Continuous or Discrete, That Is the Question: A Survey on Large Multi-Modal Models from the Perspective of Input-Output Space Extension

Zejun Li , Jiwen Zhang , Dianyi Wang , Ye Wang , Xuanjing Huang , Zhongyu Wei *

Posted Date: 11 November 2024

doi: 10.20944/preprints202411.0685.v1

Keywords: Large Multi-modal Model; Input-Output Space Extension

 doi:10.20944/preprints202411.0685.v1

Article

Continuous or Discrete, That Is the Question: A Survey on Large Multi-Modal Models from the Perspective of Input-Output Space Extension

Zejun Li [1,†], Jiwen Zhang [1,†], Dianyi Wang [2,4,†], Ye Wang [3,†], Xuanjing Huang [5] and Zhongyu Wei [1,2,4,*]

[1] School of Data Science, Fudan University
[2] Research Institute of Intelligent Complex Systems, Fudan University
[3] Academy for Engineering and Technology, Fudan University
[4] Shanghai Innovation Institute
[5] School of Computer Science, Fudan University
* zywei@fudan.edu.cn
† These authors contributed equally.

Abstract: With the success of large language models (LLMs) driving progress towards general-purpose AI, there has been a growing focus on extending these models to multi-modal domains, giving rise to large multi-modal models (LMMs). Unlike existing reviews that focuses on specific model frameworks or scenarios, this survey summarizes and provides insights into the current research on LMMs from a more general perspective, **input-output space extension**. Particularly, we discuss the following questions: (i) How to construct multi-modal input-output spaces with discretely or continuously encoded modality signals? (ii) How to design model architectures and corresponding training strategies to align the constructed multi-modal representation space? (iii) How to comprehensively evaluate LMMs based on the expanded input-output space? We hope to provide an intuitive and comprehensive overview and inspire future work.

1. Introduction

The goal of AI research is to build versatile intelligent systems capable of fulfilling tasks across diverse scenarios. Recently, the generalization and interactivity demonstrated by large language models (LLMs), which leverage language instructions as the interface between users and machines, have significantly advanced the progress towards general-purpose AI [1–4]. To extend these capabilities to multi-modal contexts, research on large multi-modal models (LMMs) is emerging, aiming to expand the input and output space of the language-based interface to more modalities. As shown in Figure 1, to extend the input space, existing methods introduce discretely or continuously encoded modality representations into the text input and learn cross-modal alignment from multi-modal intertwined data, enabling LMMs to understand multi-modal information [5–8]. Similarly, the output space can be divided into multiple subspaces of different modalities, which are further aligned with corresponding modality decoders to generate multi-modal content [9–12].

Although there are several surveys that detail the current progress in constructing LMMs, most of these works are limited to specific perspectives. (1) Some studies merely introduce the input-side extension, lacking discussion on the extension of outputs [13,14]. (2) Certain studies solely discuss specific sub-problems in the construction of LMMs, such as applications in specific modalities [15,16] and scenarios [17,18], evaluation [19,20], and data [21]. (3) Meanwhile, most existing reviews focus on a specific type of model framework: encoding information from other modalities in a continuous manner and aligning them with text embeddings through connection modules, neglecting related research on other architectures, such as unified discretely represented LMMs [12,22]. These limitations prevent existing reviews from adequately covering research problems in LMM construction and limit their applicability to a broader scope.

 doi:10.20944/preprints202411.0685.v1

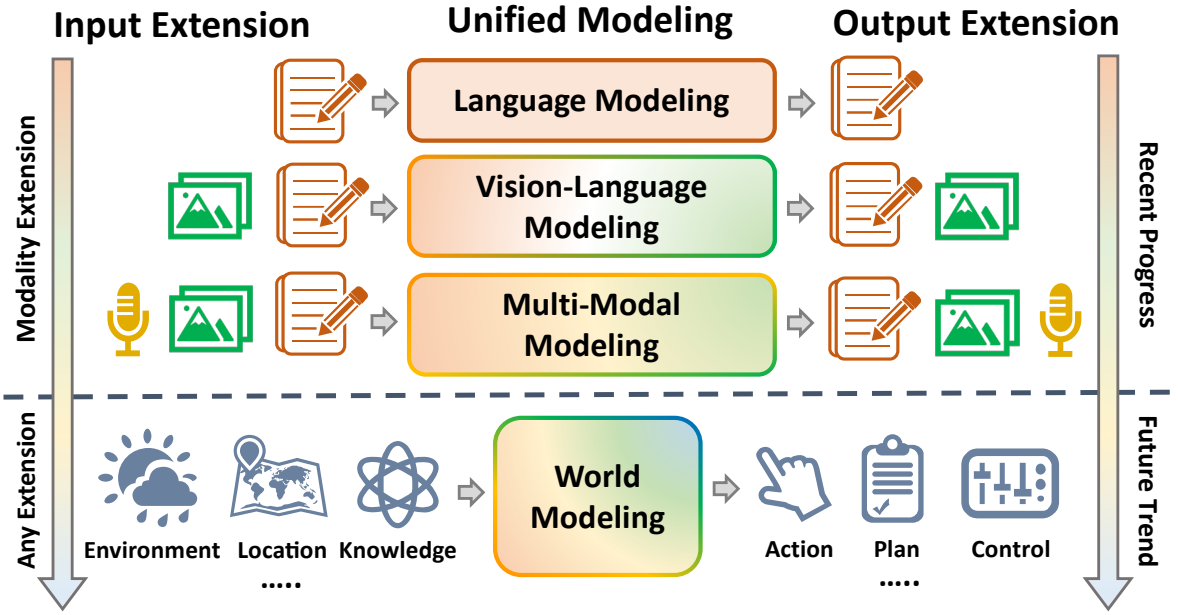

Figure 1. Illustration of the general LMM framework: expanding inputs and outputs to more modalities and aligning representations across modalities through unified multi-modal modeling.

To this end, this survey aims to summarize related works from a more general perspective: **the extension of the input-output space**. As illustrated in Figure 1, existing LMMs can be systematically summarized from this perspective, encompassing various modalities, scenarios, and model architectures, while also leaving room for further exploration to more modalities and scenarios.

To conduct a comprehensive survey, we follow a top-down logic to break down the construction of LMMs into several sub-problems, providing detailed discussions to offer insights to readers. Particularly, we try to answer the following questions. (i) How can modality signals be encoded using discrete or continuous representations, and how to construct multi-modal input-output spaces? (§3) (ii) How to design model architectures and corresponding training strategies to align the constructed multi-modal representation space? (§4) (iii) How to comprehensively evaluate LMMs based on the expanded input-output space? (§5, 6) The content of this paper focuses primarily on the extension to vision and naturally extends to audio and arbitrary-to-arbitrary modality interactions. In addition to modality extension, §7 introduces how to extend the input-output space to embodied scenarios, further demonstrating the extensibility of LMMs from the perspective discussed in this paper. In §8, we summarize the discussion on the questions raised above, providing readers with key take-home messages and an outlook on future research.

In summary, our contributions are threefold:

- Going beyond specific scenarios and model framework, we review the current LMMs from a general perspective of input-output space extension. We hope that such a broad and comprehensive survey can provide an intuitive overview to related researchers and inspire future work.
- Based on the structure of input-output spaces, we systematically review the existing models, including mainstream models based on discrete-continuous hybrid spaces and models with unified multi-modal discrete representations. Additionally, we introduce how to align the constructed multi-modal representations and conduct evaluations according to the extended input and output.
- We elaborate on how to extend LMMs to embodied scenarios to highlight the extensibility of LMMs from the input-output extension perspective. To our knowledge, this is the first article to summarize embodied LMMs.

 doi:10.20944/preprints202411.0685.v1

2. Preliminary

Before introducing LMMs, we briefly outline the evolution of multi-modal research paradigms in this section to highlight the key difference and advancements of LMMs, compared to earlier multi-modal models. As presented in Figure 2, we focus on the vision-language domain, diving the development of related research into three stages.

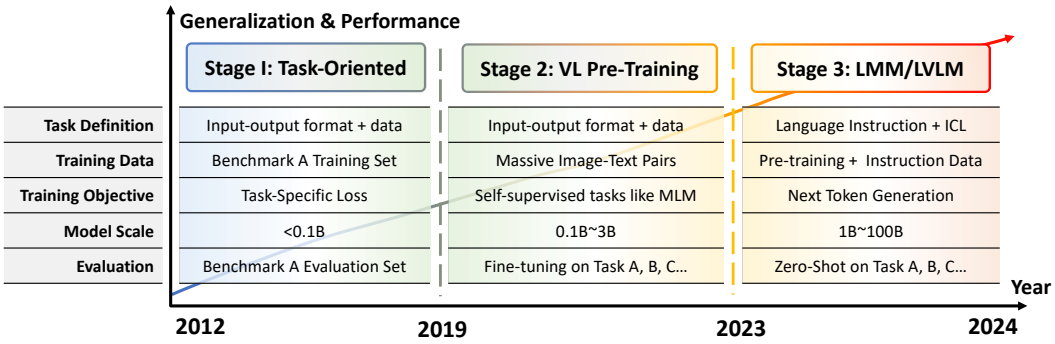

Figure 2. Illustration of the evolution of multi-modal research, focusing on the core characteristics and distinctions at different stages.

Task-Oriented Paradigm

Early multi-modal research focuses on specific scenarios and tasks. The core research problems are how to define multi-modal tasks, construct benchmarks, and design models to address these tasks. The most commonly explored tasks include VQA [23,24], image-text retrieval [25,26], image captioning [27,28], visual grounding [29], visual reasoning [30–33], and so on. Models with various architectures have been proposed for specific tasks [34–40]. The key characteristic of task-oriented methods is to define tasks through a large number of samples (training sets), limiting the generalizability and requiring high costs for transfer across tasks.

Vision-Language Pre-training (VLP)

Since task-oriented methods may introduce task-specific inductive bias and lead to overfitting, researchers explore ways to construct unified architectures and learn generalized multi-modal representations. Inspired by the pre-training techniques introduced by BERT [41], vision-language pre-trained (VLP) models are built on multi-layer Transformer [42] and trained with self-supervised tasks on large amounts of image-text pairs [43–46]. Pre-trained models provide effective initial checkpoint for fine-tuning on various downstream tasks [47–51]. VLP methods make an important step towards generalization, but they still require specific parameters and fine-tuning samples to define tasks, failing to provide a unified interface for users.

Large Multi-Modal Models (LMMs)

The success of LLMs has revealed the potential of using language-based instructions as a generalized and interactive interface [52]. Inspired by this, LMMs also leverage language as the interface between users and machines. By integrating and aligning other modalities into the input-output space, LMMs can understand multi-modal context and respond subsequent instructions from users, even in zero-shot scenarios [6,7,53]. Such generalizability and interactivity make LMMs highly applicable and versatile as multi-modal foundation models.

3. Input-Output Space Extension

In this section, we introduce prevalent solutions to construct multi-modal input-output space. As illustrated in Figure 3, existing methods can be categorized based on different input-output space structures, and the extension to other modalities can be summarized in a similar manner.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2024 doi:10.20944/preprints202411.0685.v1

3.1. Encode Multi-Modal Input Representation

Regarding the input, the core research problems involve how to code the representations of each modality and how to integrate them into a multi-modal input space (illustrated in the lower part in Figure 3).

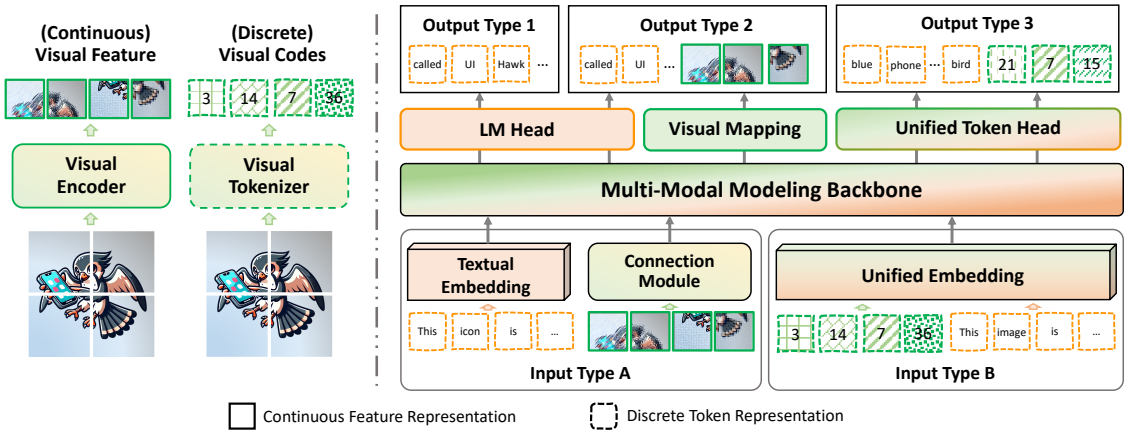

Figure 3. Summary and illustration of different input-output space structures for extension to vision modality.

3.1.1. Textual Representation

As a discrete signal, text is a sequence composed of characters. Following the practice of LLMs, LMMs typically utilize tokenizers, such as BPE [54,55], WordPiece [56], and Unigram [57], to merge characters into sub-word tokens. Ultimately, texts are represented as sequences of discrete tokens.

3.1.2. Visual Representation

For visual signals with spatial-temporal information, LMMs mainly employ pre-trained visual encoders for representing images (videos) into continuous features or discrete codes. Figure 4 illustrates the evolution of existing visual encoders.

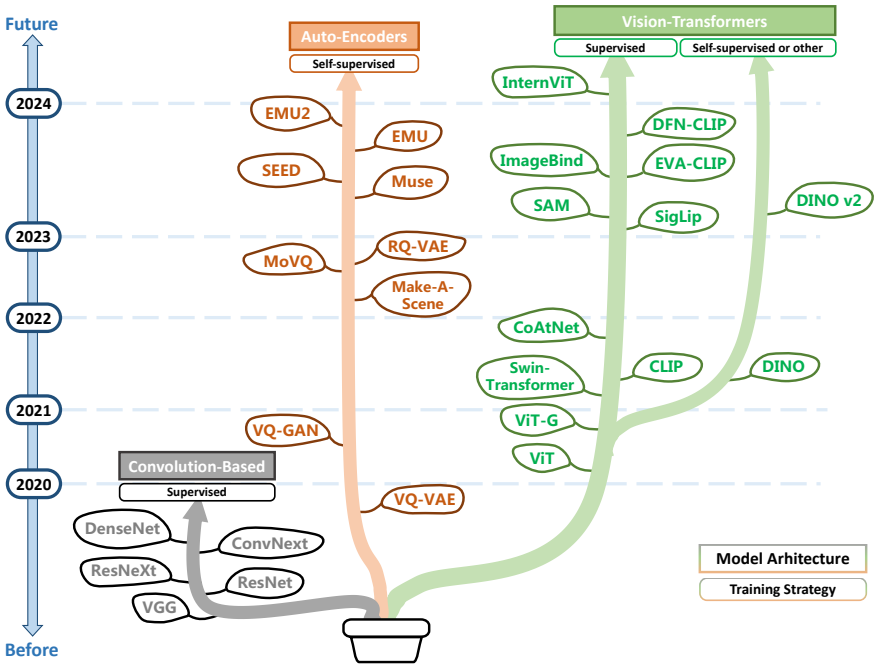

Figure 4. The evolution of commonly adopted visual encoder architectures and training strategies.

 doi:10.20944/preprints202411.0685.v1

Encoder Architecture

Commonly adopted architectures can be divided into two categories: convolution-based [58,59] and vision-Transformer-based models [60,61]. Both methods encode images into continuous 2D feature maps. These features can be further converted into discrete visual codes through vector quantization (VQ) by learning a fixed-size visual codebook [62,63]. In addition, models like Fuyu [64] do not rely on visual encoders and directly use pixel values of image patches as the visual representations.

Encoder Training

Employed visual encoders are mainly pre-trained in supervised or self-supervised manner. Early exploration utilize image categories as supervision signals [65], while CLIP-like models [66–68] use language supervision to learn generalized representations. Additionally, SAM [69] leverages segmentation tasks as training objectives. In contrast, self-supervised learning only requires images for training. Contrastive self-supervised methods train models to distinguish representations between different images [70–73]; another line of approaches construct auto-encoders, where models are demanded to reconstruct images from the encoded visual representations, which is often used to support downstream image generation [62,63,74,75].

Visual Representation Enhancement

Since most visual encoders are limited to fixed resolutions and capture specific aspects of visual features, existing LMMs proposes to enhance the input visual representations on two aspects: resolution enhancement and feature enhancement.

To support high-resolution input images, a line of methods directlty extends the visual encoder, including interpolating position embeddings in vision Transformers [7,76] and using CNN-based models to enhance the encoding efficiency of high-resolution images while compressing the size of encoded feature maps [77,78]. Another line of approaches propose to crop high-resolution images into multiple sub-images and input them into the low-resolution encoder along with the down-sampled full image [79–83]. Additionally, different sub-image partitioning templates help address issues caused by varying aspect ratios of images.

Regarding feature enhancement, common practices consider ensembing visual representations encoded by different encoders, such as combining encoders trained with different strategies [84,85], or integrating high-resolution and low-resolution encoders [86,87]. Specialized modules have been introduced to better fuse features from different encoders [87–89]

Multi-Image Input

Based on the prevalent sequence modeling framework of current LMMs, multiple images can be intuitively arranged in the input sequence [90–93]. For videos, where images (frames) are temporally related, spatial-temporal encoders such as TimeSformer [94] and VideoSwin [95] can be further used for encoding [96,97].

3.1.3. Constructing Multi-Modal Input Space

As illustrated in the lower part Figure 3, there exist two mainstream types of multi-modal input space.

Type A: Hybrid Input Space

Text are represented in a discrete form, while visual signals are encoded in continuous representations, preseving the complete visual information. However, due to the gap in the input space, connection modules are required to perform input-level cross-modal alignment, which is discussed in Section 4.

 doi:10.20944/preprints202411.0685.v1

Type B: Unified Discrete Input Space

Different from Type A, further quantizing visual representations into discrete visual codes facilitates the construction of a unified input space. A multi-model vocabulary can be intuitively integrated and directly used to support subsequent modeling.

3.1.4. Extension to More Modalities

Beyond the vision modality, signals from other modalities can be encoded and introduced into the input space following a similar paradigm. For example, various encoders can help encode audio into continuous [98–100] or discrete [101] representations. As a step further, an arbitrary-modality input space can be represented in either hybrid [11,102–104] or unified discrete forms [12].

3.2. Decode Multi-Modal Output Representation

Based on the input, backbones of LMMs present continuous multi-modal output representations which can be used to decode the output signals of different modalities. For example, with the commonly used causal modeling framework, the output representation can be leveraged to predict the signal at the next position in the sequence. Predicted token sequence can be converted to text with the tokenizer while different image generator can be adopted to decode images from output in different forms. In this section, we discuss the commonly adopted paradigms to partition the output space of different modalities and perform corresponding decoding, as shown in the upper part of Figure 3.

3.2.1. Type 1: Text-Only Output Space

If only text output is required, similar to LLMs, discrete tokens can be generated from the ouput representations through a classification-based language modeling (LM) head and specific decoding strategies [5,6].

Please note that models that first generate text descriptions and then use external tools like Stable Diffusion and CLIP to generate or retrieve content in other modalities, such as Visual ChatGPT [105], InternLM-XComposer series [106,107], and Mini-Gemini [87], are also classified as text-only output models because they are not in an end-to-end manner.

3.2.2. Type 2: Hybrid Multi-Modal Output Space

To support image generation, a series of methods first introduce special tokens, such as the start and end tokens for images, or a series of consecutive placeholder tokens to indicate where images should be generated. The continuous output representations at the corresponding positions are then connected to visual decoders (mainly Diffusion models [108]) through visual mapping modules [9,109–111]. Similar to the hybrid input space, visual mapping modules perform output-level alignment and requires further training which is described in Section 4.

3.2.3. Type 3: Unified Discrete Multi-Modal Output Space

Based on the joint multi-modal vocabulary constructed within Type B input space described in Section 3.1.3, image generation can be naturally integrated into the token decoding process. Image and text tokens in the vocabulary inherently divide the output space, and the predicted visual codes can be fed to the corresponding codebook detokenizer to generate the image [22,112].

3.2.4. Extension to More Modalities

The Type 2 and Type 3 output spaces can be further expanded to incorporate more modalities to support arbitrary-modality output. Next-GPT [11] and Codi-2 [103] further extends the hybrid output space, while AnyGPT [12] and UnifiedIO-2 [104] construct unified discrete spaces for all modalities.

 doi:10.20944/preprints202411.0685.v1

3.3. Prevalent Input-Output Paradigms

Considering the input-output space structures introduced above, most existing LMMs can be categorized to three types: (1) **Multi-modal understanding models** that rely on Type A input and Type 1 output, these models are mainly designed for understanding tasks that can be fully expressed in language [7,84,113,114]; (2) **Multi-modal generation models** which comprise of Type A input and Type 2 output, such models excel in generating multi-modal interleaved responses based on the context [9,11,75]; (3) **Unified multi-modal models** that represent and generate multiple modalities in a unified discrete form [12,22,112].

