecture and training pipeline. Subsequently, we introduce a taxonomy encompassing 122 MM-LLMs, each characterized by its specific formulations. Furthermore, we review the performance of selected MM-LLMs on mainstream benchmarks and summarize key training recipes to enhance the potency of MM-LLMs. Finally, we explore promising directions for MM-LLMs while concurrently maintaining a real-time tracking website[1] for the latest developments in the field. We hope that this survey contributes to the ongoing advancement of the MM-LLMs domain.

1 Introduction

MultiModal (MM) pre-training research has witnessed significant advancements in recent years, consistently pushing the performance boundaries across a spectrum of downstream tasks (Li et al., 2020; Akbari et al., 2021; Fang et al., 2021; Yan et al., 2021; Li et al., 2021; Radford et al., 2021; Li et al., 2022; Zellers et al., 2022; Zeng et al., 2022b; Yang et al., 2022; Wang et al., 2022a,b). However, as the scale of models and datasets continues to expand, traditional MM models incur substantial computational costs, particularly when trained from scratch. Recognizing that MM research operates at the intersection of various modalities, a logical approach is to capitalize on readily available pre-trained unimodal foundation models, with a special emphasis on powerful Large Language Models (LLMs) (OpenAI, 2022). This strategy aims to mitigate computational expenses and enhance the efficacy of MM pre-training, leading to the emergence of a novel field: MM-LLMs.

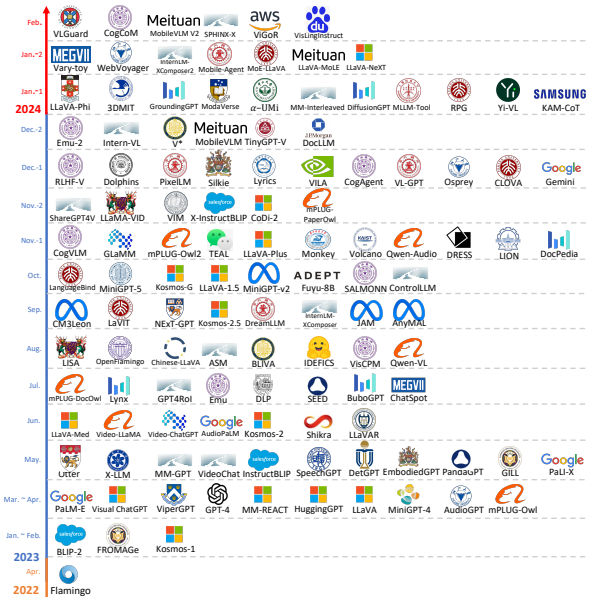

Figure 1: The timeline of MM-LLMs.

MM-LLMs harness LLMs as the cognitive powerhouse to empower various MM tasks. LLMs contribute desirable properties like robust language generation, zero-shot transfer capabilities, and In-Context Learning (ICL). Concurrently, foundation models in other modalities provide high-quality representations. Considering foundation models from different modalities are individually pre-trained, the core challenge facing MM-LLMs is how to effectively connect LLMs with models in other modalities to enable collaborative inference. The predominant focus within this field has been on refining alignment between modalities and aligning with human intent via a MM Pre-Training (PT) + MM Instruction-Tuning (IT) pipeline.

With the debut of GPT-4(Vision) (OpenAI, 2023)

[1] https://mm-llms.github.io

and Gemini (Team et al., 2023), showcasing impressive MM understanding and generation capabilities, a research fervor on MM-LLMs has been sparked. Initial research primarily focuses on MM content comprehension and text generation, encompassing tasks such as image-text understanding, exemplified by projects like BLIP-2 (Li et al., 2023e), LLaVA (Liu et al., 2023e), MiniGPT-4 (Zhu et al., 2023a), and OpenFlamingo (Awadalla et al., 2023); video-text understanding, as demonstrated by initiatives such as VideoChat (Li et al., 2023f), Video-ChatGPT (Maaz et al., 2023), and LLaMA-VID (Li et al., 2023j); and audio-text understanding, as seen in projects like Qwen-Audio (Chu et al., 2023b). Later, the capabilities of MM-LLMs have been expanded to support specific modality outputs. This includes tasks with image-text output, such as GILL (Koh et al., 2023a), Kosmos-2 (Peng et al., 2023), Emu (Sun et al., 2024), and MiniGPT-5 (Zheng et al., 2023b); as well as speech/audio-text output, exemplified by projects like SpeechGPT (Zhang et al., 2023a) and AudioPaLM (Rubenstein et al., 2023). Recent research endeavors have focused on mimicking human-like any-to-any modality conversion, shedding light on the path to artificial general intelligence. Some efforts aim to amalgamate LLMs with external tools to reach an approaching any-to-any MM comprehension and generation, such as Visual-ChatGPT (Wu et al., 2023a), HuggingGPT (Shen et al., 2023), and AudioGPT (Huang et al., 2023b). Conversely, to mitigate propagated errors in the cascade system, initiatives like NExT-GPT (Wu et al., 2023d), CoDi-2 (Tang et al., 2023c), and ModaVerse (Wang et al., 2024c) have developed end-to-end MM-LLMs of arbitrary modalities. The timeline of MM-LLMs is depicted in Figure 1.

In this paper, we present a comprehensive survey aimed at facilitating further research of MM-LLMs. To provide readers with a holistic understanding of MM-LLMs, we initially delineate general design formulations from model architecture (Section 2) and training pipeline (Section 3). We break down the general model architecture into five components: Modality Encoder (Section 2.1), Input Projector (Section 2.2), LLM Backbone (Section 2.3), Output Projector (Section 2.4), and Modality Generator (Section 2.5). The training pipeline elucidates how to enhance a pre-trained text-only LLM to support MM input or output, primarily consisting of two stages: MM PT (Section 3.1) and MM IT (Section 3.2). In that section, we also provide a summary of mainstream datasets for MM PT and MM IT. Next, we establish a taxonomy encompassing 122 State-of-the-Art (SOTA) MM-LLMs, each characterized by specific formulations, and summarize their development trends in Section 4. In Section 5, we comprehensively review the performance of major MM-LLMs on mainstream benchmarks and distill key training recipes to enhance the efficacy of MM-LLMs. In Section 6, we offer promising directions for MM-LLMs research. Moreover, we have established a website (https://mm-llms.github.io) to track the latest progress of MM-LLMs and facilitate crowdsourcing updates. Finally, we summarize the entire paper in Section 7 and discuss related surveys on MM-LLMs in Appendix A. We aspire for our survey to aid researchers in gaining a deeper understanding of this field and to inspire the design of more effective MM-LLMs.

2 Model Architecture

In this section, we provide a detailed overview of the five components comprising the general model architecture, along with the implementation choices for each component, as illustrated in Figure 2. MM-LLMs that emphasize MM understanding only include the first three components. During training, Modality Encoder, LLM Backbone, and Modality Generator are generally maintained in a frozen state. The primary optimization emphasis is on Input and Output Projectors. Given that Projectors are lightweight components, the proportion of trainable parameters in MM-LLMs is notably small compared to the total parameter count (typically around 2%). The overall parameter count is contingent on the scale of the core LLM utilized in the MM-LLMs. As a result, MM-LLMs can be efficiently trained to empower various MM tasks.

2.1 Modality Encoder

The Modality Encoder (ME) is tasked with encoding inputs from diverse modalities I_X to obtain corresponding features $\boldsymbol{F}_X$, formulated as follows:

$$\boldsymbol{F}_X = \text{ME}_X(I_X). \quad (1)$$

Various pre-trained encoder options ME_X exist for handling different modalities, where X can be image, video, audio, 3D, etc. Next, we will offer a concise introduction organized by modality.

Visual Modality For images, there are various optional encoders: **NFNet-F6** (Brock et al.,

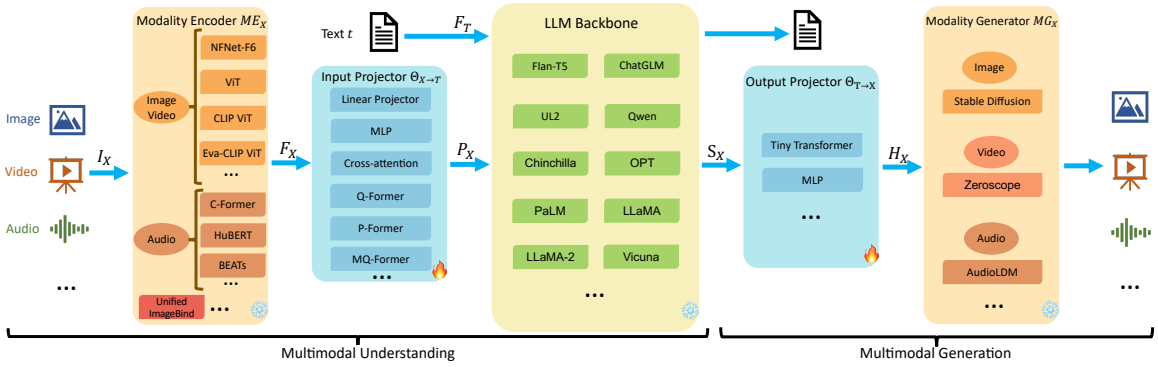

Figure 2: The general model architecture of MM-LLMs and the implementation choices for each component.

2021), **ViT** (Dosovitskiy et al., 2020), **CLIP ViT** (Radford et al., 2021), **Eva-CLIP ViT** (Fang et al., 2023), **BEiT-3** (Wang et al., 2023d), **OpenCLIP** (Cherti et al., 2023), **Grounding-DINO-T** (Zhang et al., 2022b) with Swin-T (Liu et al., 2021b) backbone, **DINOv2** (Oquab et al., 2023), **SAM-HQ** (Kirillov et al., 2023) with MAE (He et al., 2022), **RAM++** (Zhang et al., 2023i) with Swin-B backbone, **InternViT** (Chen et al., 2023j), and **VCoder** (Jain et al., 2023). For videos, they can be uniformly sampled to 5 frames, undergoing the same pre-processing as images.