Tables 1 and Table 2 list the design paradigms of currently popular LMMs, grouped according to the aforementioned classification criteria. The model alignment architectures that will be discussed in Section 4 are also included.

 doi:10.20944/preprints202411.0685.v1

Model	Input Space		Output Space		Architecture				Max Res.	Date
	Modality	Type	Modality	Type	Backbone	Modality Encoder	Connection	Internal Module		
Flamingo [115]	Text, Vision	A	Text	1	Chinchilla	NFNet	Perceiver	Cross-Attention	480	2022/04
BLIP-2 [5]	Text, Vision	A	Text	1	Flan-T5 / OPT	CLIP ViT-L/14 / Eva-CLIP ViT-G/14	Q-Former	-	224	2023/01
LLaMA-adapter [116]	Text, Vision	A	Text	1	LLaMA	CLIP-ViT-L/14	MLP	Adaption Prompt	224	2023/03
MiniGPT-4 [117]	Text, Vision	A	Text	1	Vicuna	Eva-CLIP ViT-G/14	Q-Former	-	224	2023/04
LLaVA [6]	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Linear	-	224	2023/04
mPLUG-Owl [118]	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Abstractor	-	224	2023/04
LLaMA-adapter v2 [119]	Text, Vision	A	Text	1	LLaMA	CLIP-ViT-L/14	MLP	Adaption Prompt	224	2023/04
InstructBLIP [113]	Text, Vision	A	Text	1	Flan-T5 / Vicuna	Eva-CLIP ViT-G/14	Q-Former	-	224	2023/05
Otter [92]	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Perceiver	Cross-Attention	224	2023/05
LAVIN [120]	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	MLP	MM-Adapter	224	2023/05
MultimodalGPT [121]	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Perceiver	Cross-Attention	224	2023/05
Shikra [122]	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Linear	-	224	2023/06
VideoChatGPT [123]	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Linear	-	224	2023/06
Valley [90]	Text, Vision	A	Text	1	Stable-Vicuna	CLIP ViT-L/14	Temporal Module + Linear	-	224	2023/06
Lynx [124]	Text, Vision	A	Text	1	Vicuna	EVA-1B	Resampler	Adapter	420	2023/07
Qwen-VL [7]	Text, Vision	A	Text	1	Qwen	OpenCLIP ViT-bigG	Cross-Attention	-	448	2023/08
BLIVA [125]	Text, Vision	A	Text	1	Flan-T5 / Vicuna	Eva-CLIP ViT-G/14	Q-Former + MLP	-	224	2023/08
IDEFICS [126]	Text, Vision	A	Text	1	LLaMA	OpenCLIP ViT-H/14	Perceiver	Cross-Attention	224	2023/08
OpenFlamingo [127]	Text, Vision	A	Text	1	LLaMA, MPT	CLIP ViT-L/14	Perceiver	Cross-Attention	224	2023/08
InterLM-XC [106]	Text, Vision	A	Text	1	InternLM	Eva-CLIP ViT-G/14	Perceiver	-	224	2023/09
LLaVA-1.5 [128]	Text, Vision	A	Text	1	Vicuna 1.5	CLIP ViT-L/14	MLP	-	336	2023/10
MiniGPT-v2 [129]	Text, Vision	A	Text	1	LLaMA-2	EVA	Linear	-	448	2023/10
Fuyu-8B [64]	Text, Vision	A	Text	1	Persimmon	-	Linear	-	unlimited	2023/10
UReader [79]	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Abstractor	-	224*20	2023/10
CogVLM [130]	Text, Vision	A	Text	1	Vicuna 1.5	EVA2-CLIP-E	MLP	Visual Expert	490	2023/11
Monkey [80]	Text, Vision	A	Text	1	Qwen	OpenCLIP ViT-bigG	Cross-Attention	-	896	2023/11
ShareGPT4V [131]	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14	MLP	-	336	2023/11
mPLUG-Owl2 [132]	Text, Vision	A	Text	1	LLaMA-2	CLIP ViT-L/14	Abstractor	Modality-Adaptive Module	448	2023/11
Sphinx [133]	Text, Vision	A	Text	1	LLaMA-2	CLIP ViT-L/14 + CLIP ConvNeXt-XXL + DINOv2 ViT-G/14	Linear + Q-Former	-	672	2023/11
InternVL [114]	Text, Vision	A	Text	1	Vicuna	InternViT	QLLaMA / MLP	-	336	2023/12
MobileVLM [134]	Text, Vision	A	Text	1	MobileLLaMA	CLIP ViT-L/14	LDP (conv-based)	-	336	2023/12
VILA [135]	Text, Vision	A	Text	1	LLaMA-2	CLIP ViT-L	Linear	-	336	2023/12
Osprey [77]	Text, Vision	A	Text	1	Vicuna	CLIP ConvNeXt-L	MLP	-	512	2023/12
Honeybee [136]	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14	C-Abstractor / D-Abstractor	-	336	2023/12
Omni-SMoLA [137]	Text, Vision	A	Text	1	UL2	Siglip ViT-G/14	Linear	LoRA MoE	1064	2023/12
LLaVA-Next [83]	Text, Vision	A	Text	1	Vicuna / Mistral / Hermes-2-Yi	CLIP ViT-L/14	MLP	-	672	2024/01
InterLM-XC2 [107]	Text, Vision	A	Text	1	InternLM-2	CLIP ViT-L/14	MLP	Partial LoRA	490	2024/01
Mousi [89]	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14 + MAE + LayoutLMv3 + ConvNeXt + SAM + DINOv2 ViT-G	Poly-Expert Fusion	-	1024	2024/01
LLaVA-MoLE [138]	Text, Vision	A	Text	1	Vicuna1.5	CLIP ViT-L/14	MLP	LoRA MoE	336	2024/01
MoE-LLaVA [139]	Text, Vision	A	Text	1	StableL / Qwen / Phi-2	CLIP ViT-L/14	MLP	FFN MoE	336	2024/01
MobileVLM v2 [140]	Text, Vision	A	Text	1	MobileLLaMA	CLIP ViT-L/14	LDP v2		336	2024/02
Bunny [141]	Text, Vision	A	Text	1	Phi-1.5 / LLaMA-3 StableLM-2 / Phi-2	SigLIP, EVA-CLIP	MLP	-	1152	2024/02
TinyLLaVA [142]	Text, Vision	A	Text	1	TinyLLaMA / Phi-2 / StableLM-2	SigLIP-L, CLIP ViT-L	MLP	-	336/384	2024/02
Sphinx-X [81]	Text, Vision	A	Text	1	TinyLLaMA / InternLM2 / LLaMA2 / Mixtral	CLIP ConvNeXt-XXL + DINOv2 ViT-G/14	Linear	-	672	2024/02
Mini-Gemini [87]	Text, Vision	A	Text	1	Gemma / Vicuna / Mixtral / Hermes-2-Yi	CLIP ViT-L + ConvNext-L	Cross-Attention + MLP	-	1536	2024/03
Deepseek-VL [84]	Text, Vision	A	Text	1	Deepseek LLM	SigLIP-L, SAM-B	MLP	-	1024	2024/03
LLaVA-UHD [82]	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Perceiver	-	336*6	2024/03
Yi-VL [143]	Text, Vision	A	Text	1	Yi	CLIP ViT-H/14	MLP	-	448	2024/03
MM1 [144]	Text, Vision	A	Text	1	in-house LLM	CLIP ViT-H*	C-Abstractor	-	1792	2024/03
VL Mamba [145]	Text, Vision	A	Text	1	Mamba LLM	CLIP-ViT-L / SigLIP-SO400M	VSS + MLP	-	384	2024/03
Cobra [146]	Text, Vision	A	Text	1	Mamba-Zephyr	DINOv2 + SigLIP	MLP	-	384	2024/03
InternVL 1.5 [147]	Text, Vision	A	Text	1	InternLM2	InternViT-6B	MLP	-	448*40	2024/04
Phi-3-Vision [148]	Text, Vision	A	Text	1	Phi-3	CLIP ViT-L/14	MLP	-	336*16	2024/04
PLLaVA [149]	Text, Vision	A	Text	1	Vicuna / Mistral / Hermes-2-Yi	CLIP ViT-L/14	MLP + Adaptive Pooling	-	336	2024/04
TextHawk [150]	Text, Vision	A	Text	1	InternLM-1	SigLIP-SO400M/14	Resampler + MLP	-	unlimited	2024/04
Imp [151]	Text, Vision	A	Text	1	Phi-2	SigLIP	MLP	-	384	2024/05
IDEFICS2 [152]	Text, Vision	A	Text	1	Mistral-v0.1	SigLIP-SO400M/14	Perceiver + MLP	-	384*4	2024/05
ConvLLaVA [78]	Text, Vision	A	Text	1	Vicuna-	CLIP-ConvNeXt-L*	MLP	-	1536	2024/05
Ovis [153]	Text, Vision	B	Text	1	LLaMA3 / Qwen1.5	CLIP ViT-L + Visual Embedding	-	-	336	2024/05
Deco [154]	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14	MLP + Adaptive Pooling	-	336	2024/05
CuMo [155]	Text, Vision	A	Text	1	Mistral / Mixtral	CLIP ViT-L/14	MLP	FFN + MLP MoE	336	2024/05
Cambrian-1 [88]	Text, Vision	A	Text	1	Vicuna-1.5 / LLaMA-3 / Hermes-2-Yi	CLIP ViT-L/14 + DINOv2 ViT-L/14 + SigLIP ViT-SO400M + OpenCLIP ConvNeXt-XXL	Spatial Vision Aggregator	-	1024	2024/06
GLM-4v [156]	Text, Vision	A	Text	1	GLM4	EVA-CLIP-E	Conv + SwiGLU	-	1120	2024/06
InterLM-XC2.5 [157]	Text, Vision	A	Text	1	InternLM-2	CLIP ViT-L/14	MLP	Partial LoRA	560*24	2024/07
IDEFICS3 [158]	Text, Vision	A	Text	1	LLaMA 3.1	SigLIP-SO400M/14	Perceiver + MLP	-	1820	2024/08
mPLUG-Owl3 [159]	Text, Vision	A	Text	1	Qwen2	SigLIP-SO400M/14	Linear	Hyper Attention	384*6	2024/08
CogVLM2 [156]	Text, Vision	A	Text	1	LLaMA3	EVA-CLIP-E	Conv + SwiGLU	Visual Expert	1344	2024/08
CogVLM2-video [156]	Text, Vision	A	Text	1	LLaMA3	EVA-CLIP-E	Conv + SwiGLU	-	224	2024/08
LLaVA-OV [160]	Text, Vision	A	Text	1	Qwen-2	SigLIP-SO400M/14	MLP	-	384*36	2024/09
Qwen2-VL [161]	Text, Vision	A	Text	1	Qwen-2	ViT-675M	MLP	-	unlimited	2024/09

Table 1. Summary of various frameworks of LVLMs that focus on understanding tasks with only text output (Output Type 1). If there are multiple components in a column, '+' represents a combination while '/' indicates an either-or choice. Max Res. represents the maximum resolution, the "X*Y" pattern indicates methods based on sub-image tiling, X is the base resolution while Y is the maximum number of tiles.

 doi:10.20944/preprints202411.0685.v1

Model	Input Space		Output Space		Architecture						Date
	Modality	Type	Modality	Type	Backbone	Modality Encoder	Connection	Internal Module	Mapping	Modality Decoder	
Any-Modality LMMs											
PandaGPT [162]	T, V, A...	A	T	1	Vicuna	ImageBind	Linear	-	-	-	2023/05
ImageBind -LLM [102]	T, V, A, 3D	A	T	1	Chinese -LLaMA	ImageBind + Point-Bind	Bind Network	Adaption Prompt			2023/09
Next-GPT [11]	T, V, A	A	T, V, A	2	Vicuna	ImageBind	Linear	-	Transformer	SD + AudioLDM + Zeriscope	2023/09
Codi-2 [103]	T, V, A	A	T, V, A	2	LLaMA-2	ImageBind	MLP	-	MLP	SD + AudioLDM2 + zeroscope v2	2023/11
UnifiedIO2 [104]	T, V, A	A	T, V, A	3	UnifiedIO2	OpenCLIP ViT-B + AST	Linear + Perceiver	-	-	VQ-GAN + ViT-VQGAN	2023/12
AnyGPT [12]	T, V, A	B	T, V, A	3	LLaMA-2	SEED + Encodec + SpeechTokenizer	-	-	-	SEED + Encodec + SpeechTokenizer	2024/02
Uni-MoE [163]	T, V, A	A	T	1	LLaMA	CLIP ViT-L/14 + Whisper-small + BEATs	MLP + Q-former	Modality Aware FFN MoE	-	-	2024/05
Large Audio-Language Models											
SpeechGPT [164]	T, A	B	T, A	3	LLaMA	HuBERT	-	-	-	Unit Vocoder	2023/05
Speech-LLaMA [165]	T, A	A	T	1	LLaMA	CTC compressor	Transformer	-	-	-	2023/07
SALMONN [166]	T, A	A	T	1	Vicuna	Whisper-Large-v2 + BEATs	Window-level Q-Former	-	-	-	2023/10
Qwen-Audio [167]	T, A	A	T	1	Qwen	Whisper-Large-v2	-	-	-	-	2023/11
SpeechGPT -Gen [10]	T, A	B	T, A	3	LLaMA-2	SpeechTokenizer	-	-	Flow Matching	SpeechTokenizer	2024/01
SLAM-ASR [8]	T, A	A	T	1	LLaMA-2	HuBERT	MLP + DownSample	-	-	-	2024/02
WavLLM [168]	T, A	A	T	1	LLaMA-2	Whisper-Large-v2 + WavLM-Base	Adapter + Linear	-	-	-	2024/04
SpeechVerse [169]	T, A	A	T	1	Flan-T5-XL	WavLM-Large / Best-RQ	Convolution	-	-	-	2024/05
Qwen2-Audio [170]	T, A	A	T	1	Qwen	Whisper-Large-v3	-	-	-	-	2024/07
LLaMA-Omni [171]	T, A	A	T, A	2	LLaMA-3.1	Whisper-Large-v3	MLP + DownSample	-	Transformer	Unit Vocoder	2024/09
Large Vision-Language Models for Multi-Modal Generation											
GILL [9]	T, V	A	T, V	2	OPT	CLIP ViT-L	Linear	-	Transformer	SD	2023/05
Emu [111]	T, V	A	T, V	2	LLaMA	EVA-02-CLIP-1B	Transformer	-	Linear	SD	2023/07
LaVIT [172]	T, V	A	T, V	3	LLaMA	Eva-CLIP ViT-G/14 + LaVIT Tokenizer	Linear	-	-	LaVIT De-Tokenizer	2023/09
CM3Leon [173]	T, V	B	T, V	3	CM3Leon	Make-A-Scene	-	-	-	Make-A-Scene	2023/09
DreamLLM [109]	T, V	A	T, V	2	Vicuna	CLIP ViT-L/14	Linear	-	Linear	SD	2023/09
Kosmos-G [174]	T, V	A	T, V	2	MAGNETO	CLIP ViT-L/14	Resampler	-	AlignerNet	SD	2023/10
SEED-LLaMA [112]	T, V	B	T, V	3	Vicuna / LLaMA-2	SEED Tokenizer	-	-	-	SEED De-Tokenizer	2023/10
MiniGPT-5 [110]	T, V	A	T, V	2	Vicuna	Eva-CLIP ViT-G/14	Q-Former	-	Transformer	SD	2023/10
Emu-2 [75]	T, V	A	T, V	2	LLaMA	EVA-02-CLIP-E-plus	Linear	-	Linear	SDXL	2023/12
Chameleon [22]	T, V	B	T, V	3	Chameleon	Make-A-Scene	-	-	-	Make-A-Scene	2024/05
MoMA [175]	T, V	B	T, V	3	Chamelon	Make-A-Scene	-	Modality Aware FFN MoE	-	Make-A-Scene	2024/07
Vila-U [176]	T, V	B	T, V	3	LLaMA-2	SigLIP + RQ-VAE	-	-	-	RQ-VAE	2024/09

Table 2. Supplement to Table 1. In the "Modality" column, T, V, and A are abbreviations for Text, Vision, and Audio, respectively.

4. Multi-Modal Alignment

Based on the multi-modal input-output spaces introduced in Section 3, the design of LMMs further needs to consider how to align representations across different modalities. The core research problems behind include: (1) How to design the corresponding model architecture to uniformly model multi-modal representations (Section 4.1); (2) How to train the model parameters to learn alignment and interaction across modalities (Section 4.2). Ultimately, through multi-modal alignment, LMMs can simultaneously comprehend multi-modal contexts and generate multi-modal responses.

4.1. Alignment Architecture

The current mainstream LMM architectures follow a similar paradigm: aligning inputs from all modalities to a **unified multi-modal backbone** for modeling, interaction, and generating multi-modal responses. To facilitate the unified modeling, additional modules are designed for (1) **input-level alignment** to unify the multi-modal inputs into a consistent form and space; (2) **internal alignment** of the backbone for complex cross-modal interactions; and (3) **output-level alignment** to map the outputs of the backbone to different modality decoders.

4.1.1. Multi-Modal Modeling Backbone

Typically, the backbone is based on a decoder-only architecture composed of multiple Transformer blocks [42]. To better understand language, the backbone is primarily initialized with a pre-trained LLM, such as LLaMA [2,177,178], Vicuna [179], Mistral [180], Qwen [3,181], and so on [143,182].

 doi:10.20944/preprints202411.0685.v1

In addition, LMMs for edge devices are usually initialized with smaller language models, such as MobileLLaMA [183], Phi [148], etc. [184,185]. The backbone can also inherit MoE-based language models like Mixtral 8x7B [186].

Apart from the commonly used architecture mentioned above, some LMMs adopt encoder-decoder backbones [104,187–190]. Additionally, native LMMs like Chameleon [22] are not initialized with pre-trained LLMs and trained from scratch.

4.1.2. Input-level Alignment

As introduced in Section 3.1.3, there may exist gaps between modalities in the extended multi-modal input space. To enable the backbone to process multi-modal information uniformly, it is necessary to align the form and space of inputs across modalities at the input level.

Specifically, for Type B input space, since all modalities are represented in a unified discrete token form, input-level alignment can be achieved by directly merging the vocabularies of multiple modalities and learning the token representations through subsequent alignment training [12,22,112].

Regarding Type A hybrid input space, it is required to introduce a connection module to convert inputs from other modalities into a sequential representation that matches the dimension of textual token embeddings. Commonly adopted connection modules are summarized below.

MLP Based

A typical connection module is implemented through one or more linear projection layers, connected by activation functions such as GeLU [191], resulting in a multi-layer perceptron (MLP). This approach directly aligns the dimension of representations from other modalities with text [6,128]. By further flattening the 2D or 3D features into 1D in a specific order, it allows for alignment with the text sequence [11,83,123]. The advantage of MLP-based modules lies in the simplicity, light weight, offering fast convergence during alignment training. However, MLP-based modules cannot compress redundant information, which may result in excessively long modality representation sequences (e.g., high-resolution images), reducing computational efficiency and requiring additional designs to compress the information [76,147,192].

Attention Based

Another prevalent connection modules are based on attention mechanisms. This method typically introduces a fixed number of learnable vectors as queries, which retrieve relevant information from other-modality representations (serving as keys and values) through cross-attention modules. The output representations of the queries, enriched with information from other modalities, serves as the modality input to the backbone. Representative module architectures include Q-Former [5,113], abstractor [118,132], resampler [80,124], and so on [7,106]. The query-level representations obtained from attention mechanisms effectively compress and aggregate information from other modalities. Additionally, recent works have demonstrated further extensibility, including integrating representations from multiple encoders [87,88,193], incorporating local grounding information [194], and scaling up to an 8B Q-LLaMA [114]. However, these modules mainly involve more parameters and typically require additional training [5,194]. Yao *et al.* have found that attention-based modules may result in the loss of important information.

Others

In additional to the mainstream structures mentioned above, several other connection modules have been proposed. CNN-based modules utilize the inductive bias of convolutional operations to model local information, further combined with pooling layers, the number of resulted tokens can be effectively reduced [136,140,156]. Adaptive pooling-based modules can compress features using spatial relationships without introducing additional parameters [149,154]. Furthermore, VL-Mamba

 doi:10.20944/preprints202411.0685.v1

explores to use vision selective scanning as connection to integrate representations across different modalities [145].

4.1.3. Internal Alignment

With the help of input-level alignment, vanilla Transformer based backbones can uniformly process multi-modal information. Furthermore, researchers have explored introducing additional parameter modules within the backbone to further enhance the modeling of internal interactions and alignment between modalities. In this section, we categorize and summarize commonly adopted methods for internal alignment.

Cross-Attention Layer

Flamingo [115], as a pioneering work in LMM, is the first to propose inserting cross-attention layers between the original layers of the backbone, allowing text to perceive information from the visual context. Additionally, a tanh gating mechanism is introduced to control the degree of modality fusion. Subsequently, the Flamingo architecture has been adopted by recently proposed LMMs [121, 126,127,195], CogAgent [86] further utilizes cross-attention to supplement high-resolution image information. Although effective, densely inserted cross-attention layers bring a large number of parameters. Ye *et al.* improve this by introducing sparsely inserted hyper attention, which significantly reduce extra parameters and facilitate model convergence through parallel self-attention and cross-attention calculation.

Adaption Prompt

LLaMA-Adapter incorporates modality representations into lightweight learnable adaption prompts and feed the prompts as prefix contexts to the backbone [116]. LLaMA-Adapter V2 [119] improves this method with an early knowledge fusion strategy. ImageBind-LLM [102] further extends the adaption prompts to support more modalities.

Visual Expert

To distinguish between visual and textual modeling, some LMMs introduce visual expert modules to process visual tokens specifically. CogVLM [130] adds additional attention and FFN layers to process visual tokens without compromising the original textual modeling capabilities of backbones. mPLUG-Owl2 [132] only introduces modality-specific parameter blocks in the normalization layers and the K and V mapping layers of the attention modules. InternLM-XComposer2 [107], on the other hand, designs a lightweight Partial LoRA module for additional modeling of visual tokens.

Mixture of Experts (MoE)

Unlike previously discussed modules that are densely activated, the idea of MoE is to introduce "experts" modules in the backbone which can be sparsely activated according to different inputs through gating routers [196–198]. A typical solution of introducing MoE into LMMs is based on sparse upcycling [199] to sparsify a dense checkpoint. LLaVA-MoLE [138] considers LoRA as experts and incorporates them into the FFN layers, while MoE-LLaVA [139] directly extends the FFN layers of the base model. CuMo [155] expands the FFN layers of the visual encoder and connection module with co-upcycled MoE layers. All these models utilize Top-K gating routers.

Methods mentioned above introduce MoE through implicit knowledge modeling, explicit modality-specific knowledge can also be incorporated in the design of MoE in LMMs. Uni-MoE [163] extends FFN layers and allocate specific experts for each modality. Modality-specific data are utilized to train corresponding experts for further enhancement. During inference, only relevant modality experts are activated, allowing effective utilization of modality knowledge while maintaining efficiency. Similarly, Chameleon-MoMa [175] duplicates FFN layers in Chameleon and divides experts into modality-specific groups in which routers are independently learned.

 doi:10.20944/preprints202411.0685.v1

4.1.4. Output-level Alignment

Regarding the multi-modal output space described in Section 3.2, both Type 1 and Type 3 are represented in a unified discrete token-based form, multi-modal content can be intuitively generated through a next-token prediction approach with the help of modality-specific de-tokenizer [12,22,104, 112].

For the Type 2 hybrid output space, although modality-related tokens help divide the output space into different modalities, additional mapping modules are required to align the output space of LMM backbones with the input space of corresponding modality generators. Considering image generation, commonly used modules are built on linear projection [109] or the Transformer architecture [9,110]. Similar to Q-Former, Transformer-based modules learn a fixed number of queries to retrieve information from the LMM outputs through cross-attention, serving as the condition input of image diffusion models [108]. Next-GPT [11] further extends the Transformer-based mapping modules to fit more modality diffusion generators. Additionally, Emu series [75,111] replace the linear projection of cross-attention in diffusion models to perform dimensional conversion, achieve output-level alignment.

4.2. Multi-Modal Training

In this section, we discuss how to train the model constructed in Section 4.1 to learn cross-modal alignment and modeling, facilitating the understanding and generation of multi-modal content. We first introduce commonly adopted training data, followed by an elaboration of how to utilize the data to design multi-stage training frameworks for LMMs.

4.2.1. Training Data

We separate existing training data for multi-modal alignment into two categories: pre-training data and instruction tuning data. In the following section, we delve into these data categories, providing a detailed overview of their composition.

Pre-training Data

The pre-training data mainly consists of sequences interspersed with multi-modal information, guiding LMMs to learn the multi-modal associations embedded within and align the representations. Common pre-training data typically exists in three forms, as described below.

X-text Pairs. The most typical format consists of paired X-modal data and the corresponding text, which is image-caption pair for the vision modality. Early captioning data are primarily curated by human annotators, including Flickr30K [26], COCO [203], SBU [241]. Although these datasets are of high quality, their scales are limited and the cost of further extension is affordable. To this end, subsequent works introduce methods for image-text pairs collection by crawling the web, followed by rigorous filtering and post-processing, which leads to the development of significantly larger datasets, such as CC3M [200], CC12M [242], LAION [206], COYO [202], and DataComp [209].

To process the web crawling data, filtering and deduplication are applied to ensure the datasets with high quality and broad coverage. Data filtering aims at removing undesirable content, focusing on text and image data separately. Text filtering includes language filtering, which eliminates documents below a certain language threshold [267–269], and content filtering, which keeps toxic or incomplete sentences out [270,271]. Image filtering discards low-resolution images, those with inappropriate aspect ratios, and images with unavailable URLs [213,259,261]. Deduplication is also essential, as redundant information can harm the model performance [272]. The exact deduplication method removes duplication through string matching [267,273], while URL-based deduplication identify redundant information from the same web pages [269]. In addition, locality sensitive hashing (LSH) methods can be adopted to perform approximate deduplication [274,275], and semantic-level deduplication can be managed by clustering semantic embeddings and retaining representative data [276–278]. Commonly used image deduplication methods involve removing by image URLs or pHash algorithms [259,279].

 doi:10.20944/preprints202411.0685.v1

Although the aforementioned filtering methods are effective, these large-scale datasets are generally weakly labeled, suffering from noise and sub-optimal conditions for model training, with many captions being too simple or failing to accurately describe the images.

In response, recent methods resort to synthetic re-captioning, where original images are re-captioned by advanced models to generate concise textual descriptions, represented by LAION-COCO [207] and LAION-BLIP [48]. Further advancements are then attempted, applying more sophisticated captioning models and unique prompting stratagies, to generate detailed and high-quality image. For example, LaCLIP [343] utilizes LLM to rewrite raw captions, but resulting in severe hallucination, because of limited visual information included in low-quality raw captions. VeCap [344] uses LLaVA [251] to extract all possible visual clues and leverage an LLM to do ethical check and fuse the concepts from both AltText and visual clues to generate the final caption. Subsequently, Capsfusion [345] fine-tunes LLaMA-2 [2] with training data generated by ChatGPT, and the fine-tuned LLaMA-2 [2] organically fuses and harnesses raw and synthetic captions. Monkey [80] utilize a combination of several advanced systems to collect visual descriptions which are provided to ChatGPT. Different from aforementioned automatic generation pipelines, AS-1B [221] introduces a semi-automatic data engine that efficiently leverages various foundation models as annotators, significantly reducing the enormous labeling costs to a manageable level. Additionally, CogVLM2 [156], ImageInWords [218] and Densely Captioned Images (DCI) [219] also generate and refine detailed captions with human in the loop. Recently, it has become a trend to directly utilize GPT-4V's visual perception capabilities to generate high-quality image descriptions [215,257]. Similar approaches are adopted in constructing pre-training data for Ovis [153] and LLaVA-OneVision [160]. Utilizing data generated by GPT-4V, ShareGPT-4V [131] trains captioning engines, while DenseFusion [256] further integrates visual experts as image priors to scale up hyper-detailed image-text data. In contrast to the image-to-text generation, SYNTH^2 [214] leverages a text-to-image model to generate synthetic images.

Similarly, comparable datasets can be created for other modalities. For the video modality, early caption datasets were primarily composed of manual annotations, such as YouCook2 [233], VATEX [248], and Panda-70M [226]. Additionally, web-crawled datasets like HowTo100M [222], VideoCC3M [224], HD-VILA-100M [232], and WebVid-10M [225] are commonly used for video-language alignment, while synthetic captions such as Vript [243] and VIDAL [227] are generated by advanced GPT models. For the audio modality, audio-text pairs like Clotho [228], AudioCaps [229], and AudioSet [231] are widely utilized for pre-training, with synthetic captions provided by WavCaps [230], LAION-Audio-630K [244], and AF-AudioSet [247]. Please refer to Table 3 and 4 for the summarized commonly-adopted X-text pairs.

Name	Modality	Manual Annotation	LLM / LMM Synthesis	I/V/A Source	# I/V/A	# Samples	Date
X-Text Pairs							
CC3M [200]	I			Web	3.3M	3.3M	2018/07
CC12M [201]	I			Web	12.4M	12.4M	2021/02
COYO [202]	I			Web	747M	747M	2022/08
COCO Captions [203]	I	✓		COCO	82.8K	413.9K	2015/04
Flickr-30K [26]	I	✓		Web (Flickr)	31.8K	158.9K	2014/02
WIT [204]	I			Web (Wikipedia)	11.4M	37.1M	2021/03
RedCaps [205]	I			Web (Reddit)	12M	12M	2021/11
LAION-400M [206]	I			Common Crawl	413M	413M	2021/11
LAION-2B [207]	I			Common Crawl	2.3B	2.3B	2022/03
LAION-COCO [207]	I		✓(Open Models)	LAION-2B	600M	600M	2022/09
SBU [208]	I			Web (Flickr)	1M	1M	2011/12
DataComp [209]	I			Common Crawl	1.4B	1.4B	2023/04
TextCaps [28]	I	✓		Open Images v3	22.0K	109.8K	2020/03
Capsfusion [210]	I		✓(Open Models)	LAION-COCO	120M	120M	2023/10
Taisu [211]	I		✓(Open Models)	Web	166M	219M	2022/09
VeCap-300M [212]	I		✓(Open Models)	Web	300M	300M	2023/10
Wukong [213]	I			Web	101M	101M	2022/02
GenPair [214]	I		✓(Gemini Pro)	MUSE Model	-	1M	2024/03
PixelProse [215]	I		✓(Gemini Pro Vision)	CommonPool, CC12M, RedCaps	16.4M	16.4M	2024/06
YFCC100M [216]	I			Web (Flickr)	14.8M	14.8M	2015/03
DOCCI [217]	I	✓		From Human Taken	9.6K	9.6K	2024/04
ImageInWords [218]	I	✓	✓(Open Models)	Web	-	8.6K	2024/05
DCI [219]	I	✓		SA-1B	7.8K	7.8K	2023/12
MetaCLIP [220]	I			Common Crawl	400M	400M	2023/09
AS-1B [221]	I	✓	✓(Open Models)	SA-1B	11M	1.2B	2023/08
Monkey [80]	I		✓(Open Models)	CC3M	-	213K	2023/11
HowTo100M [222]	V			Web (Youtube)	1.22M	136M	2019/07
Ego4D [223]	V	✓		From Human Taken	-	-	2021/10
VideoCC3M [224]	V, A			Web	6.3M	10.3M	2022/04
WebVid-10M [225]	V			Web	10M	10M	2021/04
Panda-70M [226]	I, V	✓		From Human Taken	-	-	2024/02
VIDAL [227]	V, A		✓(ChatGPT)	Web (Youtube, Freesound)		10M	2023/10
Clotho [228]	A	✓		Web (Freesound)	2.9K	14.4K	2019/10
AudioCaps [229]	A	✓		Web (Youtube)	38.1	38.1K	2019/06
WavCaps [230]	A		✓(ChatGPT)	FreeSound, BBC Sound Effects, SoundBible, AudioSet	403K	330.6K	2023/03
AudioSet [231]	A			Web (Youtube Videos)	1.8M	1.8M	2017/06
HD-VILA-100M [232]	V			Web (Youtube)	103M	103M	2021/11
YouCook2 [233]	V	✓		Web (Youtube)	10.3K	10.3K	2017/03
Charades [234]	V	✓		From Human Taken	8.0K	22.6K	2016/04
VidOR [235]	V	✓		YFCC-100M	7K	-	2019/06
Sth-Sth V2 [236]	V	✓		Web	168.9K	-	2017/06
Video Storytelling [237]	V	✓		Web (Youtube)	0.1K	0.4K	2018/07
HD-VG-130M [238]	V		✓(Open Models)	Web (Youtube)	130M	130M	2023/05
DocStruct4M [239]	I			Multiple Datasets	-	4.0M	2024/03
MP-DocStruct1M [240]	I			PixParse	-	1.1M	2024/09

Table 3. Summary of commonly used datasets for pre-training LMMs. In the “Modality" column, I, V, and A represent Image, Video, and Audio, respectively. “Manual Annotation" indicates whether the dataset is annotated by humans, and "LLM / LMM Synthesis" indicates whether the annotations in the dataset are synthesized by LLMs or LMMs.

 doi:10.20944/preprints202411.0685.v1

Name	Modality	Manual Annotation	LLM / LMM Synthesis	I/V/A Source	# I/V/A	# samples	Date
X-Text Pairs							
Vript [243]	V		✓(GPT-4V)	HD-VILA-100M, Web (Youtube, Tiktok)	420K	420K	2024/06
LAION-Audio-630K [244]	A		✓(Open Models)	Web	634K	634K	2022/11
VAST-27M [245]	V		✓(Open Models)	HD-VILA-100M	27M	297M	2023/05
VALOR-1M [246]	V, A	✓		AudioSet	1.2M	1.2M	2023/04
AF-AudioSet [247]	A		✓(Open Models)	AudioSet	331.4K	696.1K	2024/06
YT-Temporal-180M	V			Web (Youtube)	6M	180M	2021/06
VATEX [248]	V	✓		Kinetics-600	26.0K	519.8K	2019/04
DiDeMo [249]	V	✓		YFCC100M	21.5K	32.5K	2017/08
VILA^2 [250]	I		✓(Open Models)	MMC4, COYO	-	-	2024/07
LCS-558K [251]	I		✓(Open Models)	LAION, CC3M, SBU	-	558K	2023/05
LLaVA-CC3M [251]	I		✓(Open Models)	CC3M	-	595K	2023/05
VisText [252]	I	✓	✓(Open Models)	Web (Statista)	9.9K	9.9K	2023/06
Screen2words [253]	I	✓		Rico-SCA	15.7K	78.7K	2021/08
Librispeech [254]	A			Web (LibriVox API)	-	-	2015/04
ArxivCaps [255]	I			Web (ArXiv Papers)	6.4M	3.9M	2024/03
DenseFusion-1M [256]	I		✓(GPT-4V)	LAION-2B	1.1M	1.1M	2024/07
ShareGPT-4V [131]	I		✓(GPT-4V)	Multiple Datasets	1.2M	1.2M	2023/10
ShareGPT4Video [257]	V		✓(GPT-4V, GPT-4)	Web, Panda-70M, Ego4D, BDD100K	-	101K	2024/06
ShareGPT-4o [258]	I, V, A		✓(GPT-4o)	Multiple Datasets	-	220K	2024/05
Multi-Modal Interleaved Documents							
MMC4 [259]	I			Common Crawl	571M	101M	2023/04
OBELICS [260]	I			Common Crawl	353M	141M	2023/06
MINT-1T [261]	I			Common Crawl, PDFs, ArXiv	3.42B	1.1B	2024/06
OmniCorpus [262]	I			Common Crawl, Web, Youtube	8.6B	2.2B	2024/06
Kosmos-1 [263]	I			Common Crawl	-	71M	2023/02
InternVid-ICL [264]	V		✓(Open Models)	Youtube	7.1M	7.1M	2023/07
Howto-Interlink7M [265]	V		✓(Open Models)	HowTo100M	1M	1M	2024/01
YT-Storyboard-1B [266]	V			YT-Temporal-1B	18M	-	2023/07

Table 4. Supplement to Table 3.

Multi-modal Interleaved Documents. Although X-text pairs have been demonstrated to be effective in pre-training for cross-modal alignment, these data are usually short in length and relatively simple in form. Additionally, Single X-text pairs cannot enable LMMs to learn in-context learning capabilities in multi-modal contexts [115]. To address this, researchers propose to constructed multi-modal documents, in which multiple information units of different modalities (images, sentences, speech, etc.) are distributed in an interleaved manner. Regarding the vision modality, MMC4 [259] is the first large-scale publicly available multi-modal interleaved dataset, which is an extension of the text-only C4 dataset by gathering images from WAT files and associating each image with a sentence. OBELICS [260] is constructed from HTML files obtained from Common Crawl dumps, and the resulting documents maintain the original linearity of images and texts as they appeared on the websites, while removing spam and ads. Exposing the model to a much wider distribution of texts, MINT-1T [261] curates more diverse sources of interleaved documents from HTML documents, PDFs and ArXiv papers, with a 10x scale-up from existing open-source datasets.

As for other modalities, InternVid-ICL [264] is a large-scale interleaved video-text dataset, established by arranging clips and their descriptions in sequences according to the chronological order, connecting the interlaced multi-modal items to create video-centric dialogues. Additionally, Howto-Interlink7M [265] is a high-quality interleaved video-text dataset derived from HowTo100M [222], and YT-Storyboard-1B [266] is curated for training Emu [111] and Emu2 [75]. Compared to X-text pairs, interleaved data is relatively limited scarce. Table 4 presents relevant multi-modal interleaved datasets.

Scenario-oriented Data. While X-text data and interleaved documents effectively aid cross-modal alignment, they are limited to semantic-level information about objects and scenes, and cannot offer LMMs extensive knowledge to handle demands of diverse scenarios. Therefore, according to specific scenarios, existing methods typically aggregate relevant data to help LMMs learn particular capabilities [156,158,374]. This type of data mainly pertains to the vision-language contexts and can be obtained through methods such as manual curating [23,292], re-formulating existing data [147,157],

 doi:10.20944/preprints202411.0685.v1

or automatic synthetic data generation [255,316]. In Table 5, we summarize several scenarios of interest and related datasets. General VQA [24,282,282] focuses on visual understanding of real-world images, including identifying people and objects, scene comprehension, counting, and color recognition, often necessitating complex reasoning about visual facts. General OCR data [300,303,304] is leveraged to enhance text-rich image understanding. Document/Chart/Screen data [304,307,312] empowers LMMs to interpret complex text and structural information, enhancing their understanding of documents, tables, and screen contents. Math/Science/Code [317,320,327] involves advanced mathematical reasoning and geometry tasks, as well as code visualization tasks. Detection and Grounding data [29,330] conveys spatial information of objects through annotations like bounding boxes, endowing capabilities of visual referring and fine-grained visual perception.

Name	Modality	Scenario	Manual Annotation	LLM / LMM Synthesis	I/V/A source	# I/V/A	# Samples	Date
Scenario-Oriented								
VQAv2 [23]	I	General VQA	✓		VQA	82.8K	443.7K	2016/12
GQA [24]	I	General VQA			VG	113K	22.7M	2019/02
OKVQA [280]	I	General VQA	✓		COCO	9K	9K	2019/05
VSR [281]	I	General VQA	✓		COCO	2.2K	3.3K	2022/04
A-OKVQA [282]	I	General VQA	✓		COCO	16.5K	17.1K	2022/06
CLEVR [30]	I	General VQA			From Blender	70K	700K	2016/12
VizWiz [283]	I	General VQA	✓		Web (Vizwiz Application)	20.0K	20.0K	2018/02
Visual7W [284]	I	General VQA	✓		COCO	14.4K	69.8K	2015/11
Hateful Memes [285]	I	General VQA	✓		Web (Getty Images)	8.5K	8.5K	2020/05
TallyQA [286]	I	General VQA	✓		COCO, VG	133.0K	249.3K	2018/01
ST-VQA [287]	I	General VQA	✓		Multiple Datasets	19.0K	26.3K	2019/05
MapQA [288]	I	General VQA			Web (KFF), From Map-drawing Tools	37.4K	477.3K	2022/11
KVQA [289]	I	General VQA	✓		Wikidata	17K	130K	2019/07
ViQuAE [290]	I	General VQA	✓		Wikipedia, Wikidata, Wikimedia Commons	1.1K	1.2K	2022/07
ActivityNet-QA [291]	V	General VQA	✓		ActivityNet	3.2K	32K	2019/06
NExT-QA [292]	V	General VQA	✓		VidOR	3.9K	37.5K	2021/05
CLEVRER [293]	V	General VQA			From Bullet Physics Engine	10K	152.6K	2019/10
WebVidQA [294]	V	General VQA			WebVid2M	2.4M	3.5M	2022/05
TGIF-QA [295]	V	General VQA	✓		TGIF Dataset	62.8K	139.4K	2017/04
STAR [296]	V	General VQA	✓		Charades	13.2K	36K	2024/05
HowtoVQA69M [294]	V	General VQA			HowTo100M	62M	62M	2022/05
TVQA [297]	V	General VQA	✓		Web (TV Shows)	16.8K	121.6K	2018/09
NewsVideoQA [298]	V	General VQA	✓		Web (Youtube)	7.0K	2.4K	2022/11
IAM [299]	I	General OCR			From Human Written	5.7K	5.7K	2001/09
OCRVQA [300]	I	General OCR	✓		Book Cover Dataset	165.7K	801.6K	2019/09
TextVQA [301]	I	General OCR	✓		Open Images v3	22.0K	34.6K	2019/04
RenderedText [302]	I	General OCR			From Blender	1M	1M	2023/06
SynthDog-EN [303]	I	General OCR			From SynthDog Tools	0.5M	0.5M	2021/11
DocVQA [304]	I	Doc/Chart/Screen	✓		Web(UCSF IDL)	10.2K	39.5K	2020/07
Chart2Text [305]	I	Doc/Chart/Screen	✓		Web (Statista)	27.0K	30.2K	2022/03
DVQA [306]	I	Doc/Chart/Screen			Matplotlib tools	200K	2.3M	2018/01
ChartQA [307]	I	Doc/Chart/Screen	✓		Web	18.3K	28.3K	2022/03
PlotQA [308]	I	Doc/Chart/Screen	✓		Web, From Manual Plot	157.1K	20.2M	2019/09
FigureQA [309]	I	Doc/Chart/Screen			From Bokeh	100K	1.3M	2017/10
InfoVQA [310]	I	Doc/Chart/Screen	✓		Web	4.4K	23.9K	2021/04
ArxivQA [255]	I	Doc/Chart/Screen		✓(GPT-4V)	Web (ArXiv Papers)	28.8K	14.9K	2024/03
TabMWP [311]	I	Doc/Chart/Screen	✓		Web (IXL)	22.7K	23.1K	2022/09
ScreenQA [312]	I	Doc/Chart/Screen	✓		RICO	28.3K	68.8K	2022/09
VisualMRC [313]	I	Doc/Chart/Screen	✓		Web	7.0K	21.0K	2021/01
DUDE [314]	I	Doc/Chart/Screen	✓		Web	3.0K	23.7K	2023/05
MP-DocVQA [315]	I	Doc/Chart/Screen	✓		SingleDocVQA	4.8K	36.8K	2022/12
DocGemini [150]	I	Doc/Chart/Screen		✓(Gemini-Pro)	DocVQA, ChartQA, InfoVQA	30K	195K	2024/04
Geo170K [316]	I	Math/Science/Code		✓(ChatGPT)	GeoQA+, Geometry3k	9.1K	177.5K	2023/12
GeoQA+ [317]	I	Math/Science/Code	✓		GeoQA, Web	-	6.0K	2022/10
Geomverse [318]	I	Math/Science/Code			-	9.3K	9.3K	2023/12
RAVEN [319]	I	Math/Science/Code			From Rendering Engine	42K	42K	2019/03
ScienceQA [320]	I	Math/Science/Code			Web (IXL)	5.0K	6.2K	2022/09
Geometry3k [321]	I	Math/Science/Code	✓		Web (McGraw-Hill, Geometryonline)	1.5K	2.1K	2021/05
AI2D [322]	I	Math/Science/Code	✓		Web (Google Image Search)	3.1K	9.7K	2016/03
IconQA [323]	I	Math/Science/Code	✓		Web (IXL)	27.3K	29.9K	2021/10
TQA [324]	I	Math/Science/Code	✓		Web (CK-12)	1.5K	6.5K	2017/07
WebSight [325]	I	Math/Science/Code		✓(Open Models)	From Playwright	500K	500K	2024/03
DaTikz [326]	I	Math/Science/Code		✓(Open Models)	Web	48.0K	48.3K	2023/09
Design2Code [327]	I	Math/Science/Code	✓		C4	0.5K	0.5K	2024/03
CLEVR-MATH [328]	I	Math/Science/Code			CLEVR	70K	788.7K	2022/08
GRIT [329]	I	Detection & Grounding			COYO-700M, LAION-2B	91M	137M	2023/06
Visual Genome [330]	I	Detection & Grounding	✓		COCO, YFCC100M	64.9K	1.1M	2016/02
RefCOCO [29]	I	Detection & Grounding	✓		COCO	17.0K	120.6K	2014/10
RefCOCO+ [29]	I	Detection & Grounding	✓		COCO	17.0K	120.2K	2014/10
RefCOCOg [29]	I	Detection & Grounding	✓		COCO	21.9K	80.5K	2014/10
Objects365 [331]	I	Detection & Grounding	✓		Web (Flicker)	600K	10.1M	2019/10

Table 5. Summary of commonly used "Scenario-oriented" datasets for training LMMs. The notations follow Table 3.

 doi:10.20944/preprints202411.0685.v1

Instruction-following Data

Based on the multi-modal representations aligned using pre-training data, LMMs further leverage instruction-following data to learn how to comprehend and follow instructions in multi-modal contexts, as well as develop the ability to solve diverse tasks. As revealed by the success of instruction-tuned LMMs [383], rich and comprehensive instruction-following data is the key to help models learn generalizable capabilities. Inspired by the prevalent methods to construct textual instruction data [52,384], researchers typically create multi-modal data through two approaches.