Audio Modality is typically encoded by **C-Former** (Chen et al., 2023b), **HuBERT** (Hsu et al., 2021), **BEATs** (Chen et al., 2023g), **Whisper** (Radford et al., 2023), and **CLAP** (Wu et al., 2023e).

3D Point Cloud Modality is typically encoded by **ULIP-2** (Salesforce, 2022) with a Point-BERT (Yu et al., 2022) backbone.

Moreover, to handle numerous heterogeneous modal encoders, some MM-LLMs, particularly any-to-any ones, use **ImageBind** (Girdhar et al., 2023), a unified encoder covering six modalities, including image/video, text, audio, heat map, inertial measurement units, and depth. We provide a brief introduction to some mainstream modality encoders in Appendix B.

2.2 Input Projector

The Input Projector $\boldsymbol{\Theta}_{X\to T}$ is tasked with aligning the encoded features of other modalities $\boldsymbol{F}_X$ with the text feature space T. The aligned features as prompts $\boldsymbol{P}_X$ are then fed into the LLM Backbone alongside the textual features $\boldsymbol{F}_T$. Given X-text dataset $\{I_X, t\}$, the goal is to minimize the X-conditioned text generation loss $\mathcal{L}_{\text{txt-gen}}$:

$$\arg\min_{\boldsymbol{\Theta}_{X\to T}} \mathcal{L}_{\text{txt-gen}}(\text{LLM}(\boldsymbol{P}_X, \boldsymbol{F}_T), t), \quad (2)$$

where $\boldsymbol{P}_X = \boldsymbol{\Theta}_{X\to T}(\boldsymbol{F}_X)$.

The Input Projector can be achieved directly by a **Linear Projector** or Multi-Layer Perceptron (**MLP**), *i.e.*, several linear projectors interleaved with non-linear activation functions. There are also more complex implementations like **Cross-attention**, **Q-Former** (Li et al., 2023e), **P-Former** (Jian et al., 2023), and **MQ-Former** (Lu et al., 2023a). **Cross-attention** (Perceiver Resampler) (Alayrac et al., 2022) uses a set of trainable vectors as queries and the encoded features $\boldsymbol{F}_X$ as keys to compress the feature sequence to a fixed length. The compressed representation is then fed directly into the LLM or further used for X-Text cross-attention fusion. **Q-Former** extracts relevant features from $\boldsymbol{F}_X$ with learnable queries, and the selected features are then used as prompts $\boldsymbol{P}_X$. Meanwhile, **P-Former** generates "reference prompts", imposing an alignment constraint on the prompts produced by Q-Former. **MQ-Former** conducts a fine-grained alignment of multi-scale visual and textual signals. However, both Q-, P-, MQ-Former require an **additional PT** process for initialization.

2.3 LLM Backbone

Taking LLMs (Zhao et al., 2023c; Naveed et al., 2023; Luo et al., 2023) as the core agents, MM-LLMs can inherit some notable properties like zero-shot generalization, few-shot ICL, Chain-of-Thought (CoT), and instruction following. The LLM Backbone processes representations from various modalities, engaging in semantic understanding, reasoning, and decision-making regarding the inputs. It produces (1) direct textual outputs t, and (2) signal tokens $\boldsymbol{S}_X$ from other modalities (if any). These signal tokens act as instructions to guide the generator on whether to produce MM contents and,

if affirmative, specifying the content to produce:

$$t, \boldsymbol{S}_X = \text{LLM}(\boldsymbol{P}_X, \boldsymbol{F}_T), \quad (3)$$

where the aligned representations of other modalities $\boldsymbol{P}_X$ can be considered as soft Prompt-tuning for the LLM. Moreover, some works have introduced Parameter-Efficient Fine-Tuning (PEFT) methods, such as Prefix-tuning (Li and Liang, 2021), LoRA (Hu et al., 2021), and LayerNorm tuning (Zhao et al., 2024). In these cases, the number of additional trainable parameters is exceptionally minimal, even less than 0.1% of the total LLM parameter count. We provide an introduction to mainstream PEFT methods in Appendix C.

The commonly used LLMs in MM-LLMs incude **Flan-T5** (Chung et al., 2022), **ChatGLM** (Zeng et al., 2022a), **UL2** (Tay et al., 2022), **Persimmon** (Elsen et al., 2023), **Qwen** (Bai et al., 2023a), **Chinchilla** (Hoffmann et al., 2022), **OPT** (Zhang et al., 2022c), **PaLM** (Chowdhery et al., 2023), **LLaMA** (Touvron et al., 2023a), **LLaMA-2** (Touvron et al., 2023b), and **Vicuna** (Chiang et al., 2023). We provide a brief introduction to some representative LLMs in Appendix D.

2.4 Output Projector

The Output Projector $\boldsymbol{\Theta}_{T\rightarrow X}$ maps the signal token representations $\boldsymbol{S}_X$ from the LLM Backbone into features $\boldsymbol{H}_X$ understandable to the following Modality Generator MG_X. Given the X-text dataset $\{I_X, t\}$, t is first fed into LLM to generate the corresponding $\boldsymbol{S}_X$, then mapped into $\boldsymbol{H}_X$. To facilitate alignment of the mapped features $\boldsymbol{H}_X$, the goal is to minimize the distance between $\boldsymbol{H}_X$ and the conditional text representations of MG_X:

$$\arg\min_{\boldsymbol{\Theta}_{T\rightarrow X}} \mathcal{L}_{\text{mse}}(\boldsymbol{H}_X, \tau_X(t)). \quad (4)$$

The optimization only relies on captioning texts, without utilizing any audio or visual resources X, where $\boldsymbol{H}_X = \boldsymbol{\Theta}_{T\rightarrow X}(\boldsymbol{S}_X)$ and τ_X is the textual condition encoder in MG_X. The Output Projector is implemented by a **Tiny Transformer** with a learnable decoder feature sequence or **MLP**.

2.5 Modality Generator

The Modality Generator MG_X is tasked with producing outputs in distinct modalities. Commonly, existing works use off-the-shelf Latent Diffusion Models (LDMs) (Song et al., 2021; Bao et al., 2022; Zhao et al., 2022), *i.e.*, **Stable Diffusion** (Rombach et al., 2022) for image synthesis, **Zeroscope** (Cerspense, 2023) for video synthesis, and **AudioLDM-2** (Liu et al., 2023b,c) for audio synthesis. The features $\boldsymbol{H}_X$ mapped by the Output Projector serve as conditional inputs in the denoising process to generate MM content. During training, the ground truth content is first transformed into a latent feature z_0 by the pre-trained VAE (Kingma and Welling, 2013). Then, noise ϵ is added to z_0 to obtain the noisy latent feature z_t. A pre-trained Unet (Ronneberger et al., 2015) ϵ_X is used to compute the conditional LDM loss $\mathcal{L}_{\text{X-gen}}$ as follows:

$$\mathcal{L}_{\text{X-gen}} := \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}||\epsilon - \epsilon_X(z_t, t, \boldsymbol{H}_X)||_2^2, \quad (5)$$

which optimizes parameters $\boldsymbol{\Theta}_{X\rightarrow T}$ and $\boldsymbol{\Theta}_{T\rightarrow X}$ by minimizing $\mathcal{L}_{\text{X-gen}}$.

3 Training Pipeline

MM-LLMs' training pipeline can be delineated into two principal stages: MM PT and MM IT.

3.1 MM PT

During the PT stage, typically leveraging the X-Text datasets, Input and Output Projectors are trained to achieve alignment among various modalities by optimizing predefined objectives. For MM understanding models, optimization focuses solely on Equation (2), while for MM generation models, optimization involves Equations (2), (4), and (5). In the latter case, Equation (2) also includes the ground-truth signal token sequence.

The X-Text datasets include Image-Text, Video-Text, and Audio-Text, with Image-Text having two types: Image-Text pairs (*e.g.*, **<img1>** <txt1>) and interleaved Image-Text corpus (*e.g.*, <txt1>**<img1>**<txt2><txt3>**<img2>**<txt4>). Details of X-Text datasets are shown in Table 3 of Appendix G.

3.2 MM IT

MM IT is a method that entails fine-tuning of pre-trained MM-LLMs using instruction-formatted datasets (Wei et al., 2021). Through this process, MM-LLMs can generalize to unseen tasks by adhering to new instructions, thereby enhancing zero-shot performance. This straightforward yet impactful concept has catalyzed subsequent success in the field of NLP, exemplified by works such as InstructGPT (Ouyang et al., 2022), OPT-IML (Iyer et al., 2022), and InstructBLIP (Dai et al., 2023).

MM IT comprises Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human

Figure 3: Taxonomy for MM-LLMs. I: Image, V: Video, A/S: Audio/Speech, and T: Text. $\mathbf{I_D}$: Document understanding, $\mathbf{I_B}$: Output bounding box, $\mathbf{I_M}$: Output segmentation mask, and $\mathbf{I_R}$: Output retrieved images.

Feedback (RLHF), aiming to align with human intents and enhance the interaction capabilities of MM-LLMs. SFT converts part of the PT stage data into an instruction-aware format. Using visual Question-Answer (QA) as an example, various templates may be employed like **(1) "<Image>{Question}"** A short answer to the question is; **(2) "<Image>"** Examine the image and respond to the following question with a brief answer: **"{Question}. Answer:"**; and so on. Next, it fine-tunes pre-trained MM-LLMs using the same optimization objectives. SFT datasets can be structured as either single-turn QA or multi-turn dialogues.

After SFT, RLHF involves further fine-tuning of the model, relying on feedback regarding the MM-LLMs' responses (*e.g.*, Natural Language Feedback (NLF) labeled manually or automatically) (Sun et al., 2023b). This process employs a reinforcement learning algorithm to effectively integrate the non-differentiable NLF. The model is trained to generate corresponding responses conditioned on the NLF (Chen et al., 2023i; Akyürek et al., 2023). The statistics for SFT and RLHF datasets are presented in Table 4 of Appendix G.