Reformulating Task-oriented Datasets. As discussed in Section 2, during the development of multi-modal research, a large amount of datasets have been established for various scenarios and tasks. These datasets are typically collected, annotated, and validated by humans according to corresponding requirements, ensuring high quality. To meet the demands of the tasks, this type of data generally exists in specific input-output formats [23,25,29] and requires additional processing to be reformulated into instruction-following data. Various datasets have been reformulated as listed in Table 6.

Name	Modality	I/V/A Source	Instruction Source	Response Source	# I/V/A	# Samples
Reformulated Datasets						
MultInstruct [332]	I	Multiple Datasets	Human	Annotation	-	510K
MANTIS [333]	I, V	Multiple Datasets	-	Annotation	-	989K
X-LLM [334]	I, V, A	MiniGPT-4, AISHELL-2, ActivityNet, VSDIal-CN	Human	Annotation, Human	4.5K/1K/2K	10K
M3IT [335]	I, V	Multiple Datasets	Human	Annotation, ChatGPT	-	2.4M
InstructBLIP [113]	I, V	Multiple Datasets	Human	Annotation	-	-
OMNIINSTRUCT [336]	V, A	AVQA, Music-AVQA2.0, MSRVTT-QA	-	Annotation, InternVL-2-76B	-	93K
VideoChat2 [337]	I, V	Multiple Datasets	ChatGPT	Annotation	-	1.9M
TimeIT [338]	V	Multiple Datasets	GPT-4, Human	Annotation	-	125K
Vision-Flan [339]	I	Multiple Datasets	Human	Annotation, Human	-	1.6M
ChiMed-VL-Instruction [340]	I	PMC-Report, PMC-VQA	-	Annotation, ChatGPT	-	469K
MultiModal-GPT [341]	I	Multiple Datasets	GPT-4	Annotation	-	284.5K
The Cauldron [158]	I	Multiple Datasets	-	Annotation	-	-
MIC [342]	I, V	Multiple Datasets	ChatGPT	Annotation	-	5.8M
Video-LLaMA [91]	I, V	MiniGPT-4, LLaVA, VideoChat	-	Annotation	81K/8K	171K
PandaGPT [162]	I	LLaVA, MiniGPT-4	-	Annotation	81K	160K
mPLUG-DocOwl [118]	I	Multiple Datasets	-	Annotation	-	-
UReader [79]	I	Multiple Datasets	-	Annotation	-	-

Table 6. Summary of commonly used multi-modal instruction-following datasets curated by "Reformulation". In the "Modality" column, I, V, and A represent Image, Video, and Audio, respectively. In the "Instruction Source" column, the "-" mark means the same method is adopted as in "Response Source", or the definition of "Instruction" is not emphasized in the presented dataset. In the "Response Source" column, "Annotation" stands for the annotated labels in the original datasets.

The most common solution is to introduce templates, providing an appropriate textual description of the task (i.e., instructions) as well as a question-and-answer format, and correctly placing the input and output of each sample in their respective positions [336,385,386]. In this case, specific templates need to be designed based on the corresponding task. Early works rely on annotators for template design [113,332,334,339,387], and certain specific tasks might require image processing, such as image concatenation [388] and object annotation [342]. Later, researchers also adopt tools like ChatGPT or Gemini-Pro to assist in template design and extension [121,150,389–391], and further rephrase brief and incomplete responses into longer, more complete sentences [392,393].

Furthermore, diverse reformulated datasets can be organically integrated for joint training to empower LMMs with a wide range of capabilities. For example, Cauldron [152] converts multiple samples of the same context (such as the same image) into multi-turn conversations. In MANTIS [333], each data item contains multiple images and multiple turns of QA pairs with a suitable text-image interleaving format. LEOPARD-INSTRUCT [377] spans key domains commonly encountered in real-

 doi:10.20944/preprints202411.0685.v1

world scenarios, such as multi-page documents, charts, tables, and webpage trajectories, tailored to text-rich, multi-image contexts, while Cambrian-10M [88] considers different types of data from the perspective of capabilities, balances their proportions, and enhances model performance in knowledge-intensive tasks.

Self-Instruction Although task-oriented datasets facilitates LMMs to learn diverse capabilities, the knowledge entailed is limited to the corresponding scenarios. At the same time, the template formats are also restricted to specific tasks and mainly differ from general human-machine interaction. To further enrich the instruction data, powerful proprietary models can be leveraged to generate data according to the provided information and requirements. The most typical approach is motivated by the self instruct method [394]. To prompt ChatGPT-like models to generate instruction data based on input multi-modal information, task descriptions and specific requirements are provided through system prompts and user queries. Additionally, in-context examples can be included to offer detailed guidance that cannot be well defined through language. Table 7 includes several datasets curated with similar methods.

Name	Modality	I/V/A Source	Instruction Source	Response Source	# I/V/A	# Samples
Datasets curated by Self-Instruct						
MiniGPT-4 [76]	I	CC3M, CC12M	-	Pre-trained Model	3.5K	3.5K
DetGPT [346]	I	COCO	-	ChatGPT	5K	30K
Shikra-RD [122]	I	Flickr30K	-	GPT-4	-	0.6K
MGVLID [347]	I	Multiple Datasets	-	GPT-4	1.2M	3M
MMDU-45k [348]	I	Web (Wikipedia)	-	GPT-4o	-	45K
LLaVA [251]	I	COCO	-	GPT-4	80K	158K
PVIT [349]	I	Multiple Datasets	-	ChatGPT	-	146K
MAVIS-Instruct [350]	I	Multiple Datasets	-	GPT-4V	611K	834K
ALLaVA [351]	I	LAION, Vision-FLAN	-	GPT-4V	663K	663K
AS-V2 [352]	I	COCO	-	GPT-4V	-	127K
GPT4Tools [353]	I	-	-	ChatGPT	-	71K
LLaVA-Video-178K [354]	V	Multiple Datasets	-	GPT-4o	178K	1.3M
LLaVA-Hound [355]	V	ActivityNet, WebVid, VIDAL	-	ChatGPT	80K	240K
Square-10M [356]	I	Web	-	Gemini Pro	3.8M	9.1M
MIMIC-IT [357]	I, V	Multiple Datasets	-	ChatGPT	8.1M/502K	2.8M
Valley-Instruct [90]	V	Web	-	ChatGPT	-	73K
LLaVAR [358]	I	LAION-5B	-	GPT-4	16K	16K
LVIS-INSTRUCT4V [359]	I	LVIS	-	GPT-4V	110K	220K
LRV-Instruction [360]	I	VG, Vistext, Visualnews	-	GPT-4	-	400K
MM-Instruct [361]	I	SA-1B, DataComp-1B	ChatGPT	Mixtral-8x7b	-	234K
SVIT [362]	I	VG	-	GPT-4	108.1K	4.2M
LLaVA-Med [363]	I	PMC-15M	-	GPT-4	60K	60K
VIGC [364]	I	COCO, Objects365	-	VIG, VIC Models	-	1.8M
OphGLM [365]	I	Web	-	ChatGPT	-	20K
TEXTBIND [366]	I	CC3M	-	GPT-4	-	25.6K
Video-ChatGPT [123]	V	ActivityNet-200	-	Human, GPT-3.5	100K	100K
COSMIC [367]	A	TED-LIUM 3	-	GPT-3.5	50K	856K
SparklesDialogue [368]	I	CC3M, VG	-	GPT-4	20K	6.5K
AnyInstruct-108k [12]	I, A	From Diffusion Models	-	GPT-4	205K/616K	108K
MosIT [11]	I, V, A	Web	-	GPT-4, AIGC Tools	4K/4K/4K	5K
StableLLaVA [369]	I	From Stable Diffusion	-	ChatGPT	126K	126K
T2M [11]	I, V, A	Webvid, CC3M, AudioCap	-	GPT-4	5K/5K/5K	15K
DocReason25K [240]	I	Multiple Datasets	-	GPT-3.5, GPT-4V	8.7K	25K
Clotho-Detail [370]	A	Clotho	-	GPT-4	-	3K
MMEvol [371]	I	SEED-163K	-	GPT-4o, GPT-4o-mini	-	447K
InstructS2S-200K [171]	A	CosyVoice-300M-SFT, VITS	-	Llama-3-70B-Instruct	200K	200K
MMINSTRUCT [372]	I	Web		GPT-4V, GPT-3.5	161K	973K
VCG+ 112K [373]	V	Web		GPT-4, GPT-3.5	-	112K

Table 7. Summary of commonly used instruction-following datasets curated by "Self-instruct". The same notations are used as in Table 6.

For images, early works represented by LLaVA [6] provide visual information to ChatGPT via text descriptions including captions, objects and bounding boxes [122,347,395]. Similarly, for audio-text instruction-tuning data, audio captions and labels from the original dataset are fed into ChatGPT to generate QA pairs [166,370]. In particular, if the audio data contains talks or speeches, the

 doi:10.20944/preprints202411.0685.v1

original transcripts are also provided to ChatGPT as additional information [380,396]. Later, GPT-4V is accessible and mainly employed to directly perceive multi-modal inputs and generate fine-grained captions, diverse questions and detailed answers [351,352,359,363]. GPT-4o accepts any combination of text, audio, images, and video as input, allowing the generation based on more complex multi-modal information [156,348,354].

The general data generation process can be further optimized based on specific objectives. To acquire data of higher quality, GPT-4V and Gemini Pro can be employed to evaluate the generated data and discard samples containing hallucinatory content, meaningless questions, or erroneous answers [240,356,371]. In addition, the strong instruction-following capability of proprietary models can be utilized to generate data in specific formats such as negative instruction [360], complex questions [351], and multi-modal chain-of-thoughts [350,379,381]. Different from generating textual answers by LMMs, some multi-modal generative tools are adopted to convert textual descriptions into multi-modal elements, to meet the requirements of any-to-any LMMs [11,12].

Please notice that some datasets are constructed with both of the aforementioned methods, as presented in Table 8.

Name	Modality	I/V/A Source	Instruction Source	Response Source	# I/V/A	# Samples
Reformulation + Self-Instruct						
Cambrian-10M [88]	I	Multiple Datasets, Web	-	Annotation, GPT-3.5	-	9.8M
MULTIS [375]	I, V, A	Multiple Datasets	-	Annotation, GPT-4	-	4.6M
X-InstructBLIP [376]	I, V, A	Multiple Datasets	-	Annotation, Flan-T5-XXL	-	24K
LEOPARD-INSTRUCT [377]	I	Multiple Datasets	-	Annotation, GPT-4o	-	925K
LAMM [378]	I	Multiple Datasets	-	Annotation, GPT-API	-	186K
VoCoT [379]	I	GQA, LVIS, LLaVA-Instruct	-	Annotation, GPT-4V	-	80K
OPEN-ASQA [380]	A	Multiple Datasets	-	Annotation, GPT-3.5	1.1M	9.6M
SALMONN [166]	A	Multiple Datasets	-	Annotation, ChatGPT	-	2.3M
Visual CoT [381]	I	Multiple Datasets	-	Annotation, GPT-4	-	438K
M4-Instruct [363]	I, V	Multiple Datasets	-	Annotation, GPT-4V	-	1.2M
Web2Code [382]	I	GPT-3.5, WebSight, Pix2Code, WebSRC	-	Annotation, GPT-4	-	1.2M

Table 8. Summary of datasets with both reformulated and self-instruct generated data. The notations follow Table 6.

4.2.2. Training Stages

The training of current LMMs typically involve multiple stages, with each stage using different data to train specific parameters, gradually learning cross-modal alignment as well as multi-modal understanding and generation capabilities. Most LMMs undergo two main stages: pre-training and instruction fine-tuning. Some models also introduce additional training stages for learning specific capabilities.

Pre-training

The primary goal of pre-training is to align and associate the input representations of various modalities within the multi-modal input space, enabling the backbone to uniformly model and understand inputs across modalities. Figure 5 illustrates the commonly applied settings in the pre-training phase which is described below.

 doi:10.20944/preprints202411.0685.v1

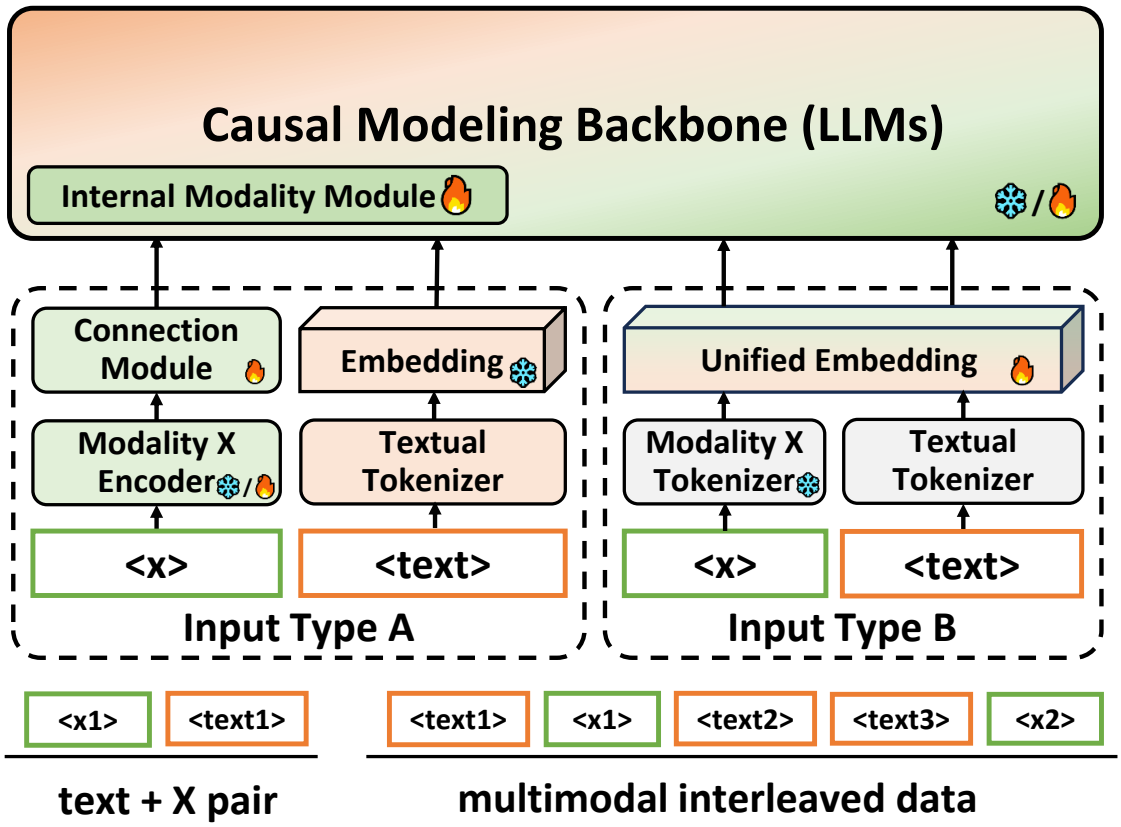

Figure 5. Illustration of common settings during the **pre-training stage**, including data and trainable parameters. "<x>" represents inputs of modalities other than text.

Training data. As mentioned in the training data section, commonly used data include X-text pairs and multi-modal interleaved documents. To improve the pre-training efficiency, existing methods propose to further enhance the quality of pre-training data through filtering and sampling [141]. Additionally, scenario-oriented data can be considered to enhance specific capabilities of LMMs. Besides multi-modal data, text-only data can be adopted to maintain the language modeling capabilities of backbones [84,106,133].

Training objective. With the input-level alignment architecture, the training data can be constructed as multi-modal sequences. For (Output Type 1) LMMs with text-only output, models are typically required to generate textual tokens in the sequence based on information from other modalities, thereby learning semantic alignment [5,6]. For LMMs with multi-modal generation capabilities, the training objectives also include predicting information in other modalities in the sequence, whether in continuous (Output Type 2) [75,109,111] or discrete (Output Type 3) [22,112,172] formats.

Trainable parameters. For LMMs with hybrid input space of Type A, input-level alignment can be efficiently achieved by merely updating the connection module [5,6,128], while some methods further unlock the backbone [135,147] or modality encoder [7,118,143] to increase the trainable model capacity, thereby enhancing specific capabilities like in-context learning [135] and adapting to multi-modal contexts. Regarding models with unified Type B input space, most methods jointly train the extended multi-modal embeddings and the backbones [12], additional training strategies may be required to stabilize the training process [22,112]. As an exception, Ovis [153] treats the visual embedding as a separate module, training only the corresponding parameters during the pre-training phase. Meanwhile, if LMMs are designed with internal alignment modules or output mapping modules, the modules are typically updated during the pre-training stage [109,115,130,397]. When modality encoders and backbones are activated for training, LoRA [398] tuning can be adopted to retain the pre-trained knowledge and improve training stability [129,152,158].

 doi:10.20944/preprints202411.0685.v1

Multi-stage pre-training. A line of studies have explored dividing pre-training into two stages, using different forms of data in each stage [84,109,152], introducing higher-quality data and task-specific data in the latter stage [7,129,130], gradually unlocking more trainable modules [7,84,153] and enhancing the input image resolution [130,143,152].

Specialized setup. Apart from general settings, some methods adopt special training strategies. InternLM-XComposer 2 [107] and mPLUG-Owl2 [132] utilize a layer-wise learning rate decay method to maintain pre-trained knowledge while updating the modality encoders. Unlike updating the entire modality encoders, ShareGPT4V [131] and TinyLLaVA [142] demonstrate that merely training the latter half of the layers is more effective and efficient. Considering the training stages, different from the aforementioned discussions, specialized pre-training stages and objectives can be used to train complex connection modules [5,114]. IDEFICS3 [158] even conducts three-stage pre-training to learn different levels of capabilities in a more refined manner.

Instruction Fine-tuning

The instruction fine-tuning stage is a crucial step in developing LMMs as versatile AI systems. It enables the model to understand and follow instructions to generate appropriate responses, thereby enhancing the interactivity. Additionally, by fine-tuning with diverse instruction data, the stage improves the generalization capabilities of LMMs, facilitating models to handle unseen scenarios and tasks in a zero-shot manner to meet real-world requirements. Figure 6 provides a straightforward illustration for this stage.

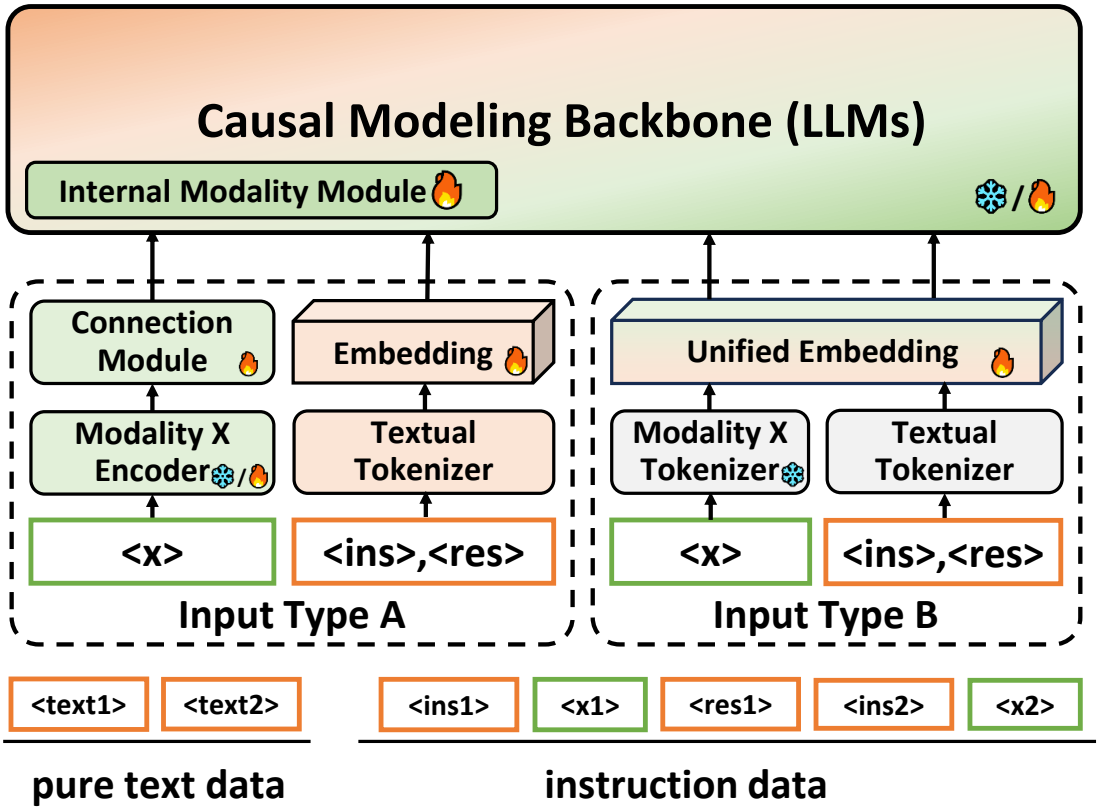

Figure 6. Illustration of common settings during the **instruction fine-tuning stage**, where <x>, <ins>, and <res> denote inputs of modalities other than text, instruction, and response, respectively.

Training data. As previously introduced in the data section, instruction-tuning data is required to be sufficiently rich. Therefore, most LMMs adopt different strategies to construct a mixed dataset based on different requirements, such as mixing task-oriented data with self-instructed data [128,152], combining general data with data from specific scenarios [122,397], unifying data from various modal-

 doi:10.20944/preprints202411.0685.v1

ities [11,12,363], integrating understanding and generation data [75,109], and blending multi-modal data and text-only data [135,144].