The datasets used by existing MM-LLMs in the MM PT and MM IT stages are diverse, but they are all **subsets** of the datasets in Tables 3 and 4.

4 SOTA MM-LLMs

As shown in Figure 3, we classify the 122 SOTA MM-LLMs from both functional and design perspectives. In the design division, "Tool-using" denotes treating the LLM as black box and providing access to certain MM expert systems to perform specific MM tasks via reasoning, while "End-to-End" signifies that the entire model is trained jointly in an end-to-end manner. Based on the previously defined design formulations, we also

Model	I→O	Modality Encoder	Input Projector	LLM Backbone	Output Projector	Modality Generator	#.PT	#.IT
Flamingo	I+V+T→T	I/V: NFNet-F6	Cross-attention	Chinchilla-1.4B/7B/70B	–	–	–	–
BLIP-2	I+T→T	I: CLIP/Eva-CLIP ViT@224	Q-Former w/ Linear Projector	Flan-T5/OPT	–	–	129M	–
LLaVA	I+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-7B/13B	–	–	–	–
MiniGPT-4	I+T→T	I: Eva-CLIP ViT-G/14	Q-Former w/ Linear Projector	Vicuna-13B	–	–	–	–
mPLUG-Owl	I+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–	–	–	–
Otter	I+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–	–	–	–
X-LLM	I+V+A+T→T	I/V: ViT-G; A: C-Former	Q-Former w/ Linear Projector	ChatGLM-6B	–	–	–	–
VideoChat	V+T→T	I: ViT-G	Q-Former w/ Linear Projector	Vicuna	–	–	–	–
InstructBLIP	I+V+T→T	I/V: ViT-G/14@224	Q-Former w/ Linear Projector	Flan-T5/Vicuna	–	–	129M	1.2M
PandaGPT	I+T→T	I: ImageBind	Linear Projector	Vicuna-13B	–	–	–	–
GILL	I+T→I+T	I: CLIP ViT-L	Linear Projector	OPT-6.7B	Tiny Transformer	I: Stable Diffusion-1.5	–	–
PaLI-X	I+T→T	I: ViT	Linear Projector	UL2-32B	–	–	–	–
Video-LLaMA	I+V+A+T→T	I/V: Eva-CLIP ViT-G/14; A: ImageBind	Q-Former w/ Linear Projector	Vicuna/LLaMA	–	–	–	–
Video-ChatGPT	V+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-v1.1	–	–	–	–
Shikra	I+T→T+I_R	I: CLIP ViT-L/14@224	Linear Projector	Vicuna-7B/13B	–	–	600K	5.5M
LLaVAR	I+T→T	I: CLIP ViT-L/14@224 & CLIP ViT-L/14@336	Linear Projector	Vicuna-13B	–	–	–	–
mPLUG-DocOwl	I_D+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–	–	–	–
Lynx	I+V+T→T	I/V: Eva-CLIP ViT-1B	Cross-attention	Vicuna	–	–	–	–
Emu	I+V+T→I+T	I/V: Eva-CLIP-1B	Cross-attention	LLaMA-13B	MLP	I: Stable Diffusion-1.5	–	–
DLP	I+T→T	I: CLIP/Eva-CLIP ViT	Q-Former+P-Former w/ Linear Projector	OPT/Flan-T5	–	–	–	–
BuboGPT	I+A+T→T+I_M	I: CLIP/Eva-CLIP ViT; A: ImageBind	Q-Former w/ Linear Projector	Vicuna	–	–	–	–
ChatSpot	I+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-7B/LLaMA	–	–	–	–
IDEFICS	I+T→T	I: OpenCLIP	Cross-attention	LLaMA	–	–	–	–
Qwen-VL-(Chat)	I+T→T	I: ViT@448 initialized from OpenClip's ViT-bigG	Cross-attention	Qwen-7B	–	–	1.4B†	50M†
LaVIT	I+T→I+T	I: ViT	Cross-attention	LLaMA-7B	–	I: Stable Diffusion	–	–
NExT-GPT	I+V+A+T→I+V+A+T	I/V/A: ImageBind	Linear Projector	Vicuna-7B	Tiny Transformer	I: Stable Diffusion; V: Zeroscope; A: AudioLDM	–	–
DreamLLM	I+T→I+T	I: CLIP ViT-L	Linear Projector	Vicuna	MLP	I: Stable Diffusion	–	–
AnyMAL	I+V+A+T→T	I: CLIP ViT/L & ViT-G & DinoV2; V: Intervideo; A: CLAP	I/V: Cross-attention; A: Linear Projector	LLaMA-2	–	–	–	–
MiniGPT-5	I+T→I+T	I: Eva-CLIP ViT-G/14	Q-Former w/ Linear Projector	Vicuna-7B	Tiny Transformer w/ MLP	I: StableDiffusion-2	–	–
LLaVA-1.5	I+T→T	I: CLIP ViT-L@336	MLP	Vicuna-v1.5-7B/13B	–	–	0.6M	0.7M
MiniGPT-v2	I+T→T	I: Eva-CLIP ViT@448	Linear Projector	LLaMA-2-Chat-7B	–	–	–	–
CogVLM	I+T→T	I: Eva-2-CLIP ViT	MLP	Vicuna-v1.5-7B	–	–	–	–
Qwen-Audio	A+T→T	A: Whisper-L-v2	Linear Projector	Qwen-7B	–	–	–	–
DRESS	I+T→T	I:Eva-CLIP ViT-G/14	Linear Projector	Vicuna-v1.5-13B	–	–	–	–
X-InstructBLIP	I+V+A+3D+T→T	I/V: Eva-CLIP ViT-G/14; A: BEATs; 3D: ULIP-2	Q-Former w/ Linear Projector	Vicuna-v1.1-7B/13B	–	–	–	–
CoDi-2	I+V+A+T→I+V+A+T	I/V/A: ImageBind	MLP	LLaMA-2-Chat-7B	MLP	I: Stable Diffusion-2.1; V: Zeroscope-v2; A: AudioLDM-2	–	–
RLHF-V	I+T→T	I: BEiT-3	Linear Projector	Vicuna-v1-13B	–	–	–	–
Silkie	I+T→T	I: ViT initialized from OpenCLIP's ViT-bigG	Cross-attention	Qwen-7B	–	–	–	–
Lyrics	I+T→T	I: CLIP ViT-L/14 & Grounding-DINO-T w/ Swin-T & SAM-HQ w/ MAE & ViT-H & RAM++ w/ Swin-B	MQ-Former w/ Linear Projection	Vicuna-13B	–	–	–	–
VILA	I+T→T	I: ViT@336	Linear Projector	LLaMA-2-7B/13B	–	–	50M	1M
IntrenVL	I+V+T→T	I/V: InternViT-6B; T: LLaMA-7B	Cross-attention w/ MLP	QLLaMA-8B & Vicuna-13B	–	–	–	–
ModaVerse	I+V+A+T→I+V+A+T	ImageBind	Linear Projector	LLaMA-2	MLP	I: Stable Diffusion; V: Videofusion; A: AudioLDM	–	–
MM-Interleaved	I+T→I+T	I: CLIP ViT-L/14	Cross-attention	Vicuna-13B	Tiny Transformer	I: Stable Diffusion-2.1	–	–

Table 1: The summary of 43 mainstream MM-LLMs. I→O: Input to Output Modalities, I: Image, V: Video, A: Audio, 3D: Point Cloud, and T: Text. In Modality Encoder, "-L" represents Large, "-G" represents Giant, "/14" indicates a patch size of 14, and "@224" signifies an image resolution of 224×224. **#.PT** and **#.IT** represent the scale of dataset during MM PT and MM IT, respectively. † includes in-house data that is not publicly accessible.

conduct a comprehensive comparison of the architectures and training dataset scales for 43 of these SOTA MM-LLMs, as illustrated in Table 1. Next, we will summarize their developmental trends and briefly introduce the core contributions of some representative models in Appendix E.

Trends in Existing MM-LLMs: **(1)** Progressing from a dedicated emphasis on MM understanding to the generation of specific modalities and further evolving into any-to-any modality conversion (*e.g.*, MiniGPT-4 → MiniGPT-5 → NExT-GPT); **(2)** Advancing from MM PT to SFT and then to RLHF, the training pipeline undergoes continuous refinement, striving to better align with human intent and enhance the model's conversational interaction capabilities (*e.g.*, BLIP-2 → InstructBLIP → DRESS); **(3)** Embracing Diversified Modal Extensions (*e.g.*, BLIP-2 → X-LLM and InstructBLIP → X-InstructBLIP); **(4)** Incorporating a Higher-Quality Training Dataset (*e.g.*, LLaVA → LLaVA-1.5); **(5)** Adopting a More Efficient Model Architecture, transitioning from complex Q- and P-Former input projector modules in BLIP-2 and DLP to a simpler yet effective linear projector in VILA.

5 Benckmarks and Performance

To offer a comprehensive performance comparison, we have compiled a table featuring major MM-LLMs across 18 VL benchmarks gathered from various papers (Li et al., 2023e; Chen et al., 2023d,f; Lin et al., 2023), as shown in Table 2. The information of these benchmarks can be found in Appendix F. Next, we will extract essential training recipes that boost the effectiveness of MM-LLMs, drawing insights from SOTA models.