Training objective. Integrated with system prompts, instruction-following data is typically formatted in specific templates to form dialogue sequences. The training objective is to generate multi-modal content within the sequence. Unlike the pre-training stage, only the contents in responses are used for gradient calculation, while the user instructions and the system prompt are masked [6].

Trainable parameters. As discussed in [128], merely fine-tuning the connection module makes it difficult for LMMs to adapt to complex multi-modal contexts. Therefore, current LMMs typically update the backbones during the instruction fine-tuning stage, either through full-parameter tuning or using PEFT (parameter-efficient fine-tuning) methods like LoRA [398]. As for small-scale LMMs, researchers have found that utilizing LoRA can reduce the risk of catastrophic forgetting [141,151]. In addition, internal alignment and output mapping modules are jointly fine-tuned in this stage, if they are included in the LMM [9,109,115,130]. As for modality encoders, most methods tend to keep them frozen in this stage, while some specific models activate them to learn specific encoding knowledge [80,141,143,153].

Specialized setup. Different from the prevalent multi-stage setting, SPHINX-X [81] and PandaGPT [162] resort to a one-stage training framework and mainly utilize instruction data to jointly learn cross-modal alignment and the ability to follow multi-modal instructions. To enhance capabilities in specific scenarios, some instruction fine-tuned models are further trained with scenario-oriented data. For instance, IDEFICS2 [152] is further adapted to long conversations, while InternLM-XComposer series [107,157] mainly focus on the article composition.

Additional Alignment Stages

In addition to the regular pre-training and instruction fine-tuning stages, some specialized models require additional training stages to achieve alignment for specific objectives.

Output-level alignment. To enable LMMs to generate multi-modal response, output-level alignment is required. Benefiting from unified multi-modal discrete representation and the pre-trained tokenizer and detokenizer for each modality, models with Type 3 output space can achieve output-level alignment directly through conventional pre-training and instruction fine-tuning [12,22,112,172]. For models with Type 2 hybrid output space, an additional alignment stage may be required. By rearranging the order of text and other-modality information in "text + X" pairs and interleaved sequences, the text-to-other-modalities generation ability can be learned in the autoregressive setting. A line of approaches keeps modality decoders frozen and train the output mapping modules through gradients passed from the decoder for alignment [103,109,110]. Since most modality decoders are originally conditioned on text for generation, the representations from the decoders' corresponding text encoder can be utilized as supervision signal [9,11]. Another line of methods, represented by Emu series [75,111], propose to construct an autoencoder architecture between modality encoders and decoders. These methods first train LMMs to align the visual input and output spaces, then align the modality decoders to this space.

Sparse internal alignment. Specifically, LMMs integrated with MoE-based internal alignment modules typically undergo an additional sparsification stage, referred as sparse upcycling [199]. During this stage, MoE layers are added to an already established dense LMMs. The parameters of MoE layers are fine-tuned using multi-modal instruction datasets, which further facilitate modality alignment and fusion while mitigating conflicts [138,139,155,163].

We summarize several LMMs and the corresponding training settings in Table 9 and 10.

 doi:10.20944/preprints202411.0685.v1

Model	Pre-training						Instruction Fine-tuning			Date
	Trainable Parameters		Training Data			Multi-stage training	Trainable Parameters		Training Data	
	Modality Encoder	LLM Backbone	Scene-oriented Data	Text-Only	Interleaved		Modality Encoder	LLM Backbone	Text-Only Data	
Flamingo [115]	✗	✗	✗	✗	✓	✗	-	-	-	2022/04
BLIP-2 [5]	✗	✗	✗	✗	✗	✓	-	-	-	2023/01
LLaMA-adapter[116]	-	-	-	-	-	-	✗	✓(P)	✓	2023/03
MiniGPT-4 [117]	✗	✗	✗	✗	✗	✗	✗	✗	✗	2023/04
LLaVA [6]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✗	2023/04
mPLUG-Owl [118]	✓	✗	✗	✗	✗	✗	✗	✓(P)	✓	2023/04
LLaMA-adapter v2[119]	-	-	-	-	-	-	✗	✓(P)	✓	2023/04
InstructBLIP [113]	✗	✗	✗	✗	✗	✓	✗	✗	✗	2023/05
Otter [92]	✗	✗	✗	✗	✓	✗	✗	✗	✗	2023/05
LAVIN [120]	-	-	-	-	-	-	✗	✓(P)	✓	2023/05
MultimodalGPT [121]	✗	✗	✗	✗	✓	✗	✗	✓(P)	✓	2023/05
Shikra [122]	✗	✓(F)	✓	✗	✗	✗	✗	✓(F)	✗	2023/06
VideoChatGPT [123]	-	-	-	-	-	-	✗	✗	✗	2023/06
Valley [90]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✗	2023/06
Lynx [124]	✗	✗	✓	✗	✗	✓	✗	✓(P)	✓	2023/07
Qwen-VL [7]	✓	✓(F)	✓	✓	✓	✓	✗	✓(F)	✓	2023/08
BLIVA [125]	✗	✗	✗	✗	✗	✗	✗	✗	✗	2023/08
IDEFICS [126]	✗	✗	✗	✗	✓	✗	-	-	-	2023/08
OpenFlamingo [127]	✗	✗	✗	✗	✓	✗	-	-	-	2023/08
InternLM-XC [106]	✗	✓(F)	✗	✓	✓	✗	✗	✓(P)	✓	2023/09
LLaVA-1.5 [128]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2023/10
MiniGPT-v2 [129]	✗	✓(P)	✓	✗	✗	✓	✗	✓(P)	✓	2023/10
Fuyu-8B [64]	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	2023/10
UReader [79]	-	-	-	-	-	-	✗	✓(P)	✗	2023/10
CogVLM [130]	✗	✗	✓	✗	✗	✓	✗	✓(F)	✗	2023/11
Monkey [80]	-	-	-	-	-	-	✓	✓(F)	✗	2023/11
ShareGPT4V [131]	✓	✓(F)	✗	✗	✗	✗	✓	✓(F)	✓	2023/11
mPLUG-Owl2 [132]	✓	✗	✗	✗	✗	✗	✓	✓(F)	✓	2023/11
Sphinx [133]	✗	✓(F)	✗	✓	✗	✗	✗	✓(F)	✓	2023/11
InternVL [114]	✓	✗	✗	✗	✗	✓	✗	✓(F)	✓	2023/12
MobileVLM [134]	✗	✗	✗	✗	✗	✗	✗	✓(P)	✗	2023/12
VILA [135]	✗	✓(F)	✗	✗	✓	✓	✗	✓(F)	✓	2023/12
Osprey [77]	✗	✗	✓	✗	✗	✓	✗	✓(F)	✗	2023/12
Honeybee [136]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2023/12
Omni-SMoLA [137]	-	-	-	-	-	-	✗	✓(P)	✗	2023/12
LLaVA-Next [83]	✗	✗	✗	✗	✗	✗	✓	✓(F)	✓	2024/01
InternLM-XC2 [107]	✓	✗	✓	UNK	✗	✗	✓	✓(F)	✓	2024/01
Mousi [89]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2024/01
LLaVA-MoLE [138]	✗	✗	✗	✗	✗	✗	✗	✓(P)	✓	2024/01
MoE-LLaVA [139]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2024/01
MobileVLM v2 [140]	✗	✓(F)	✗	✗	✗	✗	✗	✓(F)	✓	2024/02
Bunny [141]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2024/02
TinyLLaVA [142]	✓	✓(F)	✗	✗	✗	✗	✓	✓(F)	✓	2024/02
Sphinx-X [81]	-	-	-	-	-	-	✗	✓(F)	✓	2024/02
Mini-Gemini [87]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2024/03
Deepseek-VL [84]	✗	✓(F)	✓	✓	✓	✓	✓	✓(F)	✓	2024/03
LLaVA-UHD [82]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✗	2024/03
Yi-VL [143]	✓	✗	✓	✗	✗	✓	✓	✓(F)	✗	2024/03
MM1 [144]	✓	✓(F)	✗	✓	✓	✗	✓	✓(F)	✓	2024/03
VL Mamba [145]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2024/03
Cobra [146]	-	-	-	-	-	-	✗	✓(F)	✓	2024/03
InternVL 1.5 [147]	✓	✗	✓	✗	✗	✗	✓	✓(F)	✓	2024/04
Phi-3-Vision [148]	UNK	UNK	✓	✓	✓	✗	UNK	✓(F)	✓	2024/04
PLLaVA [149]	-	-	-	-	-	-	✗	✓(P)	✗	2024/04
Imp [151]	✗	✗	✗	✗	✗	✗	✗	✓(P)	✓	2024/05
IDEFICS2 [152]	✓	✓(P)	✓	✗	✓	✓	✓	✓(P)	✓	2024/05
ConvLLaVA [78]	✓	✓(F)	✗	✗	✗	✓	✗	✓(F)	✓	2024/05
Ovis [153]	✓	✗	✗	✗	✗	✓	✓	✓(F)	✓	2024/05
Deco [154]	✗	✗	✗	✗	✗	✗	✗	✓(P)	✓	2024/05
CuMo [155]	✓	✓(F)	✗	✗	✗	✓	✓	✓(F)	✓	2024/05
Cambrian-1 [88]	✗	✗	✗	✗	✗	✗	✗	✓(F)	✓	2024/06
GLM-4v [156]	✓	✓(F)	✓	✓	✗	✓	✓	✓(F)	✗	2024/06
InternLM-XC2.5 [157]	✓	✗	✓	UNK	✗	✗	✓	✓(F)	✓	2024/07
IDEFICS3 [158]	✓	✓(P)	✓	✗	✓	✓	✓	✓(P)	✓	2024/08
mPLUG-Owl3 [159]	✗	✓(F)	✓	✗	✓	✓	✗	✓(F)	✓	2024/08
CogVLM2 [156]	✓	✓(F)	✓	✓	✗	✓	✓	✓(F)	✗	2024/08
CogVLM2-video [156]	✓	✓(F)	✓	✓	✗	✓	✓	✓(F)	✗	2024/08
LLaVA-OV [160]	✓	✓(F)	✓	✓	✓	✓	✓	✓(F)	✓	2024/09
Qwen2-VL [161]	✓	✓(F)	✓	✓	✓	✓	✗	✓(F)	✓	2024/09

Table 9. Summary of specific training settings of LVLMs that focus on understanding tasks with only text output (Output Type 1). If there is "✓" in a column indicates that this special setting is enabled during the training process. Conversely, "✗" indicates that it is not enabled. '-' represents that the training stage is not applicable. "✓(P)" represents utilizing PEFT (Parameter-Efficient Fine-Tuning) method to tuning LLM typically like LoRA, "✓(F)" indicates Full Parameter Fine-Tuning, "UNK"represents an officially unknown setting.

 doi:10.20944/preprints202411.0685.v1

Model	Pre-training						Instruction Fine-tuning			Date
	Trainable Parameters		Training Data			Multi-stage training	Trainable Parameters		Training Data	
	Modality Encoder	LLM Backbone	Scene-oriented Data	Text-Only	Interleaved		Modality Encoder	LLM Backbone	Text-Only Data	
Any-Modality LMMs										
PandaGPT [162]	-	-	-	-	-	-	✗	✓(P)	✗	2023/05
ImageBind-LLM [102]	✗	✗	✗	✗	✓	✗	✗	✓(P)	✓	2023/09
Next-GPT [11]	✗	✓(P)	✗	✗	✗	✓	✗	✓(P)	✗	2023/09
Codi-2 [103]	-	-	-	-	-	-	✗	✓(P)	✓	2023/11
UnifiedIO2 [104]	✗	✓(F)	✓	✓	✓	✗	✗	✓(F)	✓	2023/12
AnyGPT [12]	✗	✓(F)	✓	✗	✓	✗	✗	✓(F)	✓	2024/02
Uni-MoE [163]	✗	✗	✓	✗	✗	✗	✗	✓(P)	✗	2024/05
Large Audio-Language Models										
SpeechGPT [164]	✗	✓(F)	✗	✗	✗	✗	✗	✓(F/P)	✓	2023/05
Speech-LLaMA [165]	-	-	-	-	-	-	✗	✓(P)	✗	2023/07
SALMONN [166]	✗	✓(P)	✗	✗	✗	✗	✗	✓(F)	✗	2023/10
Qwen-Audio [167]	✓	✗	✓	✗	✗	✗	✗	✓(F)	✗	2023/11
SpeechGPT-Gen [10]	-	-	-	-	-	-	✗	✓(F)	✗	2024/01
SLAM-ASR [8]	-	-	-	-	-	-	✗	✗	✗	2024/02
WavLLM [168]	✗	✓(P)	✓	✓	✗	✗	✗	✓(P)	✗	2024/04
SpeechVerse [169]	-	-	-	-	-	-	✗	✓(P)	✗	2024/05
Qwen2-Audio [170]	✓	✓(F)	✓	✗	✗	✗	✓	✓(F)	✗	2024/07
LLaMA-Omni [171]	-	-	-	-	-	-	✗	✓(F)	✗	2024/09
Large Vision-Language Models for Multi-Modal Generation										
GILL [9]	✗	✗	✗	✗	✓	✗	-	-	-	2023/05
Emu [111]	✓	✓(F)	✗	✗	✓	✗	✗	✓(P)	✓	2023/07
LaVIT [172]	✗	✓(F)	✗	✓	✗	✗	-	-	-	2023/09
CM3Leon [173]	✗	✓(F)	✗	✗	✓	✗	✗	✓(F)	✗	2023/09
DreamLLM [109]	✗	✓(F)	✗	✗	✓	✓	✗	✓(F)	✓	2023/09
Kosmos-G [174]	✗	✓(F)	✗	✓	✓	✗	✗	✓(F)	✗	2023/10
SEED-LLaMA [112]	✗	✓(F/P)	✗	✗	✓	✗	✗	✓(P)	✓	2023/10
MiniGPT-5 [110]	✗	✗	✗	✗	✗	✗	✗	✓(P)	✗	2023/10
Emu-2 [75]	✓	✓(F)	✓	✓	✓	✓	✓	✓(F)	✓	2023/12
Chameleon [22]	✗	✓(F)	✗	✓	✓	✓	✗	✓(F)	✓	2024/05
MoMA [175]	✗	✓(F)	✗	✓	✓	✓	-	-	-	2024/07
Vila-U [176]	✗	✓(F)	✗	✗	✓	✗	-	-	-	2024/09

Table 10. Supplement to Table 9. The notations remains the same.

5. Evaluation and Benchmarks

Multimodal evaluation benchmarks are implemented to assess and compare the performance of different LMMs on various tasks. From the perspective of input-output space extension, we categorize existing benchmarks into three types. (1) Based on the input space extended to multiple modalities, **modality comprehension benchmarks** assess perception capabilities of LMMs across multi-modal signals, covering image-to-text, video-to-text, and audio-to-text tasks. (2) With the extension of the output space, **modality generation benchmarks** evaluate the abilities of LMMs to produce multi-modal outputs through images, videos, and audio generation tasks. (3) **Hallucination benchmarks** diagnose whether representations are well aligned between modalities. In this section, we introduce modality comprehension and generation benchmarks and Section 6 focuses on the benchmarks, tasks, and methods for hallucination diagnosis.

5.1. Modality Comprehension Benchmarks

Modality comprehension benchmarks assess the ability of LMMs to understand visual or audio inputs, typically in the form of Any-to-Text tasks, outputting with text. The general evaluation framework is presented in Figure 7.

 doi:10.20944/preprints202411.0685.v1

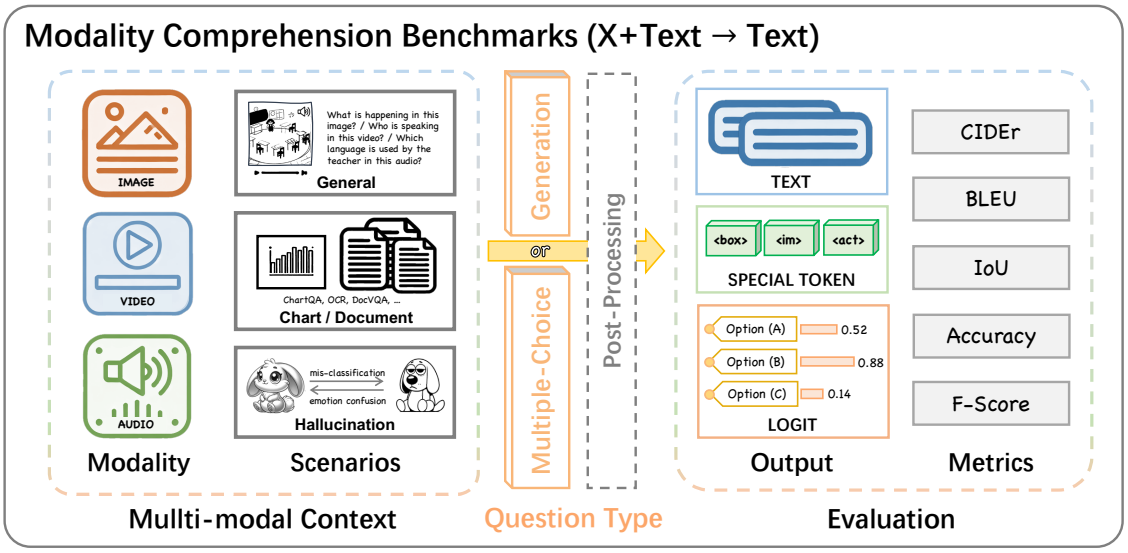

Figure 7. Illustration of the multi-modal comprehension benchmarks that mainly follow an input-output format of "X + Text → Text". Based on the target modality and the scenarios of interest, samples with multi-modal input context are constructed. Appropriate question types are designed and defined using text instructions. After model inference and post-processing, different forms of output are obtained, which are further evaluated according to the corresponding metrics for each task.

5.1.1. Image-to-Text

Currently, a large amount of multi-modal benchmarks focus on images and text. Early benchmarks aims at solving problems in specific tasks with task-oriented forms and data, comprising visual question answering (VQA), referring expression comprehension and image captioning. Specifically, VQA tasks consist of general VQA [23,24,283,287,310], knowledge-based VQA [280,320], and text-oriented VQA focusing on text understanding capabilities in images [300,301,304,307,322]. Referring expression comprehension mainly includes RefCOCO, RefCOCO+, RefCOCOg [29] and GRIT [329], requiring models to localize object queries in images. Besides, COCO [399], NoCaps [27], TextCaps [28] are commonly evaluated for image captioning. These traditional Image-to-Text benchmarks determine whether the output is accurate by fully matching the output with the ground truth. For image captioning tasks, conventional metrics like BLEU [400], ROUGE [401], CIDEr [402], METEOR [403] are implemented to evaluate the quality of the model output, referring annotated captions. Additionally, these benchmarks may require specific output formats, such as single words or phrases, necessitating additional requirements for the model [128] or corresponding evaluation methods [404,405].

As LMMs achieve substantial progress across various downstream tasks, benchmarks specifically for their evaluation have been designed and proposed, which adapt to the free-form text output by these models, comprising multiple-choice questions and open-ended generation tasks. Concretely, input prompts for multiple-choice questions consist of designed instructions, questions, images, and options. The options select by LMMs can be detected from the generated responses through option symbols or text matching methods. MME [406] formalizes the questions into binary terms and evaluates the perceptual and cognitive capabilities of LMMs by asking the model to answer yes or no, while M3Exam [407], MMT-Bench [408], and SEED-Bench-1 [409] further expand the number of options in each question, posing greater challenges to LMMs. Additionally, MMBench [410] proposes to use LLM-based choice extractors to convert the free-form text into a specific choice (A, B, C, etc.), and MMStar [411] introduces new metrics for evaluating multi-modal gain and multi-modal leakage. Besides general scenarios, researchers have also constructed relevant benchmarks targeting specific capabilities and contexts [412–415].

Furthermore, beyond designing questions to guide LMMs to produce outputs in specific formats for evaluation, it is necessary to assess the model's ability when responding with free-form texts in real-world scenarios (open-ended generation). Under this setting, rule-based or LLM-based evalu-

 doi:10.20944/preprints202411.0685.v1

ation pipelines are mainly employed. In rule-based cases, robust regular expressions are deployed to extract key phrases, such as numbers and conclusion phrases, from the responses for accurate answer matching, this method is adopted in MMMU [416], OCR-Bench [417], and MathVista [418]. Similarly, ReForm-Eval [405] and SciGraphQA [419] implement rule-based metrics to process free-form predictions. In addition, LLM-based evaluation methods are more suited for assessing long and complex outputs. These methods typically use powerful large models, such as ChatGPT, to score the model output based on the question description, instructions, visual information, and designed scoring ruless [251,378,420–422]. In addition to automated evaluation pipelines, LVLM-eHub [404] and OwlEval [118] resort to manual evaluation methods, which are directly in line with human preferences but significantly increase the evaluation cost.

5.1.2. Video-to-Text

According to the Any-to-Text paradigm, Video-to-Text comprehension benchmarks are also considered. Video reasoning [291,292,295,423] is a commonly adopted task for evaluation. In related benchmarks, QA pairs are typically generated from existing video descriptions, and the capabilities of LMMs are measured in terms of precision. Similar to image-to-text benchmarks, some benchmarks are designed in the form of multiple-choice questions for efficient and reliable evaluation [296,391,424–427]. Furthermore, researchers are also interested in the video captioning task which is evaluated using traditional metrics [233,248,428]. However, these traditional evaluation metrics rely on the exact matching between the generated and ground-truth captions, which are limited in capturing the richness of video content. Thus, the ChatGPT-assisted evaluation method is applied in VCGBench [123], VCGBench-Diverse [373], MSVC [429], and MLVU [430].

5.1.3. Audio-to-Text

Most existing Audio-to-Text comprehension benchmarks focus on task-oriented evaluation, including Automatic Speech Recognition (ASR), Automatic Speech Translation (AST), Speech Emotion Recognition (SER), Audio Question Answer (AQA), and Audio Captionning (AC). ASR tasks report a word error rate (WER) metric, where a lower number is better [254,431,432], and CoVoST2 [433] is commonly adopted for the translation task with a BLEU [400] score. Meld [434] evaluates SER and ClothoAQA [435] evaluates AQA with accuracy. AudioCaps [229] and Clotho [228] are widely used for the AC task, with CIDEr [402], SPICE [436], or SPIDEr [437] as the metric. Although these benchmarks reveal the capabilities of LMMs from multiple perspectives, they are limited to specific tasks and cannot adequately reflect performance in real-world scenarios. Therefore, AIR-Bench [438] is proposed to conduct assessment that aligns closely with the actual user interaction experience. In addition, AudioBench [439] is introduced as a comprehensive evaluation benchmark specifically designed for general instruction-following audio-language models, and Dynamic-SUPERB [440] is a benchmark that covers comprehensive diverse speech tasks, designed for building universal speech models.