Training Recipes **Firstly**, higher image resolution can incorporate more visual details for the model, benefiting tasks that require fine-grained details. For example, LLaVA-1.5 and VILA employ a resolution of 336×336, while Qwen-VL and MiniGPT-v2 utilize 448×448. However, higher resolutions lead to longer token sequences, incurring additional training and inference costs. MiniGPT-v2 addresses this by concatenating 4 adjacent visual tokens in the embedding space to reduce length. Recently, Monkey (Li et al., 2023l) proposed a solution to enhance the resolution of input images without retraining a high-resolution visual encoder, utilizing only a low-resolution visual encoder, supporting resolutions up to 1300×800. To enhance the understanding of rich-text images, tables, and document content, DocPedia (Feng et al., 2023) introduced a method to increase the visual encoder resolution to 2560×2560, overcoming the limitations of poorly performing low resolutions in open-sourced ViT. **Secondly**, the incorporation of high-quality SFT data can significantly improve performance in specific tasks, as evidenced

Model	LLM Backbone	OKVQA	IconVQA	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME^{P}	MME^{C}	MMB	MMB^{CN}	SEED^{I}	LLaVA^{W}	MM-Vet	QBench	HM	VSR
Flamingo	Chinchilla-7B	44.7	–	–	–	28.8	–	–	–	–	–	–	–	–	–	–	–	57.0	31.8
BLIP-2	Flan-T5$_{XXL}$(13B)	45.9	40.6	65.0	44.7	19.6	61.0	42.5	85.3	1293.8	290.0	–	–	46.4	38.1	22.4	–	53.7	50.9
LLaVA	Vicuna-13B	54.4	43.0	–	41.3	–	–	38.9	–	–	–	–	–	–	–	–	–	–	51.2
MiniGPT-4	Vicuna-13B	37.5	37.6	–	30.8	–	–	19.4	–	–	–	–	–	–	–	–	–	–	41.6
InstructBLIP	Vicuna-7B	–	–	–	49.2	34.5	60.5	50.1	–	–	–	36.0	23.7	53.4	60.9	26.2	56.7	–	–
InstructBLIP	Vicuna-13B	–	44.8	–	49.5	33.4	63.1	50.7	78.9	1212.8	291.8	–	–	–	58.2	25.6	–	57.5	52.1
Shikra	Vicuna-13B	47.2	–	77.4*	–	–	–	–	–	–	–	58.8	–	–	–	–	54.7	–	–
IDEFICS-9B	LLaMA-7B	–	–	50.9	38.4	35.5	–	25.9	–	–	–	48.2	25.2	–	–	–	–	–	–
IDEFICS-80B	LLaMA-65B	–	–	60.0	45.2	36.0	–	30.9	–	–	–	54.5	38.1	–	–	–	–	–	–
Qwen-VL	Qwen-7B	–	–	78.8*	59.3*	35.2	67.1	63.8	–	–	–	38.2	7.4	56.3	–	–	59.4	–	–
Qwen-VL-Chat	Qwen-7B	–	–	78.2*	57.5*	38.9	68.2	61.5	–	1487.5	**360.7**	60.6	56.7	58.2	–	–	–	–	–
LLaVA-1.5	Vicuna-1.5-7B	–	–	78.5*	62.0*	50.0	66.8	58.2	**85.9**	1510.7	316.1‡	64.3	58.3	58.6	63.4	30.5	58.7	–	–
+ShareGPT4V	Vicuna-1.5-7B	–	–	**80.6**	–	57.2	68.4	–	–	**1567.4**	**376.4**	68.8	62.2	**69.7**	72.6	37.6	**63.4**	–	–
LLaVA-1.5	Vicuna-1.5-13B	–	–	80.0*	**63.3***	53.6	71.6	61.3	**85.9**	1531.3	295.4‡	67.7	63.6	61.6	70.7	35.4	**62.1**	–	–
MiniGPT-v2	LLaMA-2-Chat-7B	**56.9**	**47.7**	–	60.3	30.3	–	51.9	–	–	–	–	–	–	–	–	–	**58.2**	**60.6**
MiniGPT-v2-Chat	LLaMA-2-Chat-7B	**55.9**	**49.4**	–	58.8	42.4	–	52.3	–	–	–	–	–	–	–	–	–	**59.5**	**63.3**
VILA-7B	LLaMA-2-7B	–	–	79.9*	62.3*	57.8	68.2	64.4	**85.5**	1533.0	–	68.9	61.7	61.1	69.7	34.9	–	–	–
VILA-13B	LLaMA-2-13B	–	–	**80.8***	**63.3***	**60.6**	**73.7**	**66.6**	84.2	**1570.1**	–	**70.3**	**64.3**	**62.8**	**73.0**	**38.8**	–	–	–
+ShareGPT4V	LLaMA-2-13B	–	–	**80.6***	**63.2***	**62.4**	**73.1**	**65.3**	84.8	1556.5	–	**70.8**	**65.4**	61.4	**78.4**	**45.7**	–	–	–

Table 2: Comparison of mainstream MM-LLMs on 18 VL benchmarks. The **red** denotes the highest result, and the **blue** denotes the second highest result. ‡ indicates ShareGPT4V's (Chen et al., 2023f) re-implemented test results, which are missed in benchmarks or origin papers. * indicates that training images are observed during training.

by the addition of ShareGPT4V data to LLaVA-1.5 and VILA-13B, as shown in Table 2. **Moreover**, VILA reveals several key findings: (1) Performing PEFT on the LLM Backbone promotes deep embedding alignment, crucial for ICL; (2) Interleaved Image-Text data proves beneficial, whereas Image-Text pairs alone are sub-optimal; (3) Re-blending text-only instruction data (*e.g.*, unnatural instruction (Honovich et al., 2022)) with image-text data during SFT not only addresses the degradation of text-only tasks but also enhances VL task accuracy.

6 Future Directions

In this section, we explore promising future directions for MM-LLMs across the following aspects:

More Powerful Models We can enhance the MM-LLMs' strength from the following four key avenues: **(1) Expanding Modalities**: Current MM-LLMs mainly support the following modalities: image, video, audio, 3D, and text. However, the real world involves a broader range of modalities. Extending MM-LLMs to accommodate additional modalities (*e.g.*, web pages, heat maps, and figures&tables) will increase the model's versatility, making it more universally applicable; **(2) Diversifying LLMs**: Incorporating various types and sizes of LLMs provides practitioners with the flexibility to select the most appropriate one based on their specific requirements; **(3) Improving MM IT Dataset Quality**: Current MM IT datasets have ample room for improvement and expansion. Diversifying the range of instructions can enhance the effectiveness of MM-LLMs in understanding and executing user commands. **(4) Strengthening MM Generation Capabilities**: Most current MM-LLMs are predominantly oriented towards MM understanding. Although some models have incorporated MM generation capabilities, the quality of generated responses may be constrained by the capacities of the LDMs. Exploring the integration of retrieval-based approaches (Asai et al., 2023; Gao et al., 2023a) holds significant promise in complementing the generative process, potentially enhancing the overall performance of the model.

More Challenging Benchmarks Existing benchmarks might not adequately challenge the capabilities of MM-LLMs, given that many datasets have previously appeared to varying degrees in the PT or IT sets. This implies that the models may have learned these tasks during training. Moreover, current benchmarks predominantly concentrate on the VL sub-field. Thus, it is crucial for the development of MM-LLMs to construct a more challenging, larger-scale benchmark that includes more modalities and uses a unified evaluation standard. For instance, GOAT-Bench (Lin et al., 2024b) is introduced to assess the capability of various MM-LLMs in discerning and responding to nuanced aspects of social abuse depicted in memes. MathVista (Lu et al., 2024) evaluates the math reasoning ability of MM-LLMs within visual contexts. Moreover, MMMU (Yue et al., 2023) and CMMMU (Zhang et al., 2024) have respectively introduced English and Chinese versions of the massive multi-discipline MM understanding and reasoning benchmark for expert artificial general intelligence. Fan et al. have also challenged MM-LLMs with multipanel VQA. BenchLMM (Cai et al., 2023) benchmarks the cross-style visual capability of MM-LLMs. Additionally, Liu et al. have conducted an in-depth study on the optical character recognition capabilities of MM-LLMs.

Mobile/Lightweight Deployment To deploy MM-LLMs on resource-constrained platforms and

achieve optimal performance meanwhile, such as low-power mobile and IoT devices, lightweight implementations are of paramount importance. A notable advancement in this realm is MobileVLM (Chu et al., 2023a). This approach strategically downscales LLaMA, allowing for seamless off-the-shelf deployment. MobileVLM further introduces a lightweight downsample projector, consisting of fewer than 20 million parameters, contributing to improved computational speed. Recently, there have been many similar studies on lightweighting MM-LLMs, achieving efficient computation and inference with comparable performance or minimal loss, including TinyGPT-V (Yuan et al., 2023b), Vary-toy (Wei et al., 2024), Mobile-Agent (Wang et al., 2024b), MoE-LLaVA (Lin et al., 2024a), and MobileVLM V2 (Chu et al., 2024). Nevertheless, this avenue necessitates additional exploration for further advancements in development.

Embodied Intelligence The embodied intelligence aims to replicate human-like perception and interaction with the surroundings by effectively understanding the environment, recognizing pertinent objects, assessing their spatial relationships, and devising a comprehensive task plan (Firoozi et al., 2023). Embodied AI tasks, such as embodied planning, embodied visual question answering, and embodied control, equip robots to autonomously implement extended plans by leveraging real-time observations. Some typical works in this area are PaLM-E (Driess et al., 2023) and EmbodiedGPT (Mu et al., 2023). PaLM-E introduces a multi-embodiment agent through the training of a MM-LLM. Beyond functioning solely as an embodied decision maker, PaLM-E also demonstrates proficiency in handling general VL tasks. EmbodiedGPT introduces an economically efficient method characterized by a CoT approach, enhancing the capability of embodied agents to engage with the real world and establishing a closed loop that connects high-level planning with low-level control. While MM-LLM-based Embodied Intelligence has made advancements in integrating with robots, further exploration is needed to enhance the autonomy of robots.

Continual Learning Due to the large training costs associated with their massive scale, MM-LLMs are not amenable to frequent re-training. However, updates are necessary to endow MM-LLMs with new skills and keep them up-to-date with rapidly evolving human knowledge (Wu et al., 2024). Thus, Continual Learning (CL) is needed to make the model flexible enough to efficiently and continually leverage emerging data while avoiding the substantial cost of retraining MM-LLMs. CL for MM-LLMs can be classified into two stages: continual PT and continual IT. Recently, a continual MM IT benchmark has been proposed to continuously fine-tune MM-LLMs for new MM tasks while maintaining superior performance on tasks learned during the original MM IT stage (He et al., 2023). It introduces two primary challenges: (1) catastrophic forgetting, where models forget previous knowledge when learning new tasks (Robins, 1995; McCloskey and Cohen, 1989; Goodfellow et al., 2013; Zhang et al., 2023d,c,b; Zheng et al., 2023a), and (2) negative forward transfer, indicating that the performance of unseen tasks declines when learning new ones (Zheng et al., 2024)

Mitigating Hallucination Hallucinations entail generating textual descriptions of nonexistent objects without visual cues, which manifest in diverse categories (Liu et al., 2024a) such as misjudgments and inaccuracies in descriptions. The origins of these hallucinations are multifaceted (Liu et al., 2024a), including biases and annotation errors in training data. Additionally, Skip $\backslash n$ (Han et al., 2024) highlights semantic drift biases associated with paragraph separators, which can induce hallucinations when deliberately inserted. Current methods to mitigate these hallucinations involve leveraging self-feedback as visual cues (Lee et al., 2023). However, challenges persist, necessitating nuanced discernment between accurate and hallucinatory outputs, as well as advancements in training methodologies to enhance output reliability.

7 Conclusion

In this paper, we have presented a comprehensive survey of MM-LLMs with a focus on recent advancements. Initially, we categorize the model architecture into five components, providing a detailed overview of general design formulations and training pipelines. Subsequently, we introduce various SOTA MM-LLMs, each distinguished by its specific formulations. Our survey also sheds light on their capabilities across diverse MM benchmarks and envisions future developments in this rapidly evolving field. We hope this survey can provide insights for researchers, contributing to the ongoing advancements in the MM-LLMs domain.

Limitations

In this paper, we embark on a comprehensive exploration of the current MM-LLMs landscape, presenting a synthesis from diverse perspectives enriched by our insights. Acknowledging the dynamic nature of this field, it is plausible that certain aspects may have eluded our scrutiny, and recent advances might not be entirely encapsulated. To tackle this inherent challenge, we've established a dedicated website for real-time tracking, using crowdsourcing to capture the latest advancements. Our goal is for this platform to evolve into a continuous source of contributions propelling ongoing development in the field. Given the constraints of page limits, we are unable to delve into all technical details and have provided concise overviews of the core contributions of mainstream MM-LLMs. Looking ahead, we commit to vigilant monitoring and continual enhancement of relevant details on our website, incorporating fresh insights as they emerge.

References

2023. Bliva: A simple multimodal llm for better handling of text-rich visual questions. *arXiv preprint arXiv:2308.09936*.

Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. 2023. Jointly Training Large Autoregressive Multimodal Models. *arXiv preprint arXiv:2309.15564*.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs. *arXiv preprint arXiv:2305.08844*.

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. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.

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 Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. 2022. Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models. In *International Conference on Learning Representations*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our Multimodal Models.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16548–16558.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. 2021. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR.

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. *Advances in neural information processing systems*, 33:1877–1901.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. 2023. BenchLMM: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*.

Cerspense. 2023. Zeroscope: Diffusion-based text-to-video synthesis.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023a. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023b. 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. 2023c. 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. 2023d. 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. 2023e. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023f. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023g. BEATs: Audio Pre-Training with Acoustic Tokenizers. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 5178–5193.

Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024. LLaVA-MoLE: Sparse Mixture of LoRA Experts for Mitigating Data Conflicts in Instruction Finetuning MLLMs. *arXiv preprint arXiv:2401.16160*.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022a. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678.

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, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023i. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023j. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

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. *Journal of Machine Learning Research*, 24(240):1–113.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023a. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

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*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023b. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

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*.

XTuner Contributors. 2023. XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Linhao Dong and Bo Xu. 2020. Cif: Continuous integrate-and-fire for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083. IEEE.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2024a. Dreamllm: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024b. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

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*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022a. A Survey of Vision-Language Pre-Trained Models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5436–5443.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022b. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Erich Elsen, Augustus Odena, Maxwell Nye, Sağnak Taşırlar, Tri Dao, Curtis Hawthorne, Deepak Moparthi, and Arushi Somani. 2023. Releasing Persimmon-8B.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. 2024. Muffin or Chihuahua? Challenging Large Vision-Language Models with Multipanel VQA. *arXiv preprint arXiv:2401.15847*.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

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*.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. 2023. Foundation Models in Robotics: Applications, Challenges, and the Future. *arXiv preprint arXiv:2312.07843*.

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*.

Chin-Lun Fu, Zih-Ching Chen, Yun-Ru Lee, and Hung-Yi Lee. 2022. AdapterBias: Parameter-efficient Token-dependent Representation Shift for Adapters in NLP Tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2608–2621.

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. *arXiv preprint arXiv:2304.14108*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.05935*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023a. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, and Qing Li. 2023b. CLOVA: A Closed-Loop Visual Assistant with Tool Usage and Update. *arXiv preprint arXiv:2312.10908*.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

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*.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Minglun Han, Feilong Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Knowledge Transfer from Pre-trained Language Models to Cif-based Speech Recognizers via Hierarchical Distillation. *arXiv preprint arXiv:2301.13003*.

Minglun Han, Linhao Dong, Zhenlin Liang, Meng Cai, Shiyu Zhou, Zejun Ma, and Bo Xu. 2022. Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8532–8536. IEEE.

Zongbo Han, Zechen Bai, Haiyang Mei, Qianli Xu, Changqing Zhang, and Mike Zheng Shou. 2024. Skip $\backslash n$: A simple method to reduce hallucination in large vision-language models. *arXiv preprint arXiv:2402.01345*.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. *arXiv preprint arXiv:2401.13919*.

Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. 2023. Continual instruction tuning for large multimodal models. *arXiv preprint arXiv:2311.16206*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a Unified View of Parameter-Efficient Transfer Learning. In *International Conference on Learning Representations*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

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 *International Conference on Learning Representations*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023b. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

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*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023c. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

IDEFICS. 2023. Introducing IDEFICS: An Open Reproduction of State-of-the-Art Visual Language Model.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2023. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping Vision-Language Learning with Decoupled Language Pre-training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. 2024. Unified language-vision pretraining with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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 *Thirty-seventh Conference on Neural Information Processing Systems*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR.

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. *International journal of computer vision*, 123:32–73.

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, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

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*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023e. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023f. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023g. Silkie: Preference Distillation for Large Visual Language Models. *arXiv preprint arXiv:2312.10665*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023h. M^3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv preprint arXiv:2306.04387*.

Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024a. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

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*.

Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, and Xiangde Liu. 2024b. 3DMIT: 3D Multi-modal Instruction Tuning for Scene Understanding. *arXiv preprint arXiv:2401.03201*.

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*.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. 2024c. LEGO: Language Enhanced Multi-modal Grounding Model. *arXiv preprint arXiv:2401.06071*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024a. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. *arXiv preprint arXiv:2401.15947*.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024b. GOAT-Bench: Safety Insights to Large Multimodal Models through Meme-Based Social Abuse. *arXiv preprint arXiv:2401.01523*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

LinkSoul-AI. 2023. Chinese-LLaVA.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. 2023b. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 21450–21474.

Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2023c. AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining. *CoRR*, abs/2308.05734.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023d. Improved Baselines with Visual Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

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 *Thirty-seventh Conference on Neural Information Processing Systems*.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023f. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023g. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023h. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Zhiheng Li, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, et al. 2023i. Controlllm: Augment language models with tools by searching on graphs. *arXiv preprint arXiv:2310.17796*.

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-Language Pretrained Models: A Survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5530–5537.

Junyu Lu, Ruyi Gan, Dixiang Zhang, Xiaojun Wu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2023a. Lyrics: Boosting Fine-grained Language-Vision Alignment and Comprehension via Semantic-aware Visual Objects. *arXiv preprint arXiv:2312.05278*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yujie Lu, Xiujun Li, William Yang Wang, and Yejin Choi. 2023b. Vim: Probing multimodal large language models for visual embedded instruction following. *arXiv preprint arXiv:2311.17647*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. *arXiv preprint arXiv:2306.08568*.

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*.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2023. 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*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. KAM-CoT: Knowledge Augmented Multimodal Chain-of-Thoughts Reasoning. *arXiv preprint arXiv:2401.12863*.

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*.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Ziyi Ni, Minglun Han, Feilong Chen, Linghui Meng, Jing Shi, Pin Lv, and Bo Xu. 2024. VILAS: Exploring the Effects of Vision and Language Context in Automatic Speech Recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

OpenAI. 2022. OpenAI: Introducing ChatGPT.

OpenAI. 2023. GPT-4 Technical Report.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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*.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. 2024. CogCoM: Train Large Vision-Language Models Diving into Details through Chain of Manipulations. *arXiv preprint arXiv:2402.04236*.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. 2024. DiffusionGPT: LLM-Driven Text-to-Image Generation System. *arXiv preprint arXiv:2401.10061*.

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 *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 28492–28518.

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. *The Journal of Machine Learning Research*, 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*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

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

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Ludan Ruan and Qin Jin. 2022. Survey: Transformer based video-language pre-training. *AI Open*, 3:1–13.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. AudioPaLM: A Large Language Model That Can Speak and Listen. *arXiv preprint arXiv:2306.12925*.

Salesforce. 2022. Ulip.

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. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. 2022b. Laion coco: 600m synthetic captions from laion2b-en.

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. *arXiv preprint arXiv:2111.02114*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024. Small llms are weak tool learners: A multi-llm agent. *arXiv preprint arXiv:2401.07324*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Shezheng Song, Xiaopeng Li, and Shasha Li. 2023. How to Bridge the Gap between Modalities: A Comprehensive Survey on Multimodal Large Language Model. *arXiv preprint arXiv:2311.07594*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

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. 2024. Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023b. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023a. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

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

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023c. Any-to-Any Generation via Composable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. 2022. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Gemini Team, 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*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Yi Team. 2023. Yi-VL.