5.2. Modality Generation Benchmarks

Based on the output spaces that are extended to multiple modalities, modality generation benchmarks evaluate the abilities of LMMs to generate multi-modal content. These benchmarks can be categorized into image, video, and audio generation tasks depending on the target modality.

We illustrate the general evaluation framework of generation benchmarks in Figure 8. According to different scenarios, various conditional generation tasks provide multi-modal contexts, and the generated outputs are assessed from different perspectives based on different reference information.

 doi:10.20944/preprints202411.0685.v1

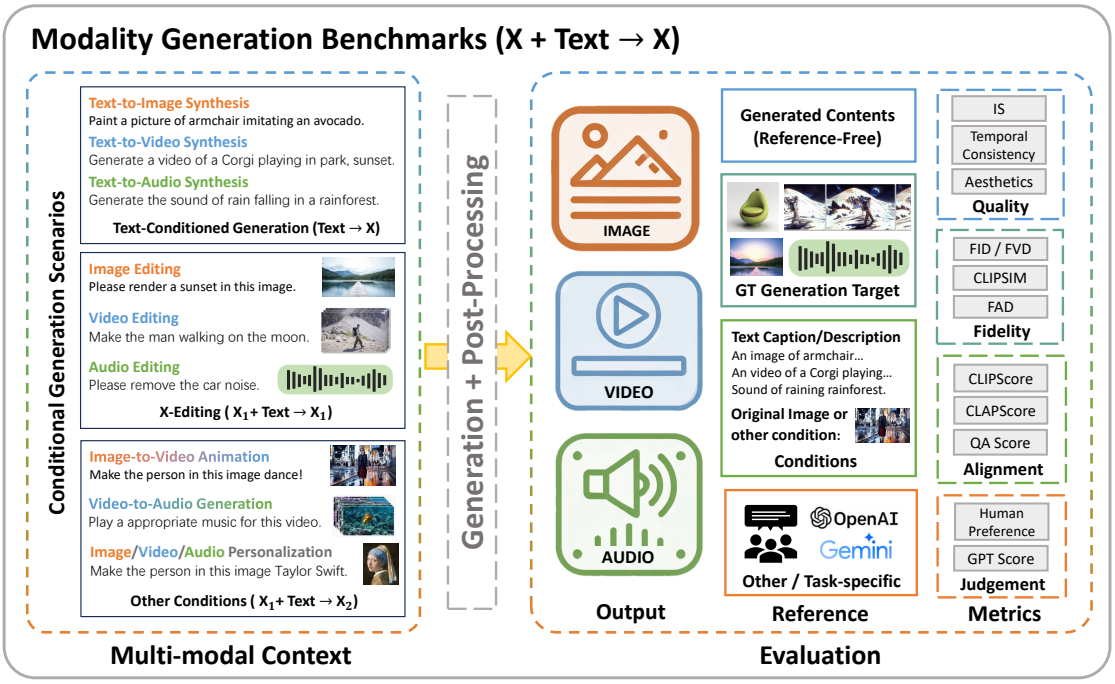

Figure 8. Illustration of the multi-modal generation benchmarks that mainly follow an input-output format of "X + Text → X".

5.2.1. Image Generation

Conditional image synthesis involves generating visuals based on specific conditions or inputs, such as text descriptions or visual prompts, with the aim of creating high-quality, contextually accurate images that fulfill the given conditions. This process encompasses tasks such as text-to-image generation [441–443], image restoration [444,445], and image editing [446,447]. Furthermore, certain purely visual tasks, such as object detection, depth estimation, and segmentation, can also be framed within the context of image generation [448–450].

For the synthesized images, evaluations are conducted based on various forms of reference information. (1) When the target image is used as a reference, the evaluation can be carried out by measuring the differences between the generated and real images [25,451], often quantified by metrics such as FID [452], KID [453], SSIM [454], and CLIP-I [66]. MJHQ-30K [455] employs FID to evaluate the generation capability against a dataset of 30K high-quality images. Meanwhile, leveraging GPT-4V's visual perception capabilities, VIESCORE [456] assesses the semantic alignment between reference and generated images, evaluating their adherence to input instructions. Similarly, in DREAMBENCH++[457], GPT-4o is prompted with the task definition, target image, and output image, then assigns a final score based on the alignment. (2) In cases where text serves as the reference, the alignment between generated images and the textual descriptions is measured through metrics like CLIPScore[458] and BLIP-CLIP [48]. Additionally, GenAI-Bench [459] employs VQAScore by inputting both images and formatted text questions into a VQA model, calculating the probability of the model predicting a "yes" answer. As widely used comprehensive benchmarks for image generation, T2I-Compbench [460] further utilizes MiniGPT-4 [117] with Chain-of-Thought reasoning for evaluating semantic alignment, and GENEval [461] utilizes an object detection model to identify objects in the image, then processes each bounding box through a classification model to determine the corresponding category, which is subsequently matched with the original text description. (3) When no explicit reference is available, human evaluation is often employed to assess the intrinsic diversity and clarity of synthetic images [462–464]. However, to mitigate the high cost of human evaluation, automated metrics like the Inception Score (IS) [465], Aesthetic Predictor's Score (AP) are introduced for independent quality assessment. To further align with human preferences, some approaches leverage reward-model-based methods [466–468], while others automate the process

 doi:10.20944/preprints202411.0685.v1

using LMMs [469–471]. DPG-Bench employs mPLUG-large [472] as its adjudicator, evaluating the generated images based on specific questions and calculating scores in accordance with DSG [473]. (4) With reference information for specific dimensions such as segmentation maps, bounding boxes, key points, and depth maps, task-specific evaluations are also applicable. This process involves extracting the relevant information from the generated image and comparing it to the reference. Specifically, metrics such as mIoU [474], cIoU [475], and gIoU [474] are calculated in segmentation tasks [476–478]. Similarly, for detection tasks, including object detection and keypoint detection, AP and mAP serve as critical performance metrics [385,450,479]. In depth estimation, discrepancies are quantified using absolute error metrics [480,481], including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

5.2.2. Video Generation

Similar to image generation, video generation can also be categorized into various tasks based on different generation conditions, including the commonly used text-to-video [225,428,482,483], image-to-video [484,485], and video editing [486] tasks. In addition, videos further enable temporal-related generation tasks such as future prediction, frame interpolation, and video looping.

The generated videos are evaluated from various perspectives with respect to different references. Since videos can be regarded as sequences of multiple frames, previously introduced image-level metrics, such as FID, IS, PSNR, SSIM, and CLIPSIM, can be used to assess the quality of video frames. Additionally, researchers have developed several video-level metrics. (1) Without considering the reference, the focus is on the perceptual quality of the video itself. With the help of video encoders like C3D [487], IS can be improved to Video Inception Score [488]. EvalCrafter [489] and VBench [490] respectively utilize the Dover [491] model and the LAION aesthetic predictor to assess video quality in terms of aesthetics. Similarly, Wu *et al.* mix several experts with different biases towards a comprehensive quality score. Meanwhile, temporal consistency is another crucial aspect of videos, which is characterized at the pixel level [489,493,494] and the semantic level [495,496]. Furthermore, several benchmarks [489,490,497,498] provide a more systematic assessment of temporal consistency, including aspects from the perspective of subjects, backgrounds, actions, and dynamics. (2) Given ground-truth target videos as references, researchers design FVD [499], adapting FID to video scenarios by utilizing the I3D video encoder [500]. There also exist several variants such as KVD [501] and UMT-FVD [483]. (3) When evaluating conditional generation, it is necessary to assess the alignment between conditions and generated videos. To measure the video-text alignment, CLIPScore [458] is widely used, while FETV [483] explores several variants of CLIPScore. EvalCrafter [489] additionally designs SD-Score and BLIP-BLEU. For fine-grained consistency evaluation, VBench [337] and EvalCrafter [489] utilize various tools for multiple-perspective assessment, whereas other researchers [483,492] build a series of QA pairs based on text descriptions for evaluation. Specifically, for edit tasks, Frame-ACC [496] can be used to determine whether effective editing has been made. Regarding the image condition, image-video conformity can be measure by PIC [502] (4) Although automated metrics are efficient, they may lose important information due to the complicated contents contained in videos, human evaluation is widely adopted for video generation evaluation [483,485,496,502]. Recently, GPT-4o is employed to reduce the need for manual intervention [498].

5.2.3. Audio Generation

Audio generation tasks involve crafting audio content either from text descriptions or existing audio recordings. This includes text-to-audio synthesis, wherein audio is generated based on textual prompts [503–505], as well as audio-to-audio generation, which focuses on manipulating audio by adding, removing, or replacing elements within the tracks [103,164].

Regarding the evaluation on generated audio: (1) When considering ground-truth audio as a reference, subjective evaluation is typically conducted through human scoring, which includes Mean Opinion Scores (MOS) to assess the overall quality of speech. Similarity MOS (SMOS) is used to

 doi:10.20944/preprints202411.0685.v1

evaluate speaker similarity between the speech prompt and the generated output, while Comparative MOS (CMOS) assesses the relative naturalness of synthesized speech compared to the original ground truth audio [506,507]. Given that subjective metrics often require considerable time and workload, various objective metrics are applied to streamline the process, such as FAD [508], KL [509], and CLAP Score [99,510], which can also be utilized for evaluating music generation quality [511–513]. For speech synthesis tasks, Speaker Similarity measures the consistency of timbre between the generated and prompt speech by comparing speaker embeddings, with similarity scores predicted by the speaker verification model WavLM-TDNN [514,515]. Additionally, [516] introduces the TDOA benchmark to assess spatial audio quality by computing TDOA distributions and comparing the generated audio to the ground truth using Mean Absolute Error (MAE). (2) In cases where reference text descriptions are provided, Word Error Rate (WER) and Character Error Rate (CER) are commonly used to evaluate the content accuracy of synthesized speech by calculating the distance between the transcription of the synthesized speech and the input text conditions [12,517,518]. By converting speech into text, ChatGPT Score is employed to evaluate the response quality [10], while models like LLAMA-OMNI [171] and EMOVA [519] utilize GPT-4o to evaluate transcription accuracy and score model responses. Diffsound [520] further adopts a pre-trained audio caption transformer (ACT) [521] to compute a sound-caption-based loss. (3) When only the generated audio is analyzed through discriminate models, besides the fundamental metric like IS [465], Liu *et al.* train a sound classifier to verify sample quality. Moreover, the UTMOS model [523] is specifically designed to predict the Mean Opinion Score (MOS) for speech, allowing for an assessment of its naturalness.

6. Diagnostics: Benchmarks for Hallucination Evaluation

Despite the powerful capabilities LMMs have demonstrated across various scenarios, there may still exists misalignment within the models. Pre-trained LMMs may not fully and accurately understand multi-modal information, leading to the risk of generating incorrect information, also known as hallucinations. Therefore, it is necessary to diagnose these symptoms and analyze the internal mechanisms of LMMs more deeply. In this section, we focus on the misalignment between texts and images, introducing benchmarks and methods utilized to diagnose hallucinations in LVLMs. Corresponding to the hallucination symptoms in description and judgement tasks, current hallucination detection methods can be classified into two major types: (1) evaluating LMMs' capabilities of hallucination discrimination, and (2) assessing the model's ability of non-hallucinatory content generation.

6.1. Evaluation on Hallucination Discrimination

Hallucination discrimination evaluation approach is designed to assess the ability of LMMs to distinguish between accurate and fabricated content. Methods following this approach typically adopt a question-answering format, where LMMs are asked questions based on descriptions that either align with or contradict the content of a given image, and their responses are evaluated accordingly. For instance, POPE [524] employs binary questions about the presence of objects in images to assess the hallucination discrimination capability of LMMs. CIEM [525], similar to POPE [524], automates the object selection process by utilizing ChatGPT for prompting. Another method, NOPE [526], is also VQA-based and specifically designed to evaluate the models' ability to recognize the absence of objects in visual queries, with correct responses being negative statements.

 doi:10.20944/preprints202411.0685.v1

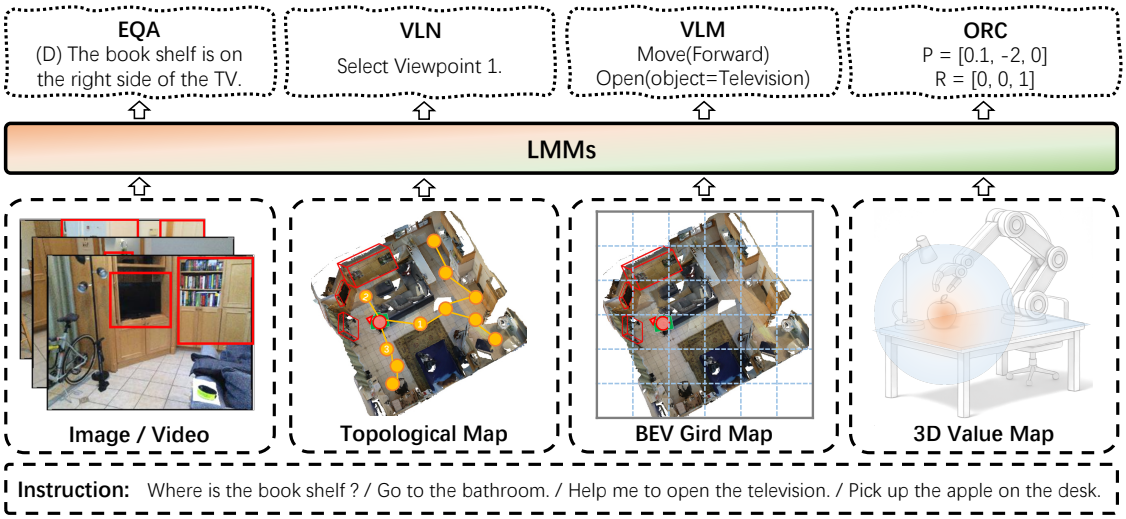

Figure 9. Examples of the input-output space for embodied tasks. Typically, for embodied tasks, the input includes the user instruction, the current observation (image or video), the environment and the history (optional). We omit the history here, as for different tasks, the content of history vary from pure texts to observation sequences, action sequences and/or updated environment representations.

6.2. Evaluation on Hallucination Generation

Evaluating hallucination generation aims to measure the proportion of hallucinated content in the outputs. Currently, there are primarily two main types: rule-based and model-based methods. Hand-crafted rule-based methods are characterized by their strong interpretability, achieved by manually designing multiple evaluation steps with specific and clear objectives. Typical benchmarks include CHAIR [527], CCEval [528], and FAITHSCORE [529]. Additionally, AMBER [530] can be used to evaluate both generative and discriminative tasks through several rule-based metrics, enabling the detection of existence, attribute, and relation hallucinations. Model-based methods directly assess the performance of LMMs by evaluating their responses with the help of intelligent models. Based on the model used, these methods can be categorized into two types: LLM-based evaluation and hallucination-data-driven model evaluation. For LLM-based evaluation, GPT-4 is mainly employed to assess contents generated by LVLMs, focusing on hallucination levels, capitalizing on the robust natural language understanding and processing capabilities of advanced LLMs [360,531]. In practice, LLMs are prompted to evaluate and score these responses by comprehensively considering visual information with dense captions, object bounding boxes, user instructions, and model responses. Likewise, hallucination-data-driven model evaluation methods build labelled hallucination datasets for fine-tuning models to detect hallucinations [532,533].

7. Extension to Embodied Agents

Embodied AI is a rapidly advancing field of research that explores how agents develop intelligence through interaction with environments. This interaction encompasses not only the perception and understanding of the environment, but also the decision-making of future actions [534]. In this section, we will firstly introduce several categories of embodied tasks, then delve into how to adapt LVLMs to embodied tasks by extending the input-output spaces.

7.1. Embodied Tasks

Tasks are referred to as "embodied" because the agent needs to interact with a real or virtual environment. Based on the complexity of the interaction actions, we can categorize embodied tasks and corresponding datasets as follows: (1) ***Embodied Question Answering (EQA)*** [535,536]: In these tasks, the agent is required only to answer user questions. Broadly speaking, we can consider such action spaces as discrete vocabularies. (2) ***Vision-and-Language Navigation (VLN)*** [537,538]: These tasks

 doi:10.20944/preprints202411.0685.v1

involve navigation within an environment based on user instructions. However, this tasks does not require interactions with objects. Therefore, the action space is either discrete directional movements, such as forward, backward, left, and right, or it can involve continuous control parameters, such as speed and direction. ***Graphical User Interface (GUI) Navigation*** [539,540] is a specialized category of VLN tasks where the agent needs to operate on a computer or mobile phone based on user instructions. Action spaces are often discretized screen operations, such as clicking, swiping, and text-editing. (3) ***Vision-and-Language Manipulation (VLM)*** [541–543]: These tasks require the agent to not only engage in question-answer dialogues with the user, but also navigate the environment and interact with objects based on user instructions. This action space builds upon the action space of VLN tasks by adding object manipulation actions. (4) ***Open-World Robot Control (ORC)*** [544–546]: In these tasks, agents are equipped with high-degree-of-freedom robotic arms, capable of performing precise object manipulations, such as grasping and moving objects. Both indoor environments, such as household scenarios, and outdoor settings, such as material transport scenarios, are involved. The action space for ORC tasks is continuous, determined by the complexity of the robotic arm movements.

7.2. Input Extension: Environment Representation

Since embodied agents interact with the environment as the subject, the egocentric observation becomes an essential choice [413,537,542,547,548]. Under egocentric observations, the environment is often represented as a local image [549–551] corresponding to the current orientation or by rotating 360 degrees, which could be satisfactory for EQA tasks. However, VLN and VLM tasks require an integrated understanding of observed environments. When the agent operates from a single egocentric view, its understanding of the environment is inherently partial and localized. To obtain a complete picture, the agent must engage in thorough and repeated exploration of the environment [552,553]. Therefore, the ability to integrate temporal local information and transform it into a long-term global perspective is crucial for embodied agents. Several works utilize topological map [554,555] to record the spatial semantics during navigation, either for obtaining a better visual representation for the environment [556], or for constructing reasoning chains [557]. Others employ bird's-eye-view grid maps to structure the visited environment [558–560]. For ORC tasks, a detailed 3D modeling of the environment is essential for executing precise actions with a robotic arm. For example, VoxPoser [561] take the 3D value map derived from interactions between a LLM and a vision-language model to enable exact and efficient object manipulations.

7.3. Output Extension: Action Representation

As stated in Section 7.1, different embodied tasks have distinct action spaces, necessitating the extensions to model outputs to accommodate the specific demands of each task.

Discrete Action Space

For embodied tasks of VLN and VLM with discrete action spaces, embodied actions are divided into a fixed set of categories. The output of agents select one action category for execution. One line of work, i.e. LLaRP [562], utilizes an additional action prediction module specifically designed to decode discrete actions, which could generalize to unseen tasks better. Another line of work leverages the powerful language decoding capabilities of LLMs. For example, NavGPT [563] and NaviLLM [564] predict actions as plain-text, which is then parsed into specific action commands. This design simplifies the complexity of action decision, but limits the granularity of actions for complex operations like robotic arm control in ORC tasks. To mitigate this issue, RT-2 [565] introduce special action tokens into the vocabulary. These discrete tokens are then de-tokenized into continuous control signals of robot action.

 doi:10.20944/preprints202411.0685.v1

Continuous Action Space

To better adapt to ORC tasks, the extension to continuous actions is necessary. A continuous action space is represented by a set of continuous values, such as the joint angles or velocities of a robotic arm, allowing the agent to move or adjust freely within the control space. Since the direct outputs of LVLMs are discrete tokens, decoding continuous actions typically requires an extra action decoding head. RoboFlamingo [566] experiments with different action decoding head architectures (e.g., MLP, RNN, and Transformer) to enable language-conditioned robotic control. Octo [567] employs a modular framework, integrating diffusion model-based action policies to predict continuous actions. Unlike RoboFlamingo, the advantage of Octo lies in its ability to flexibly connect different task encoders, observation encoders, and action decoders, making it highly adaptable.

Hierarchical Action Space

Hierarchical action space separates the level of action control into high-level task planning and low-level control policies (could be either discrete or continuous), each handled by separate modules or models. Specifically, PALM-E [551] and LEO [568] uses high-level instructions generated by LVLMs to guide low-level control policies in executing specific actions. LEO [568] further enhances its understanding on 3D world by utilizing a 3D encoder and crafting large-scale datasets for training.

7.4. Multi-Modal Alignment

Input-level Alignment

To bridge the gap between the newly introduced environment representation and other modalities, SMNet [554], GridMM [560] and Trans4Map [558] employ end-to-end imitation learning, continuously adjusting the model parameters to optimize allocentric map generation and updating processes. However, these obtained map representations are highly dependent on the UNet and GRU modules nested within the model architecture, lacking the ability to transfer between different language backbones. To address this issue, Ego^2-Map [556] takes a self-supervised contrastive learning strategy, comparing egocentric view features with their corresponding semantic maps. Such representations encompass rich spatial information from a map, exhibiting strong generalizable capability on various environments for both high-level and low-level action spaces.