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

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chenyu Wang, Weixin Luo, Qianyu Chen, Haonan Mai, Jindi Guo, Sixun Dong, Zhengxin Li, Lin Ma, Shenghua Gao, et al. 2024a. Tool-LMM: A Large Multi-Modal Model for Tool Agent Learning. *arXiv preprint arXiv:2401.10727*.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023a. DocLLM: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024b. Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. *arXiv preprint arXiv:2401.16158*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. 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. 2023c. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023d. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

Xinyu Wang, Bohan Zhuang, and Qi Wu. 2024c. ModaVerse: Efficiently Transforming Modalities with LLMs. *arXiv preprint arXiv:2401.06395*.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Small Language Model Meets with Reinforced Vision Vocabulary. *arXiv preprint arXiv:2401.12503*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

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*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023b. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. 2017. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023c. Multimodal large language models: A survey. *arXiv preprint arXiv:2311.13165*.

Penghao Wu and Saining Xie. 2023. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 17.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023d. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual Learning for Large Language Models: A Survey. *arXiv preprint arXiv:2402.01364*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023e. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Rui Yan, Mike Zheng Shou, Yixiao Ge, Alex Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. 2021. Video-text pre-training with learned regions. *arXiv preprint arXiv:2112.01194*.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. ViGoR: Improving Visual Grounding of Large Vision Language Models with Fine-Grained Reward Modeling. *arXiv preprint arXiv:2402.06118*.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. *arXiv preprint arXiv:2401.11708*.

Zhen Yang, Yingxue Zhang, Fandong Meng, and Jie Zhou. 2023a. TEAL: Tokenize and Embed ALL for Multi-modal Large Language Models. *arXiv preprint arXiv:2311.04589*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. 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*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. 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. 2023c. mplug-owl2: Revolutionizing multi-modal 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*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023b. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*.

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. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

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*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023c. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2023a. Osprey: Pixel Understanding with Visual Instruction Tuning. *arXiv preprint arXiv:2312.10032*.

Zhengqing Yuan, Zhaoxu Li, and Lichao Sun. 2023b. TinyGPT-V: Efficient Multimodal Large Language Model via Small Backbones. *arXiv preprint arXiv:2312.16862*.

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*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022a. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. 2023. What Matters in Training a GPT4-Style Language Model with Multimodal Inputs? *arXiv preprint arXiv:2307.02469*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022b. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773.

Duzhen Zhang, Wei Cong, Jiahua Dong, Yahan Yu, Xiuyi Chen, Yonggang Zhang, and Zhen Fang. 2023b. Continual Named Entity Recognition without Catastrophic Forgetting. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Duzhen Zhang, Hongliu Li, Wei Cong, Rongtao Xu, Jiahua Dong, and Xiuyi Chen. 2023c. Task relation distillation and prototypical pseudo label for incremental named entity recognition. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3319–3329.

Duzhen Zhang, Yahan Yu, Feilong Chen, and Xiuyi Chen. 2023d. Decomposing Logits Distillation for Incremental Named Entity Recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1919–1923.

Duzhen Zhang, Tielin Zhang, Shuncheng Jia, Qingyu Wang, and Bo Xu. 2022a. Recent Advances and New Frontiers in Spiking Neural Networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5670–5677.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. 2024. CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark. *arXiv preprint arXiv:2401.11944*.

Hang Zhang, Xin Li, and Lidong Bing. 2023e. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. 2022b. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *The Eleventh International Conference on Learning Representations*.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. 2020. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 698–714. Springer.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023f. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

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. 2022c. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023h. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023i. Recognize Anything: A Strong Image Tagging Model. *arXiv preprint arXiv:2306.03514*.

Bingchen Zhao, Haoqin Tu, Chen Wei, and Cihang Xie. 2024. Tuning LayerNorm in Attention: Towards Efficient Multimodal LLM Finetuning. In *The Twelfth International Conference on Learning Representations*.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

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*.

Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. 2022. EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023c. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023d. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.

Junhao Zheng, Qianli Ma, Zhen Liu, Binquan Wu, and Huawen Feng. 2024. Beyond Anti-Forgetting: Multimodal Continual Instruction Tuning with Positive Forward Transfer. *arXiv preprint arXiv:2401.09181*.

Junhao Zheng, Shengjie Qiu, and Qianli Ma. 2023a. Learn or Recall? Revisiting Incremental Learning with Pre-trained Language Models. *arXiv preprint arXiv:2312.07887*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023b. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2024a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*.

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*.

Dongsheng Zhu, Xunzhu Tang, Weidong Han, Jinghui Lu, Yukun Zhao, Guoliang Xing, Junfeng Wang, and Dawei Yin. 2024b. VisLingInstruct: Elevating Zero-Shot Learning in Multi-Modal Language Models with Autonomous Instruction Optimization. *arXiv preprint arXiv:2402.07398*.

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*.

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

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. 2024c. LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model. *arXiv preprint arXiv:2401.02330*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. 2023. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. *arXiv preprint arXiv:2402.02207*.

A Related Surveys

Prior to the emergence of LLMs, several surveys on traditional MM PT have been conducted (Ruan and Jin, 2022; Du et al., 2022a; Long et al., 2022; Chen et al., 2023a). Most of these models entail a substantial computational cost during the PT phase, attributable to end-to-end training using large-scale models and datasets. As a consequence of not incorporating LLMs, these models suffer from deficiencies in instruction following, ICL, CoT, and interactive capabilities. Moreover, the training pipeline solely encompasses the PT phase without the inclusion of an IT stage.

In recent times, several surveys have emerged on MM-LLMs. Yin et al. and Wu et al. exclusively delve into early VL understanding models. Huang et al. place a primary emphasis on visual IT, while Song et al. focus on modal alignment methods. Lastly, Cui et al. provide a comprehensive review of the applications of MM-LLMs within the realm of autonomous driving.

Compared with their works, the main distinctions are outlined as follows:

- We have comprehensively covered nearly all MM-LLMs over the past year, totaling around 120 or more, including not only understanding models but also generative models. Our coverage extends beyond VL modalities to encompass various modes such as audio and 3D point cloud;
- To offer readers a comprehensive understanding of MM-LLMs, we have introduced a general model architecture that incorporates any-to-any modality transformations, offering a detailed overview of the functional roles and implementation choices for each component;
- We have summarized the developmental trends of existing MM-LLMs and provided some training recipes that can enhance effectiveness;
- We have established an open-source website for MM-LLMs researchers, supporting crowd-sourced updates and aiming to facilitate collaboration in the MM-LLMs field. We anticipate that this survey will illuminate future research in the MM-LLMs domain.

B Modality Encoder

In the following, we provide a brief introduction to some mainstream modality encoders.

B.1 Visual Modality

NFNet-F6 (Brock et al., 2021) is a normalizer-free ResNet (He et al., 2016), showcasing an adaptive gradient clipping that allows training on extensively augmented datasets while achieving SOTA levels of image recognition.

ViT (Dosovitskiy et al., 2020) applies the Transformer (Vaswani et al., 2017) to images by first dividing the image into patches. It then undergoes linear projection to flatten the patches, followed by encoding via Transformer blocks.

CLIP ViT (Radford et al., 2021) builds connections between text and images, comprising a ViT and a text encoder. With a vast amount of text-image pairs, it optimizes ViT by contrastive learning, treating paired text and images as positive samples and others as negative ones.

Eva-CLIP ViT (Fang et al., 2023) stabilizes the training and optimization process of the massive CLIP, offering new directions in expanding and accelerating the expensive training of MM base models.

B.2 Audio Modality

C-Former (Chen et al., 2023b) employs the CIF (Dong and Xu, 2020; Zhang et al., 2022a; Han et al., 2022, 2023) for sequence transduction and a Transformer to extract audio features.

HuBERT (Hsu et al., 2021) is a self-supervised speech representation learning framework based on BERT (Kenton and Toutanova, 2019), achieved by the masked prediction of discrete hidden units. It has the capability to convert continuous speech signals into a sequence of discrete units.

BEATs (Chen et al., 2023g) is an iterative audio pre-training framework designed to learn Bidirectional Encoder representations from Audio Transformers.

C Mainstream PEFT Methods

PEFT entails maintaining the pre-trained LLM in a frozen state while adjusting a small number of additional trainable parameters. In the following section, we revisit several representative PEFT methods, where $\boldsymbol{x}$ and $\boldsymbol{h}$ represent the input and output

of the original module, and $\boldsymbol{h}'$ signifies the output of this module when attached with PEFT.

Prefix-tuning (Li and Liang, 2021; Lester et al., 2021) involves the addition of learnable tokens to the keys and values of the attention module. This process is formulated as follows:

$$\mathbf{h}' = \text{Attn}\left(\mathbf{x}\mathbf{W}_q, [\mathbf{P}_k, \mathbf{x}\mathbf{W}_k], [\mathbf{P}_v, \mathbf{x}\mathbf{W}_v]\right), \quad (6)$$

with $\mathbf{P}_k, \mathbf{P}_v \in \mathbb{R}^{l \times d}$ representing two sets of prefix tokens. $[\cdot, \cdot]$ denotes concatenation, and Attn is defined as:

$$\text{Attn}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) := \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}.$$

Adapter (Houlsby et al., 2019; He et al., 2021; Rebuffi et al., 2017; Zhang et al., 2020) is typically a residual block consisting of a down-projection matrix $\mathbf{A}$, a nonlinear activation function $\sigma(\cdot)$, and an up-projection matrix $\mathbf{B}$. It can be inserted into any layer of the pre-trained LLM, formulated as follows:

$$\mathbf{h}' = \mathbf{h} + \sigma(\mathbf{x}\mathbf{A})\mathbf{B}. \quad (7)$$

LoRA (Hu et al., 2021) is the most commonly used PEFT method. It assumes that the change in parameters occurs within a low-rank space. Given a pre-trained matrix $\boldsymbol{W} \in \mathbb{R}^{c \times d}$, LoRA learns an incremental update $\Delta\boldsymbol{W}$ and decomposes $\Delta\boldsymbol{W}$ into a matrix multiplication between two low-rank matrices $\boldsymbol{A} \in \mathbb{R}^{c \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r \times d}$, where $r \ll \min(c, d)$. LoRA follows the forward process as outlined below:

$$\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x} + \Delta\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}. \quad (8)$$

QLoRA (Dettmers et al., 2023) is a quantized LoRA. The underlying principle of QLoRA includes the quantization of pre-trained weights to 4 bits, followed by the execution of PEFT using LoRA.

LayerNorm tuning (Zhao et al., 2024) presents an efficient strategy to transform LLMs into MM-LLMs, which tunes LayerNorm in attention block yielding strong MM performance compared with full parameter finetuning or LoRA.

In addition to the aforementioned PEFT methods, there are several others, including P-tuning (Liu et al., 2022), P-tuning v2 (Liu et al., 2021a), AdaptBias (Fu et al., 2022), Compacter (Karimi Mahabadi et al., 2021), AdapterFormer (Chen et al., 2022a), XTuner (Contributors, 2023), P-LoRA (Dong et al., 2024b), MoLE (Chen et al., 2024), and Delta-LoRA (Zi et al., 2023).

D Representative LLMs

The representative LLM Backbones in existing MM-LLMs research are as follows:

Flan-T5 (Chung et al., 2022) investigates IT for T5 (Raffel et al., 2020), an encoder-decoder architecture using unified text-to-text training for all natural language processing issues, exhibiting robust zero-shot and CoT capabilities.

ChatGLM is a Chinese-English bilingual dialogue model,[2] optimized by an auto-regressive mask infilling objective. It is based on the GLM (Du et al., 2022b; Zeng et al., 2022a) architecture, optimized for Chinese question answering and dialogues.

InternLM (Team, 2023) is a multilingual trillion-parameter foundation model trained on over a trillion tokens of data. Based on this foundation, the model utilizes high-quality human-annotated dialogue data combined with RLHF to respond to complex instructions during human interactions, exhibiting responses that align with human ethics and values.

UL2 (Tay et al., 2022) is an encoder-decoder model trained utilizing a mixture of denoisers objectives, surpassing T5 on numerous benchmarks.

Qwen (Bai et al., 2023a) is trained on large-scale and diverse datasets, with a primary focus on Chinese and English. It employs SFT and RLHF techniques for alignment, resulting in dialogue models like Qwen-Chat.

Chinchilla (Hoffmann et al., 2022) is a causal decoder, trained on extensive text data. It posits that model size should double for every doubling of training tokens.

OPT (Zhang et al., 2022c) is a GPT-3 (Brown et al., 2020) clone, striving to release an open-source model that replicates the performance of GPT-3.

PaLM (Chowdhery et al., 2023) is a causal decoder structure with parallel attention and feed-forward layers, enabling training speeds up to 15 times faster. Notable changes contain RoPE embeddings, SwiGLU activation, multi-query attention, and etc.

[2] https://github.com/THUDM/ChatGLM-6B

LLaMA (Touvron et al., 2023a) comprises decoder-only models with efficient causal attention.

LLaMA-2 (Touvron et al., 2023b) focuses on fine-tuning a superior and safer LLaMA-2-Chat model for conversation generation, incorporating 40% more training data with grouped-query attention and a larger context length.

Vicuna (Chiang et al., 2023) is a model built on top of LLaMA, utilizing user dialogue data obtained from ShareGPT.com and trained by SFT.

E SOTA MM-LLMs

In the following, we will provide a brief introduction to the core contributions of some representative MM-LLMs.

Flamingo (Alayrac et al., 2022) represents a series of Visual Language (VL) Models designed for processing interleaved visual data and text, generating free-form text as the output.

BLIP-2 (Li et al., 2023e) introduces a more resource-efficient framework, comprising the lightweight Q-Former to bridge modality gaps and the utilization of frozen LLMs. Leveraging LLMs, BLIP-2 can be guided for zero-shot image-to-text generation using natural language prompts.

LLaVA (Liu et al., 2023e) pioneers the transfer of IT techniques to the MM domain. Addressing data scarcity, LLaVA introduces a novel open-source MM instruction-following dataset created using ChatGPT/GPT-4, alongside the MM instruction-following benchmark, LLaVA-Bench.

MiniGPT-4 (Zhu et al., 2023a) proposes a streamlined approach where training only one linear layer aligns the pre-trained vision encoder with the LLM. This efficient method enables the replication of the exhibited capabilities of GPT-4.

mPLUG-Owl (Ye et al., 2023b) presents a novel modularized training framework for MM-LLMs, incorporating the visual context. To assess different models' performance in MM tasks, the framework includes an instructional evaluation dataset called OwlEval.

X-LLM (Chen et al., 2023b) is expanded to various modalities, including audio, and demonstrates strong scalability. Leveraging the language transferability of the Q-Former, X-LLM is successfully applied in the context of Sino-Tibetan Chinese.

VideoChat (Li et al., 2023f) pioneers an efficient chat-centric MM-LLM for video understanding dialogue, setting standards for future research in this domain and offering protocols for both academia and industry.

InstructBLIP (Dai et al., 2023) is trained based on the pre-trained BLIP-2 model, updating only the Q-Former during MM IT. By introducing instruction-aware visual feature extraction and corresponding instructions, the model enables the extraction of flexible and diverse features.

PandaGPT (Su et al., 2023) is a pioneering general-purpose model with the capability to comprehend and act upon instructions across 6 different modalities: text, image/video, audio, thermal, depth, and inertial measurement units.

(PaLI-X (Chen et al., 2023h) is trained using mixed VL objectives and unimodal objectives, including prefix completion and masked-token completion. This approach proves effective for both downstream task results and achieving the Pareto frontier in the fine-tuning setting.

Video-LLaMA (Zhang et al., 2023e) introduces a multi-branch cross-modal PT framework, enabling LLMs to simultaneously process the vision and audio content of a given video while engaging in conversations with humans. This framework aligns vision with language as well as audio with language.

Video-ChatGPT (Maaz et al., 2023) is a model specifically designed for video conversations, capable of generating discussions about videos by integrating spatiotemporal vision representations.

Shikra (Chen et al., 2023e) introduces a simple and unified pre-trained MM-LLM tailored for Referential Dialogue, a task involving discussions about regions and objects in images. This model demonstrates commendable generalization ability, effectively handling unseen settings.

DLP (Jian et al., 2023) proposes the P-Former to predict the ideal prompt, trained on a dataset of single-modal sentences. This showcases the feasibility of single-modal training to enhance MM learning.

BuboGPT (Zhao et al., 2023d) is a model constructed by learning a shared semantic space for a comprehensive understanding of MM content. It

explores fine-grained relationships among different modalities such as image, text, and audio.

ChatSpot (Zhao et al., 2023b) introduces a simple yet potent method for finely adjusting precise referring instructions for MM-LLM, facilitating fine-grained interactions. The incorporation of precise referring instructions, consisting of image- and region-level instructions, enhances the integration of multi-grained VL task descriptions.

Qwen-VL (Bai et al., 2023b) is a multi-lingual MM-LLM that supports both English and Chinese. Qwen-VL also allows the input of multiple images during the training phase, improving its ability to understand the vision context.

NExT-GPT (Wu et al., 2023d) is an end-to-end, general-purpose any-to-any MM-LLM that supports the free input and output of image, video, audio, and text. It employs a lightweight alignment strategy, utilizing LLM-centric alignment in the encoding phase and instruction-following alignment in the decoding phase.

MiniGPT-5 (Zheng et al., 2023b) is an MM-LLM integrated with inversion to generative vokens and integration with Stable Diffusion. It excels in performing interleaved VL outputs for MM generation. The inclusion of classifier-free guidance during the training phase enhances the quality of generation.

LLaVA-1.5 (Liu et al., 2023d) reports simple modifications to the LLaVA framework, including applying an MLP projection and introducing VQA data tailored for academic tasks, along with simple response formatting prompts. These adjustments result in enhanced capabilities for MM understanding.

MiniGPT-v2 (Chen et al., 2023d) is an MM-LLM designed as a unified interface for diverse VL multi-task learning. To create a single model proficient in handling multiple VL tasks, identifiers are incorporated for each task during both training and inference. This facilitates clear task distinction, ultimately enhancing learning efficiency.

CogVLM (Wang et al., 2023b) is an open-source MM-LLM that bridges the gap between modalities via a trainable visual expert module within the attention and feedforward layers. This allows for a deep fusion of MM features without compromising performance on NLP downstream tasks.

DRESS (Chen et al., 2023i) introduces a method using natural language feedback to enhance alignment with human preferences. DRESS extends the conditional reinforcement learning algorithm to integrate non-differentiable natural language feedback, training the model to generate appropriate responses based on feedback.

X-InstructBLIP (Panagopoulou et al., 2023) introduces a cross-modal framework with instruction-aware representations, scalable enough to empower LLMs to handle diverse tasks across multiple modalities, including image/video, audio, and 3D. Notably, it achieves this without the need for modality-specific PT.

CoDi-2 (Tang et al., 2023b) is a MM generation model excelling in modality-interleaved instruction following, in-context generation, and user-model interaction by multi-turn conversations. It enhances CoDi (Tang et al., 2023c) to process intricate modality-interleaved inputs and instructions, generating latent features autoregressively.

VILA (Lin et al., 2023) outperforms in vision tasks and shows remarkable reasoning ability while maintaining text-only capabilities. It achieves this by harnessing the full capabilities of LLM learning, using the interleaved attributes of image-text pairs, and implementing meticulous text data re-blending.

F VL Benchmarks

The 18 VL benchmarks presented in Table 2 include **OKVQA** (Schwenk et al., 2022), **IconVQA** (Lu et al., 2021), **VQA^{v2}** (Goyal et al., 2017), **GQA** (Hudson and Manning, 2019), **VizWiz** (Gurari et al., 2018), **SQA^{I}**: ScienceQA-IMG (Lu et al., 2022), **VQA^{T}**: TextVQA (Singh et al., 2019), **POPE** (Li et al., 2023k), **MME^{P}**: MME Perception (Fu et al., 2023), **MME^{C}**: MME Cognition (Fu et al., 2023), **MMB**: MMBenchmark (Liu et al., 2023g), **MMB^{CN}**: MMBench-Chinese (Liu et al., 2023g), **SEED^{I}**: SEED-Bench (Image) (Li et al., 2023c), **LLaVA^{W}**: LLaVA-Bench (In-the-Wild) (Liu et al., 2023a), **MM-Vet** (Yu et al., 2023c), **QBench** (Wu et al., 2023b), **HM**: HatefulMemes (Kiela et al., 2020), and **VSR** (Liu et al., 2023a).

G Training Dataset

The statistics for MM PT and MM IT dataset are presented in Table 3 and Table 4, respectively.

Dataset Name	X Modality	#.X	#.T	#.X-T
ALIGN (Jia et al., 2021)	Image	1.8B	1.8B	1.8B
LTIP (Alayrac et al., 2022)	Image	312M	312M	312M
MS-COCO (Lin et al., 2014)	Image	124K	620K	620K
Visual Genome (Krishna et al., 2017)	Image	108K	4.5M	4.5M
CC3M (Sharma et al., 2018)	Image	3.3M	3.3M	3.3M
CC12M (Changpinyo et al., 2021)	Image	12.4M	12.4M	12.4M
SBU (Ordonez et al., 2011)	Image	1M	1M	1M
LAION-5B (Schuhmann et al., 2022)	Image	5.9B	5.9B	5.9B
LAION-400M (Schuhmann et al., 2021)	Image	400M	400M	400M
LAION-en (Schuhmann et al., 2022)	Image	2.3B	2.3B	2.3B
LAION-zh (Schuhmann et al., 2022)	Image	142M	142M	142M
LAION-COCO (Schuhmann et al., 2022b)	Image	600M	600M	600M
Flickr30k (Young et al., 2014)	Image	31K	158K	158K
AI Challenger Captions (Wu et al., 2017)	Image	300K	1.5M	1.5M
COYO (Byeon et al., 2022)	Image	747M	747M	747M
Wukong (Gu et al., 2022)	Image	101M	101M	101M
COCO Caption (Chen et al., 2015)	Image	164K	1M	1M
WebLI (Chen et al., 2022b)	Image	10B	12B	12B
Episodic WebLI (Chen et al., 2023h)	Image	400M	400M	400M
CC595k (Liu et al., 2023e)	Image	595K	595K	595K
RefCOCO (Kazemzadeh et al., 2014)	Image	20K	142K	142K
RefCOCO+ (Yu et al., 2016)	Image	20K	142K	142K
Visual-7W (Zhu et al., 2016)	Image	47.3K	328K	328K
OCR-VQA (Mishra et al., 2019)	Image	207K	1M	1M
ST-VQA (Biten et al., 2022)	Image	23K	32K	32K
DocVQA (Mathew et al., 2021)	Image	12K	50K	50K
TextVQA (Singh et al., 2019)	Image	28.4K	45.3K	45.3K
DataComp (Gadre et al., 2023)	Image	1.4B	1.4B	1.4B
GQA (Hudson and Manning, 2019)	Image	113K	22M	22M
VGQA (Krishna et al., 2017)	Image	108K	1.7M	1.7M
VQA^{v2} (Goyal et al., 2017)	Image	265K	1.4M	1.4M
DVQA (Kafle et al., 2018)	Image	300K	3.5M	3.5M
OK-VQA (Schwenk et al., 2022)	Image	14K	14K	14K
A-OKVQA (Schwenk et al., 2022)	Image	23.7K	24.9K	24.9K
Text Captions (Sidorov et al., 2020)	Image	28K	145K	145K
M3W (Interleaved) (Alayrac et al., 2022)	Image	185M	182GB	43.3M (Instances)
MMC4 (Interleaved) (Zhu et al., 2023c)	Image	571M	43B	101.2M (Instances)
Obelics (Interleaved) (Laurençon et al., 2023)	Image	353M	115M	141M (Instances)
MSRVTT (Xu et al., 2016)	Video	10K	200K	200K
WebVid (Bain et al., 2021)	Video	10M	10M	10M
VTP (Alayrac et al., 2022)	Video	27M	27M	27M
AISHELL-1 (Chen et al., 2023b)	Audio	–	–	128K
AISHELL-2 (Chen et al., 2023b)	Audio	–	–	1M
WaveCaps (Mei et al., 2023)	Audio	403K	403K	403K
VSDial-CN (Ni et al., 2024)	Image, Audio	120K (Image), 1.2M(Audio)	120K	1.2M

Table 3: The statistics for MM PT datasets. **#.X** represents the quantity of X, **#.T** represents the quantity of Text, and **#.X-T** represents the quantity of X-Text pairs, where X can be Image, Video, or Audio.

Dataset Name	Type	I→O	Source	Method	Multi-Turn	#.I/V/A	#.Dialog Turn	#.Instance
MiniGPT-4's IT (Zhu et al., 2023a)	SFT	I+T→T	CC3M, CC12M	Auto.	✗	134M/–/–	1	5K
StableLLaVA (Li et al., 2023i)	SFT	I+T→T	SD (Rombach et al., 2022)	Auto.+Manu.	✗	126K/–/–	1	126K
LLaVA's IT (Zhang et al., 2023h)	SFT	I+T→T	MS-COCO	Auto.	✓	81K/–/–	2.29	150K
SVIT (Zhao et al., 2023a)	SFT	I+T→T	MS-COCO, Visual Genome	Auto.	✓	108K/–/–	5	3.2M
LLaVAR's IT (Zhang et al., 2023h)	SFT	I+T→T	MS-COCO, CC3M, LAION	LLaVA+Auto.	✓	20K/–/–	2.27	174K
ShareGPT4V's IT (Chen et al., 2023f)	SFT	I+T→T	LCS, COCO, SAM, TextCaps, WikiArt	Auto.+Manu.	✗	100K/–/–	–	–
DRESS's IT (Chen et al., 2023i)	SFT	I+T→T	LLaVA's IT, VLSafe	Auto.+Manu.	✓	193K/–/–	∼4	–
VideoChat's IT (Li et al., 2023f)	SFT	V+T→T	WebVid	Auto.	✓	–/8K/–	1.82	11K
Video-ChatGPT's IT (Maaz et al., 2023)	SFT	V+T→T	ActivityNet (Caba Heilbron et al., 2015)	Inherit	✓	–/100K/–	1	100K
Video-LLaMA's IT (Zhang et al., 2023e)	SFT	I/V+T→T	MiniGPT-4, LLaVA, and VideoChat's IT	Auto.	✓	81K/8K/–	2.22	171K
InstructBLIP's IT (Dai et al., 2023)	SFT	I/V+T→T	Multiple (InstructBLIP's Figure 2)	Auto.	✗	–	–	∼1.6M
X-InstructBLIP's IT (Panagopoulou et al., 2023)	SFT	I/V/A/3D+T→T	Multiple (X-InstructBLIP's Figure 4)	Auto.	✗	–	–	∼1.8M
MIMIC-IT (Li et al., 2023a)	SFT	I/V+T→T	Multiple	Auto.	✗	8.1M/502K/–	1	2.8M
PandaGPT's IT (Su et al., 2023)	SFT	I+T→T	MiniGPT-4 and LLaVA's IT	Inherit	✓	81K/–/–	2.29	160K
MGVLID (Zhao et al., 2023b)	SFT	I+B+T→T	Multiple	Auto.+Manu.	✗	108K/–/–	–	108K
M^3IT (Li et al., 2023h)	SFT	I/V/B+T→T	Multiple	Auto.+Manu.	✗	–/–/–	1	2.4M
LAMM (Yin et al., 2023b)	SFT	I+3D+T→T	Multiple	Auto.+Manu.	✓	91K/–/–	3.27	196K
BuboGPT's IT (Zhao et al., 2023d)	SFT	(I+A)/A+T→T	Clotho, VGGSS	Auto.	✗	5K/–/9K	–	9K
mPLUG-DocOwl's IT (Ye et al., 2023b)	SFT	I/Tab/Web+T→T	Multiple	Inherit	✗	–	–	–
T2M (Wu et al., 2023d)	SFT	T→I/V/A+T	WebVid, CC3M, AudioCap	Auto.	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT (Wu et al., 2023d)	SFT	I+V+A+T→I+V+A+T	Youtube, Google, Flickr30k, Midjourney, etc.	Auto.+Manu.	✓	4K/4K/4K	4.8	5K
Osprey's IT (Yuan et al., 2023a)	SFT	I+T→T	MS-COCO, RefCOCO, RefCOCO+, LLaVA's IT etc. (fine-grained region-text dataset)	Auto.+Manu.	✓	–/–/–	∼4	724K
LLaVA-RLHF (Sun et al., 2023b)	RLHF	I+T→T	Collected human preference	Manu.	✗	–/–/–	–	10K
DRESS's IT (Chen et al., 2023i)	RLHF	I+T→T	LLaVA's IT, VLSafe	Auto.+Manu.	✓	33K/–/–	∼4	–
RLHF-V's IT (Yu et al., 2023b)	RLHF	I+T→T	Collected human preference	Manu.	✗	–/–/–	–	1.4K
VLFeedback (Li et al., 2023g)	RLHF	I+T→T	Responses generated by 12 MM-LLMs	Auto.	✗	–/–/–	–	80K
RTVLM (Li et al., 2024a)	RLHF	I+T→T	New question-image pairs based on publicly available images or originally diffusion-generated images (Gallegos et al., 2023)	Auto.+Manu.	✗	–/–/–	–	5K
VLGuard's IT (Zong et al., 2024)	RLHF	I+T→T	Source image data from various datasets	Auto.	✗	3K/–/–	–	3K
MMViG (Yan et al., 2024)	RLHF	I+T→T	MS-COCO	Manu.	✗	16K/–/–	–	16K

Table 4: The statistics for MM IT datasets. I→O: Input to Output Modalities, T: Text, I: Image, V: Video, A: Audio, B: Bounding box, 3D: Point Cloud, Tab: Table, and Web: Web page.