Output-level Alignment

Adapting the outputs to different action spaces is essential for agents to understand and execute complex tasks. There are two major strategies: (1) *Direct Alignment*: This approach maps instructions directly to executable actions in an end-to-end manner, as exemplified by RoboFlamingo [566] and Octo [567]. RoboFlamingo uses an MLP-based action decoder to convert hidden states into specific control signals, while Octo employs a conditional diffusion decoder as the action policy module. This decoder predicts continuous action distributions, transforming Gaussian noise into desired action outputs through a series of denoising steps. During training, both RoboFlamingo and Octo collect sequential actions covering various scenarios and tasks, enhancing the model's generalization capability during pre-training. They also allow the policy module to be fine-tuned with a small amount of trajectory data so as to quickly adapt to new tasks. (2) *Indirect Alignment*: This method breaks down user instructions into language plans that can be understood by downstream models, with representative works as PALM-E [551] and LEO [568]. PALM-E pre-trains on large datasets of robotic manipulation planning, visual question answering and captioning, converting complex environmental perceptions into multi-step task planning. It integrates the task plans with SayCan [569] for specific action execution. While LEO adopts a two-stage training process involving 3D vision-language alignment and fine-tuning on 3D vision-language-action instructions, enhancing the agent's adaptability to different action spaces.

 doi:10.20944/preprints202411.0685.v1

7.5. Evaluation

Task-Specific Benchmarks

Different embodied AI tasks involve distinct scenarios, each with specific evaluation datasets. These datasets fall under the category of modality comprehension datasets, as they involve interpreting multiple types of input, such as vision, language, and history actions. Based on the task categories in Section 7.1, task-specific datasets are: (1) For EQA tasks, representative datasets include EQA [535], IQUAD [536], and SQA3D [570]. (2) For VLN tasks, notable benchmarks include MultiON [571], R2R [537], R2R-CE [538], SOON [572], and REVERIE [548]. For GUI navigation tasks, relevant datasets involve MinoWeb++ [539], Mind2Web [573], AITZ [574], and GUI-Odyssey [575]. Specifically, there is a bilingual evaluation benchmark FunUI [576] to access the basic UI understandings of GUI agents. (3) For VLM tasks, representative datasets include ALFRED [541], TEACH [542], and HomeRobot OVMM [543]. (4) For ORC tasks, notable datasets include Franka Kitchen [544], Open X-Embodiment [546] and CALVIN [545]. Generally, for EQA tasks, the primary evaluation metric is accuracy, while for action prediction, i.e. VLN, VLM an ORC tasks, both single-step action prediction accuracy and episodic task execution success rate are used as evaluation metrics.

Comprehensive Benchmarks

Comprehensive evaluation of agent performance across various embodied scenarios is challenging. To address this, Visual-AgentBench (VAB) [577] has been developed as a pioneering benchmark specifically designed to access LMMs as visually-grounded agents. VAB covers a wide range of environments, including embodied, graphical user interfaces, and visual design. It evaluates the understanding and interaction capabilities of LMMs through multi-task assessments and offers a hybrid trajectory training set built using methods such as program solvers, LMM agent bootstrapping, and human demonstrations to enhance the performance of LMMs in behavior cloning. Success rate is adopted as the evaluation metric.

8. Discussion and Outlook

8.0.1. How to construct multi-modal input-output spaces with discretely or continuously encoded modality signal?

Currently, mainstream LMMs follow the hybrid structure, where modality signals are continuously encoded and integrated into the text space. This method is simple yet effective, leveraging encoders like CLIP [66] and CLAP [99], which are aligned with text through large-scale pre-training, to achieve impressive performance in comprehension tasks. However, this approach introduce additional design costs for corresponding alignment modules for the input and output ends.

Meanwhile, hybrid input spaces cannot directly support multi-modal content generation. This necessitates the design of more complex output layers and decoding strategies for LMMs with multi-modal generation capabilities, leading to a significant gap between the input and output spaces.

On the other hand, the unified discrete space structure is more straightforward, supporting both comprehension and generation tasks through a unified approach (e.g., next-token prediction). However, they are currently limited by the absence of strong discrete encoders across various modalities, akin to CLIP, resulting in slightly weaker performance on comprehension tasks compared to hybrid models. Ovis [153], however, has shown that by carefully designing and expanding the visual vocabulary, discrete models can also perform well on comprehension tasks. Additionally, due to the competitive relationship between modalities, improving training stability is also a challenge that needs to be addressed for unified discrete representation models.

In conclusion, both approaches have their strengths and weaknesses, with significant room for optimization. At the same time, we believe that the current training strategies of discrete and continuous encoders are not mutually exclusive, the development and approaches of both methods

 doi:10.20944/preprints202411.0685.v1

can learn from each other. The research community eagerly anticipates an effective modality encoding method that unifies understanding and generation.

Furthermore, there is a noticeable granularity gap between textual and modal representations, whether the modality signals are encoded continuously or discretely. Text tokens carry explicit semantics, while individual modality tokens might only contain limited information. A single text token may correspond to multiple tokens in an image, leading to excessively long token sequences for modality signals in current LMMs. In the future, can we build modality representations that carry semantics at specific levels?

8.0.2. How to design model architectures and training strategies to align the constructed multi-modal space?

The architectures should to be designed based on the input and output space. Most LMMs are built on a backbone, usually initialized from a pre-trained LLM to gain better text understanding capabilities and initial representations. For hybrid spaces, additional design is required for input and output alignment modules. Although the LLM backbone can perform unified multi-modal modeling through training, relatively complex internal alignment modules can be introduced to model complex cross-modal interactions.

As introduced in Section 4.1, there is a variety of designs for each module, with different structures having trade-offs across various dimensions. No structure consistently performs better across different scenarios and requirements.

Regarding training strategies, most models undergo two stages: pre-training and instruction fine-tuning. The former aligns modal representations, while the latter enhances instruction-following capabilities in multi-modal scenarios. The scale and quality of training data are critical. Ensuring the data with a wide coverage and high-quality information is the key to effectively improving LMM's performance as the data scale up. During this process, it is necessary to select appropriate parameter settings.

In summary, the design of current multi-modal alignment frameworks—including structure, data, and training parameters—remains a research problem with a vast feasible space. Since training LMMs requires extensive computation, empirical experiments demand significant computational resources. Finding ways to quickly validate the effectiveness of an optimization direction is essential. Additionally, there have already been relevant explorations to provide some general conclusions [144,152], helping researchers offer heuristic approaches to narrow down the model design space.

8.0.3. How to comprehensively evaluate LMMs based on the expanded input-output space?

Based on the extended input space, text instructions can be leveraged as the interface for evaluating the LMM's capabilities of understanding multi-modal information through X-to-Text tasks. Specifically, depending on the task, language can be used to describe the corresponding requirements, guiding the model to respond to the questions in the desired form. Evaluation data can be constructed based on different scenarios. Combining data from multiple scenarios leads to a comprehensive evaluation of the model.

With output extension, the model can perform generation tasks in the target modality based on different multimodal contexts, such as text-to-image/video/audio generation and image/video/audio editing tasks. The generated images, videos, and audio are further evaluated according to different metrics with respect to the reference information.

Most current evaluation benchmarks aim to assess zero-shot capabilities, aligning with real-world application scenarios. Additionally, another effective way to define tasks is in-context learning (ICL), where examples can more effectively convey task requirements when it is difficult to describe them in text. However, this method has yet to be fully explored in multi-modal scenarios. At the same time, some studies have found a mutually exclusive phenomenon between zero-shot and ICL capabilities in LMMs [144].

 doi:10.20944/preprints202411.0685.v1

Furthermore, two major limitations of current benchmarks are: (1) the gap between the evaluation tasks and realistic queries from users; and (2) the misalignment between the scores evaluated and human preferences from the real world, while extensive manual evaluation introduces higher costs. Defining appropriate questions and metrics can prevent LMMs from overfitting to certain scores. Moreover, it is also crucial to reveal the risks of LMMs in practical applications.

8.0.4. A promising way towards world models.

As demonstrated in Section 7, the perspective of expanding the input-output space is not limited to modalities. Any form of information and signals can be considered. By encoding them into the input-output space and aligning them through the design of model architecture and training strategies, the trained models an be applied and evaluated in downstream tasks. We believe that this framework further reveals the possibility of building models capable of understanding the physical world. The claim, "predicting the next token is to understand the world", could be validated with a premise that the defined token space has been expanded to cover a sufficient amount of information and signals from the world.

9. Conclusion

In this paper, we summarize the current methods of large multi-modal model (LMM) construction from the perspective of input-output space extension. We further break down and provide detailed discussion of the key research problems in the construction process, including the structure of multi-modal input and output spaces, multi-modal representation alignment frameworks, and comprehensive evaluation of generated multi-modal content. Our summarizing framework is not only straightforward but also effectively encapsulates the mainstream approaches while offering potential for further extension. This paper will continue to be updated, and we hope it can provide an intuitive and comprehensive overview for related researchers and inspire future work.

References

1. OpenAI. ChatGPT (August 3 version), 2023.
2. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; others. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288* **2023**.
3. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; others. Qwen technical report. *arXiv preprint arXiv:2309.16609* **2023**.
4. AI@Meta. Llama 3 Model Card **2024**.
5. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597* **2023**.
6. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2024**, *36*.
7. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* **2023**.
8. Ma, Z.; Yang, G.; Yang, Y.; Gao, Z.; Wang, J.; Du, Z.; Yu, F.; Chen, Q.; Zheng, S.; Zhang, S.; others. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. *arXiv preprint arXiv:2402.08846* **2024**.
9. Koh, J.Y.; Fried, D.; Salakhutdinov, R.R. Generating images with multimodal language models. *Advances in Neural Information Processing Systems* **2024**, *36*.
10. Zhang, D.; Zhang, X.; Zhan, J.; Li, S.; Zhou, Y.; Qiu, X. SpeechGPT-Gen: Scaling Chain-of-Information Speech Generation. *arXiv preprint arXiv:2401.13527* **2024**.
11. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T.S. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519* **2023**.
12. Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; Yan, H.; Fu, J.; Gui, T.; Sun, T.; Jiang, Y.; Qiu, X. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling, 2024, [arXiv:cs.CL/2402.12226].

13. Wu, J.; Gan, W.; Chen, Z.; Wan, S.; Philip, S.Y. Multimodal large language models: A survey. 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023, pp. 2247–2256.
14. Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; Cucchiara, R. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451* **2024**.
15. Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; others. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432* **2023**.
16. Latif, S.; Shoukat, M.; Shamshad, F.; Usama, M.; Ren, Y.; Cuayáhuitl, H.; Wang, W.; Zhang, X.; Togneri, R.; Cambria, E.; others. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792* **2023**.
17. Xiao, H.; Zhou, F.; Liu, X.; Liu, T.; Li, Z.; Liu, X.; Huang, X. A comprehensive survey of large language models and multimodal large language models in medicine. *arXiv preprint arXiv:2405.08603* **2024**.
18. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.D.; others. A survey on multimodal large language models for autonomous driving. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.
19. Li, J.; Lu, W. A Survey on Benchmarks of Multimodal Large Language Models. *arXiv preprint arXiv:2408.08632* **2024**.
20. Huang, J.; Zhang, J. A Survey on Evaluation of Multimodal Large Language Models. *arXiv preprint arXiv:2408.15769* **2024**.
21. Bai, T.; Liang, H.; Wan, B.; Yang, L.; Li, B.; Wang, Y.; Cui, B.; He, C.; Yuan, B.; Zhang, W. A Survey of Multimodal Large Language Model from A Data-centric Perspective. *arXiv preprint arXiv:2405.16640* **2024**.
22. Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818* **2024**.
23. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6904–6913.
24. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6700–6709.
25. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. ECCV. Springer, 2014, pp. 740–755.
26. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* **2014**, *2*, 67–78.
27. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel object captioning at scale. ICCV, 2019, pp. 8948–8957.
28. Sidorov, O.; Hu, R.; Rohrbach, M.; Singh, A. Textcaps: a dataset for image captioning with reading comprehension. ECCV. Springer, 2020, pp. 742–758.
29. Kazemzadeh, S.; Ordonez, V.; Matten, M.; Berg, T. Referitgame: Referring to objects in photographs of natural scenes. EMNLP, 2014, pp. 787–798.
30. Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2901–2910.
31. Suhr, A.; Lewis, M.; Yeh, J.; Artzi, Y. A Corpus of Natural Language for Visual Reasoning. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Barzilay, R.; Kan, M.Y., Eds.; Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 217–223. doi:10.18653/v1/P17-2034.
32. Xie, N.; Lai, F.; Doran, D.; Kadav, A. Visual entailment: A novel task for fine-grained image understanding. *arXiv:1901.06706* **2019**.
33. Zellers, R.; Bisk, Y.; Farhadi, A.; Choi, Y. From recognition to cognition: Visual commonsense reasoning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6720–6731.
34. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning; Bach, F.; Blei, D., Eds.; PMLR: Lille, France, 2015; Vol. 37, *Proceedings of Machine Learning Research*, pp. 2048–2057.

35. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* **2017**.
36. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
37. Li, Z.; Wei, Z.; Fan, Z.; Shan, H.; Huang, X. An Unsupervised Sampling Approach for Image-Sentence Matching Using Document-level Structural Information. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *35*, 13324–13332. doi:10.1609/aaai.v35i15.17573.
38. Wang, R.; Wei, Z.; Li, P.; Zhang, Q.; Huang, X. Storytelling from an Image Stream Using Scene Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence* **2020**, *34*, 9185–9192. doi:10.1609/aaai.v34i05.6455.
39. Nagaraja, V.K.; Morariu, V.I.; Davis, L.S. Modeling context between objects for referring expression understanding. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, 2016, pp. 792–807.
40. Yue, S.; Tu, Y.; Li, L.; Yang, Y.; Gao, S.; Yu, Z. I3n: Intra-and inter-representation interaction network for change captioning. *IEEE Transactions on Multimedia* **2023**, *25*, 8828–8841.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, [arXiv:cs.CL/1810.04805].
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.
43. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. ECCV. Springer, 2020, pp. 104–120.
44. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-training of Generic Visual-Linguistic Representations, 2020, [arXiv:cs.CV/1908.08530].
45. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. Advances in Neural Information Processing Systems; Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; Vaughan, J.W., Eds. Curran Associates, Inc., 2021, Vol. 34, pp. 9694–9705.
46. Li, Z.; Fan, Z.; Tou, H.; Chen, J.; Wei, Z.; Huang, X. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4395–4405.
47. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; Gao, J. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX; Springer-Verlag: Berlin, Heidelberg, 2020; p. 121–137. doi:10.1007/978-3-030-58577-8_8.
48. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. ICML. PMLR, 2022, pp. 12888–12900.
49. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners are Image-Text Foundation Models, 2022, [arXiv:cs.CV/2205.01917].
50. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision, 2022, [arXiv:cs.CV/2108.10904].
51. Li, Z.; Fan, Z.; Chen, J.; Zhang, Q.; Huang, X.; Wei, Z. Unifying Cross-Lingual and Cross-Modal Modeling Towards Weakly Supervised Multilingual Vision-Language Pre-training. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds.; Association for Computational Linguistics: Toronto, Canada, 2023; pp. 5939–5958. doi:10.18653/v1/2023.acl-long.327.
52. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* **2021**.
53. OpenAI. GPT-4 Technical Report. *arXiv:2303.08774* **2023**.
54. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* **2015**.

55. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; others. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
56. Wu, Y. Google's Neural Machine Translation System: Bridging the Gap between human and machine translation. *arXiv preprint arXiv:1609.08144* **2016**.
57. Kudo, T. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959* **2018**.
58. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
59. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. arXiv e-prints, 2022.
60. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
61. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
62. Van Den Oord, A.; Vinyals, O.; others. Neural discrete representation learning. *Advances in neural information processing systems* **2017**, *30*.
63. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.
64. Bavishi, R.; Elsen, E.; Hawthorne, C.; Nye, M.; Odena, A.; Somani, A.; Taşırlar, S. Fuyu-8B, 2023.
65. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021, [arXiv:cs.CV/2010.11929].
66. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others. Learning transferable visual models from natural language supervision. ICML. PMLR, 2021, pp. 8748–8763.
67. Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389* **2023**.
68. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid loss for language image pre-training. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11975–11986.
69. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; others. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
70. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.
71. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
72. Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832* **2021**.
73. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; others. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* **2023**.
74. Ge, Y.; Ge, Y.; Zeng, Z.; Wang, X.; Shan, Y. Planting a SEED of Vision in Large Language Model, 2023, [arXiv:cs.CV/2307.08041].
75. Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Luo, Z.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; Wang, X. Generative Multimodal Models are In-Context Learners, 2024, [arXiv:cs.CV/2312.13286].
76. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592* **2023**.

77. Yuan, Y.; Li, W.; Liu, J.; Tang, D.; Luo, X.; Qin, C.; Zhang, L.; Zhu, J. Osprey: Pixel understanding with visual instruction tuning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 28202–28211.
78. Ge, C.; Cheng, S.; Wang, Z.; Yuan, J.; Gao, Y.; Song, J.; Song, S.; Huang, G.; Zheng, B. ConvLLaVA: Hierarchical Backbones as Visual Encoder for Large Multimodal Models. *arXiv preprint arXiv:2405.15738* **2024**.
79. Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Xu, G.; Li, C.; Tian, J.; Qian, Q.; Zhang, J.; others. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126* **2023**.
80. Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; Bai, X. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607* **2023**.
81. Gao, P.; Zhang, R.; Liu, C.; Qiu, L.; Huang, S.; Lin, W.; Zhao, S.; Geng, S.; Lin, Z.; Jin, P.; others. SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.05935* **2024**.
82. Xu, R.; Yao, Y.; Guo, Z.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.S.; Liu, Z.; Sun, M.; Huang, G. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. *ArXiv* **2024**, *abs/2403.11703*.
83. Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; Lee, Y.J. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.
84. Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Sun, Y.; others. Deepseek-vl: Towards real-world vision-language understanding. *arXiv:2403.05525* **2024**.
85. Zhao, H.; Zhang, M.; Zhao, W.; Ding, P.; Huang, S.; Wang, D. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference **2024**.
86. Hong, W.; Wang, W.; Lv, Q.; Xu, J.; Yu, W.; Ji, J.; Wang, Y.; Wang, Z.; Dong, Y.; Ding, M.; others. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914* **2023**.
87. Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; Jia, J. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814* **2024**.
88. Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S.C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; others. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860* **2024**.
89. Fan, X.; Ji, T.; Jiang, C.; Li, S.; Jin, S.; Song, S.; Wang, J.; Hong, B.; Chen, L.; Zheng, G.; others. MouSi: Poly-Visual-Expert Vision-Language Models. *arXiv preprint arXiv:2401.17221* **2024**.
90. Luo, R.; Zhao, Z.; Yang, M.; Dong, J.; Qiu, M.; Lu, P.; Wang, T.; Wei, Z. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207* **2023**.
91. Zhang, H.; Li, X.; Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* **2023**.
92. Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; Liu, Z. Otter: A Multi-Modal Model with In-Context Instruction Tuning, 2023, [arXiv:cs.CV/2305.03726].
93. Yu, Y.Q.; Liao, M.; Zhang, J.; Wu, J. TextHawk2: A Large Vision-Language Model Excels in Bilingual OCR and Grounding with 16x Fewer Tokens. *arXiv preprint arXiv:2410.05261* **2024**.
94. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? ICML, 2021, Vol. 2, p. 4.
95. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3202–3211.
96. Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; Qiao, Y. Videochat: Chat-centric video understanding. *arXiv:2305.06355* **2023**.
97. Xu, H.; Ye, Q.; Wu, X.; Yan, M.; Miao, Y.; Ye, J.; Xu, G.; Hu, A.; Shi, Y.; Xu, G.; others. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362* **2023**.
98. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing* **2021**, *29*, 3451–3460.

99. Elizalde, B.; Deshmukh, S.; Al Ismail, M.; Wang, H. Clap learning audio concepts from natural language supervision. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
100. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. Imagebind: One embedding space to bind them all. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15180–15190.
101. Zhang, X.; Zhang, D.; Li, S.; Zhou, Y.; Qiu, X. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692* **2023**.
102. Han, J.; Zhang, R.; Shao, W.; Gao, P.; Xu, P.; Xiao, H.; Zhang, K.; Liu, C.; Wen, S.; Guo, Z.; others. ImageBind-LLM: Multi-modality Instruction Tuning. *arXiv:2309.03905* **2023**.
103. Tang, Z.; Yang, Z.; Khademi, M.; Liu, Y.; Zhu, C.; Bansal, M. CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation, 2023, [arXiv:cs.CV/2311.18775].
104. Lu, J.; Clark, C.; Lee, S.; Zhang, Z.; Khosla, S.; Marten, R.; Hoiem, D.; Kembhavi, A. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action, 2023, [arXiv:cs.CV/2312.17172].
105. Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* **2023**.
106. Zhang, P.; Wang, X.; Cao, Y.; Xu, C.; Ouyang, L.; Zhao, Z.; Ding, S.; Zhang, S.; Duan, H.; Yan, H.; Zhang, X.; Li, W.; Li, J.; Chen, K.; He, C.; Zhang, X.; Qiao, Y.; Lin, D.; Wang, J. InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition. *ArXiv* **2023**, *abs/2309.15112*.
107. Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; others. InternLM-XComposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420* **2024**.
108. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis With Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
109. Dong, R.; Han, C.; Peng, Y.; Qi, Z.; Ge, Z.; Yang, J.; Zhao, L.; Sun, J.; Zhou, H.; Wei, H.; Kong, X.; Zhang, X.; Ma, K.; Yi, L. DreamLLM: Synergistic Multimodal Comprehension and Creation, 2024, [arXiv:cs.CV/2309.11499].
110. Zheng, K.; He, X.; Wang, X.E. MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens, 2024, [arXiv:cs.CV/2310.02239].
111. Sun, Q.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; Wang, Y.; Gao, H.; Liu, J.; Huang, T.; Wang, X. Emu: Generative Pretraining in Multimodality, 2024, [arXiv:cs.CV/2307.05222].
112. Ge, Y.; Zhao, S.; Zeng, Z.; Ge, Y.; Li, C.; Wang, X.; Shan, Y. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218* **2023**.
113. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023, [arXiv:cs.CV/2305.06500].
114. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Muyan, Z.; Zhang, Q.; Zhu, X.; Lu, L.; others. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238* **2023**.
115. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; others. Flamingo: a visual language model for few-shot learning. *NIPS* **2022**, *35*, 23716–23736.
116. Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199* **2023**.
117. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592* **2023**.
118. Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; others. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178* **2023**.
119. Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; others. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010* **2023**.
120. Luo, G.; Zhou, Y.; Ren, T.; Chen, S.; Sun, X.; Ji, R. Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models. *arXiv:2305.15023* **2023**.

121. Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790* **2023**.
122. Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; Zhao, R. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv:2306.15195* **2023**.
123. Maaz, M.; Rasheed, H.; Khan, S.; Khan, F.S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* **2023**.
124. Zeng, Y.; Zhang, H.; Zheng, J.; Xia, J.; Wei, G.; Wei, Y.; Zhang, Y.; Kong, T. What Matters in Training a GPT4-Style Language Model with Multimodal Inputs? *arXiv:2307.02469* **2023**.
125. Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; Tu, Z. BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions. *arXiv:2308.09936* **2023**.
126. IDEFICS. Introducing IDEFICS: An Open Reproduction of State-of-the-Art Visual Language Model. https://huggingface.co/blog/idefics, 2023.
127. Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; others. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* **2023**.
128. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv:2310.03744* **2023**.
129. Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; Elhoseiny, M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478* **2023**.
130. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; others. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* **2023**.
131. Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; Lin, D. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793* **2023**.
132. Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; Zhou, J. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. *ArXiv* **2023**, *abs/2311.04257*.
133. Lin, Z.; Liu, C.; Zhang, R.; Gao, P.; Qiu, L.; Xiao, H.; Qiu, H.; Lin, C.; Shao, W.; Chen, K.; others. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575* **2023**.
134. Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; others. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886* **2023**.
135. Lin, J.; Yin, H.; Ping, W.; Lu, Y.; Molchanov, P.; Tao, A.; Mao, H.; Kautz, J.; Shoeybi, M.; Han, S. VILA: On Pre-training for Visual Language Models, 2024, [arXiv:cs.CV/2312.07533].
136. Cha, J.; Kang, W.; Mun, J.; Roh, B. Honeybee: Locality-enhanced projector for multimodal llm. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13817–13827.
137. Wu, J.; Hu, X.; Wang, Y.; Pang, B.; Soricut, R. Omni-SMoLA: Boosting Generalist Multimodal Models with Soft Mixture of Low-rank Experts. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14205–14215.
138. Chen, S.; Jie, Z.; Ma, L. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160* **2024**.
139. Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Zhang, J.; Ning, M.; Yuan, L. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947* **2024**.
140. Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; Sun, X.; Hu, Y.; Lin, X.; Zhang, B.; others. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *arXiv preprint arXiv:2402.03766* **2024**.
141. He, M.; Liu, Y.; Wu, B.; Yuan, J.; Wang, Y.; Huang, T.; Zhao, B. Efficient Multimodal Learning from Data-centric Perspective. *arXiv preprint arXiv:2402.11530* **2024**.
142. Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; Huang, L. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289* **2024**.
143. Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; others. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652* **2024**.

144. McKinzie, B.; Gan, Z.; Fauconnier, J.P.; Dodge, S.; Zhang, B.; Dufter, P.; Shah, D.; Du, X.; Peng, F.; Weers, F.; others. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611* **2024**.
145. Qiao, Y.; Yu, Z.; Guo, L.; Chen, S.; Zhao, Z.; Sun, M.; Wu, Q.; Liu, J. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600* **2024**.
146. Zhao, H.; Zhang, M.; Zhao, W.; Ding, P.; Huang, S.; Wang, D. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520* **2024**.
147. Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; others. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821* **2024**.
148. Abdin, M.; Jacobs, S.A.; Awan, A.A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; others. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* **2024**.
149. Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S.K.; Feng, J. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994* **2024**.
150. Yu, Y.; Liao, M.; Wu, J.; Liao, Y.; Zheng, X.; Zeng, W. TextHawk: Exploring Efficient Fine-Grained Perception of Multimodal Large Language Models. *CoRR* **2024**, *abs/2404.09204*.
151. Shao, Z.; Yu, Z.; Yu, J.; Ouyang, X.; Zheng, L.; Gai, Z.; Wang, M.; Ding, J. Imp: Highly Capable Large Multimodal Models for Mobile Devices. *arXiv preprint arXiv:2405.12107* **2024**.
152. Laurençon, H.; Tronchon, L.; Cord, M.; Sanh, V. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246* **2024**.
153. Lu, S.; Li, Y.; Chen, Q.G.; Xu, Z.; Luo, W.; Zhang, K.; Ye, H.J. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv preprint arXiv:2405.20797* **2024**.
154. Yao, L.; Li, L.; Ren, S.; Wang, L.; Liu, Y.; Sun, X.; Hou, L. DeCo: Decoupling Token Compression from Semantic Abstraction in Multimodal Large Language Models. *arXiv preprint arXiv:2405.20985* **2024**.
155. Li, J.; Wang, X.; Zhu, S.; Kuo, C.W.; Xu, L.; Chen, F.; Jain, J.; Shi, H.; Wen, L. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv preprint arXiv:2405.05949* **2024**.
156. Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; others. CogVLM2: Visual Language Models for Image and Video Understanding. *arXiv preprint arXiv:2408.16500* **2024**.
157. Zhang, P.; Dong, X.; Zang, Y.; Cao, Y.; Qian, R.; Chen, L.; Guo, Q.; Duan, H.; Wang, B.; Ouyang, L.; Zhang, S.; Zhang, W.; Li, Y.; Gao, Y.; Sun, P.; Zhang, X.; Li, W.; Li, J.; Wang, W.; Yan, H.; He, C.; Zhang, X.; Chen, K.; Dai, J.; Qiao, Y.; Lin, D.; Wang, J. InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output, 2024, [arXiv:cs.CV/2407.03320].
158. Laurençon, H.; Marafioti, A.; Sanh, V.; Tronchon, L. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637* **2024**.
159. Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; Zhou, J. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. *arXiv preprint arXiv:2408.04840* **2024**.
160. Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; Li, C. LLaVA-OneVision: Easy Visual Task Transfer, 2024, [arXiv:cs.CV/2408.03326].
161. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; others. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* **2024**.
162. Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; Cai, D. Pandagpt: One model to instruction-follow them all. *arXiv:2305.16355* **2023**.
163. Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; Zhang, M. Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts. *arXiv preprint arXiv:2405.11273* **2024**.
164. Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; Qiu, X. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000* **2023**.
165. Wu, J.; Gaur, Y.; Chen, Z.; Zhou, L.; Zhu, Y.; Wang, T.; Li, J.; Liu, S.; Ren, B.; Liu, L.; others. On decoder-only architecture for speech-to-text and large language model integration. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023, pp. 1–8.

166. Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; Zhang, C. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289* **2023**.
167. Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919* **2023**.
168. Hu, S.; Zhou, L.; Liu, S.; Chen, S.; Hao, H.; Pan, J.; Liu, X.; Li, J.; Sivasankaran, S.; Liu, L.; others. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656* **2024**.
169. Das, N.; Dingliwal, S.; Ronanki, S.; Paturi, R.; Huang, D.; Mathur, P.; Yuan, J.; Bekal, D.; Niu, X.; Jayanthi, S.M.; others. SpeechVerse: A Large-scale Generalizable Audio Language Model. *arXiv preprint arXiv:2405.08295* **2024**.
170. Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; others. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759* **2024**.
171. Fang, Q.; Guo, S.; Zhou, Y.; Ma, Z.; Zhang, S.; Feng, Y. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. *arXiv preprint arXiv:2409.06666* **2024**.
172. Jin, Y.; Xu, K.; Xu, K.; Chen, L.; Liao, C.; Tan, J.; Mu, Y.; others. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. *arXiv preprint arXiv:2309.04669* **2023**.
173. Yu, L.; Shi, B.; Pasunuru, R.; Muller, B.; Golovneva, O.; Wang, T.; Babu, A.; Tang, B.; Karrer, B.; Sheynin, S.; others. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591* **2023**, *2*.
174. Pan, X.; Dong, L.; Huang, S.; Peng, Z.; Chen, W.; Wei, F. Kosmos-G: Generating Images in Context with Multimodal Large Language Models, 2024, [arXiv:cs.CV/2310.02992].
175. Lin, X.V.; Shrivastava, A.; Luo, L.; Iyer, S.; Lewis, M.; Gosh, G.; Zettlemoyer, L.; Aghajanyan, A. MoMa: Efficient Early-Fusion Pre-training with Mixture of Modality-Aware Experts. *arXiv preprint arXiv:2407.21770* **2024**.
176. Wu, Y.; Zhang, Z.; Chen, J.; Tang, H.; Li, D.; Fang, Y.; Zhu, L.; Xie, E.; Yin, H.; Yi, L.; others. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation. *arXiv preprint arXiv:2409.04429* **2024**.
177. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; others. Llama: Open and efficient foundation language models. *arXiv:2302.13971* **2023**.
178. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; others. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* **2024**.
179. Chiang, W.L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J.E.; Stoica, I.; Xing, E.P. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023.
180. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; others. Mistral 7B. *arXiv preprint arXiv:2310.06825* **2023**.
181. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; others. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* **2024**.
182. Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; others. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954* **2024**.
183. Kan, K.B.; Mun, H.; Cao, G.; Lee, Y. Mobile-LLaMA: Instruction Fine-Tuning Open-Source LLM for Network Analysis in 5G Networks. *IEEE Network* **2024**.
184. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; others. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* **2024**.
185. Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; others. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395* **2024**.
186. MistralAITeam. Mixtral of experts A high quality Sparse Mixture-of-Experts. [EB/OL], 2023. https://mistral.ai/news/mixtral-of-experts/ Accessed December 11, 2023.
187. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S.S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E.H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q.V.; Wei, J. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* **2024**, *25*, 1–53.

 doi:10.20944/preprints202411.0685.v1

188. Chen, X.; Djolonga, J.; Padlewski, P.; Mustafa, B.; Changpinyo, S.; Wu, J.; Ruiz, C.R.; Goodman, S.; Wang, X.; Tay, Y.; Shakeri, S.; Dehghani, M.; Salz, D.; Lucic, M.; Tschannen, M.; Nagrani, A.; Hu, H.; Joshi, M.; Pang, B.; Montgomery, C.; Pietrzyk, P.; Ritter, M.; Piergiovanni, A.; Minderer, M.; Pavetic, F.; Waters, A.; Li, G.; Alabdulmohsin, I.; Beyer, L.; Amelot, J.; Lee, K.; Steiner, A.P.; Li, Y.; Keysers, D.; Arnab, A.; Xu, Y.; Rong, K.; Kolesnikov, A.; Seyedhosseini, M.; Angelova, A.; Zhai, X.; Houlsby, N.; Soricut, R. PaLI-X: On Scaling up a Multilingual Vision and Language Model, 2023, [arXiv:cs.CV/2305.18565].
189. Bachmann, R.; Kar, O.F.; Mizrahi, D.; Garjani, A.; Gao, M.; Griffiths, D.; Hu, J.; Dehghan, A.; Zamir, A. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, 2024, [arXiv:cs.CV/2406.09406].
190. Mizrahi, D.; Bachmann, R.; Kar, O.F.; Yeo, T.; Gao, M.; Dehghan, A.; Zamir, A. 4M: Massively Multimodal Masked Modeling, 2023, [arXiv:cs.CV/2312.06647].
191. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs), 2023, [arXiv:cs.LG/1606.08415].
192. Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Zhang, S.; Duan, H.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Chen, Z.; Zhang, X.; Li, W.; Li, J.; Wang, W.; Chen, K.; He, C.; Zhang, X.; Dai, J.; Qiao, Y.; Lin, D.; Wang, J. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD, 2024, [arXiv:cs.CV/2404.06512].
193. Kar, O.F.; Tonioni, A.; Poklukar, P.; Kulshrestha, A.; Zamir, A.; Tombari, F. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204* **2024**.
194. Lu, J.; Gan, R.; Zhang, D.; Wu, X.; Wu, Z.; Sun, R.; Zhang, J.; Zhang, P.; Song, Y. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *arXiv preprint arXiv:2312.05278* **2023**.
195. Chen, K.; Shen, D.; Zhong, H.; Zhong, H.; Xia, K.; Xu, D.; Yuan, W.; Hu, Y.; Wen, B.; Zhang, T.; others. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177* **2024**.
196. FedusF, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **2022**, *23*, 1–39.
197. Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; Huang, J. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204* **2024**.
198. Shen, S.; Hou, L.; Zhou, Y.; Du, N.; Longpre, S.; Wei, J.; Chung, H.W.; Zoph, B.; Fedus, W.; Chen, X.; others. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705* **2023**.
199. Komatsuzaki, A.; Puigcerver, J.; Lee-Thorp, J.; Ruiz, C.R.; Mustafa, B.; Ainslie, J.; Tay, Y.; Dehghani, M.; Houlsby, N. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055* **2022**.
200. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. ACL, 2018, pp. 2556–2565.
201. Changpinyo, S.; Sharma, P.; Ding, N.; Soricut, R. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3557–3567. doi:10.1109/CVPR46437.2021.00356.
202. Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; Kim, S. Coyo-700m: Image-text pair dataset, 2022.
203. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325* **2015**.
204. Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 2443–2449.
205. Desai, K.; Kaul, G.; Aysola, Z.; Johnson, J. RedCaps: Web-curated image-text data created by the people, for the people. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks; Vanschoren, J.; Yeung, S., Eds., 2021, Vol. 1.
206. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114* **2021**.
207. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; others. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **2022**, *35*, 25278–25294.

208. Ordonez, V.; Kulkarni, G.; Berg, T. Im2Text: Describing Images Using 1 Million Captioned Photographs. Advances in Neural Information Processing Systems; Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; Weinberger, K., Eds. Curran Associates, Inc., 2011, Vol. 24.
209. Gadre, S.Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; others. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* **2024**, *36*.
210. Yu, Q.; Sun, Q.; Zhang, X.; Cui, Y.; Zhang, F.; Wang, X.; Liu, J. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550* **2023**.
211. Liu, Y.; Zhu, G.; Zhu, B.; Song, Q.; Ge, G.; Chen, H.; Qiao, G.; Peng, R.; Wu, L.; Wang, J. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems* **2022**, *35*, 16705–16717.
212. Lai, Z.; Zhang, H.; Zhang, B.; Wu, W.; Bai, H.; Timofeev, A.; Du, X.; Gan, Z.; Shan, J.; Chuah, C.N.; Yang, Y.; Cao, M. VeCLIP: Improving CLIP Training via Visual-enriched Captions, 2024, [arXiv:cs.CV/2310.07699].
213. Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; others. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems* **2022**, *35*, 26418–26431.
214. Sharifzadeh, S.; Kaplanis, C.; Pathak, S.; Kumaran, D.; Ilic, A.; Mitrovic, J.; Blundell, C.; Banino, A. Synth ^{2}: Boosting Visual-Language Models with Synthetic Captions and Image Embeddings. *arXiv preprint arXiv:2403.07750* **2024**.
215. Singla, V.; Yue, K.; Paul, S.; Shirkavand, R.; Jayawardhana, M.; Ganjdanesh, A.; Huang, H.; Bhatele, A.; Somepalli, G.; Goldstein, T. From Pixels to Prose: A Large Dataset of Dense Image Captions. *arXiv preprint arXiv:2406.10328* **2024**.
216. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.J. Yfcc100m: The new data in multimedia research. *Communications of the ACM* **2016**, *59*, 64–73.
217. Onoe, Y.; Rane, S.; Berger, Z.; Bitton, Y.; Cho, J.; Garg, R.; Ku, A.; Parekh, Z.; Pont-Tuset, J.; Tanzer, G.; others. DOCCI: Descriptions of Connected and Contrasting Images. *arXiv preprint arXiv:2404.19753* **2024**.
218. Garg, R.; Burns, A.; Ayan, B.K.; Bitton, Y.; Montgomery, C.; Onoe, Y.; Bunner, A.; Krishna, R.; Baldridge, J.; Soricut, R. ImageInWords: Unlocking Hyper-Detailed Image Descriptions. *arXiv preprint arXiv:2405.02793* **2024**.
219. Urbanek, J.; Bordes, F.; Astolfi, P.; Williamson, M.; Sharma, V.; Romero-Soriano, A. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26700–26709.
220. Xu, H.; Xie, S.; Tan, X.E.; Huang, P.Y.; Howes, R.; Sharma, V.; Li, S.W.; Ghosh, G.; Zettlemoyer, L.; Feichtenhofer, C. Demystifying clip data. *arXiv preprint arXiv:2309.16671* **2023**.
221. Wang, W.; Shi, M.; Li, Q.; Wang, W.; Huang, Z.; Xing, L.; Chen, Z.; Li, H.; Zhu, X.; Cao, Z.; others. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907* **2023**.
222. Miech, A.; Zhukov, D.; Alayrac, J.B.; Tapaswi, M.; Laptev, I.; Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2630–2640.
223. Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; others. Ego4d: Around the world in 3,000 hours of egocentric video. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18995–19012.
224. Nagrani, A.; Seo, P.H.; Seybold, B.; Hauth, A.; Manen, S.; Sun, C.; Schmid, C. Learning audio-video modalities from image captions. European Conference on Computer Vision. Springer, 2022, pp. 407–426.
225. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1728–1738.
226. Chen, T.S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.w.; Jeon, B.E.; Fang, Y.; Lee, H.Y.; Ren, J.; Yang, M.H.; others. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13320–13331.

227. Zhu, B.; Lin, B.; Ning, M.; Yan, Y.; Cui, J.; Wang, H.; Pang, Y.; Jiang, W.; Zhang, J.; Li, Z.; others. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852* **2023**.
228. Drossos, K.; Lipping, S.; Virtanen, T. Clotho: An audio captioning dataset. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 736–740.
229. Kim, C.D.; Kim, B.; Lee, H.; Kim, G. Audiocaps: Generating captions for audios in the wild. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 119–132.
230. Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M.D.; Zou, Y.; Wang, W. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2024**.
231. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017, pp. 776–780.
232. Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; Guo, B. Advancing high-resolution video-language representation with large-scale video transcriptions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5036–5045.
233. Zhou, L.; Xu, C.; Corso, J. Towards automatic learning of procedures from web instructional videos. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.
234. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in homes: Crowdsourcing data collection for activity understanding. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 510–526.
235. Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; Chua, T.S. Annotating Objects and Relations in User-Generated Videos. Proceedings of the 2019 on International Conference on Multimedia Retrieval. ACM, 2019, pp. 279–287.
236. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; others. The" something something" video database for learning and evaluating visual common sense. Proceedings of the IEEE international conference on computer vision, 2017, pp. 5842–5850.
237. Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, M.S. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia* **2019**, *22*, 554–565.
238. Wang, W.; Yang, H.; Tuo, Z.; He, H.; Zhu, J.; Fu, J.; Liu, J. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. *arXiv preprint arXiv:2305.10874* **2023**.
239. Hu, A.; Xu, H.; Ye, J.; Yan, M.; Zhang, L.; Zhang, B.; Li, C.; Zhang, J.; Jin, Q.; Huang, F.; Zhou, J. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. *ArXiv* **2024**, *abs/2403.12895*.
240. Hu, A.; Xu, H.; Ye, J.; Yan, M.; Zhang, L.; Zhang, B.; Li, C.; Zhang, J.; Jin, Q.; Huang, F.; others. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895* **2024**.
241. Ordonez, V.; Kulkarni, G.; Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* **2011**, *24*.
242. Changpinyo, S.; Sharma, P.; Ding, N.; Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3558–3568.
243. Yang, D.; Huang, S.; Lu, C.; Han, X.; Zhang, H.; Gao, Y.; Hu, Y.; Zhao, H. Vript: A Video Is Worth Thousands of Words. *arXiv preprint arXiv:2406.06040* **2024**.
244. Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
245. Chen, S.; Li, H.; Wang, Q.; Zhao, Z.; Sun, M.; Zhu, X.; Liu, J. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems* **2023**, *36*, 72842–72866.

246. Chen, S.; He, X.; Guo, L.; Zhu, X.; Wang, W.; Tang, J.; Liu, J. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345* **2023**.
247. Kong, Z.; Lee, S.g.; Ghosal, D.; Majumder, N.; Mehrish, A.; Valle, R.; Poria, S.; Catanzaro, B. Improving text-to-audio models with synthetic captions. *arXiv preprint arXiv:2406.15487* **2024**.
248. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.F.; Wang, W.Y. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4581–4591.
249. Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; Russell, B. Localizing moments in video with natural language. Proceedings of the IEEE international conference on computer vision, 2017, pp. 5803–5812.
250. Fang, Y.; Zhu, L.; Lu, Y.; Wang, Y.; Molchanov, P.; Cho, J.H.; Pavone, M.; Han, S.; Yin, H. VILA^2: VILA Augmented VILA. *arXiv preprint arXiv:2407.17453* **2024**.
251. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *arXiv:2304.08485* **2023**.
252. Tang, B.J.; Boggust, A.; Satyanarayan, A. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356* **2023**.
253. Wang, B.; Li, G.; Zhou, X.; Chen, Z.; Grossman, T.; Li, Y. Screen2words: Automatic mobile UI summarization with multimodal learning. The 34th Annual ACM Symposium on User Interface Software and Technology, 2021, pp. 498–510.
254. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.
255. Li, L.; Wang, Y.; Xu, R.; Wang, P.; Feng, X.; Kong, L.; Liu, Q. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231* **2024**.
256. Li, X.; Zhang, F.; Diao, H.; Wang, Y.; Wang, X.; Duan, L.Y. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303* **2024**.
257. Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; others. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325* **2024**.
258. Erfei, C.; Yinan, H.; Zheng, M.; Zhe, C.; Hao, T.; Weiyun, W.; Kunchang, L.; Yi, W.; Wenhai, W.; Xizhou, Z.; Lewei, L.; Tong, L.; Yali, W.; Limin, W.; Yu, Q.; Jifeng, D. Comprehensive Multimodal Annotations With GPT-4o, 2024.
259. Zhu, W.; Hessel, J.; Awadalla, A.; Gadre, S.Y.; Dodge, J.; Fang, A.; Yu, Y.; Schmidt, L.; Wang, W.Y.; Choi, Y. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NeurIPS* **2024**, *36*.
260. Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; others. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* **2024**, *36*.
261. Awadalla, A.; Xue, L.; Lo, O.; Shu, M.; Lee, H.; Guha, E.K.; Jordan, M.; Shen, S.; Awadalla, M.; Savarese, S.; others. MINT-1T: Scaling Open-Source Multimodal Data by 10x: A Multimodal Dataset with One Trillion Tokens. *arXiv preprint arXiv:2406.11271* **2024**.
262. Li, Q.; Chen, Z.; Wang, W.; Wang, W.; Ye, S.; Jin, Z.; Chen, G.; He, Y.; Gao, Z.; Cui, E.; others. Omni-Corpus: An Unified Multimodal Corpus of 10 Billion-Level Images Interleaved with Text. *arXiv preprint arXiv:2406.08418* **2024**.
263. Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O.K.; Patra, B.; others. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems* **2023**, *36*, 72096–72109.
264. Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; others. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942* **2023**.
265. Wang, A.J.; Li, L.; Lin, K.Q.; Wang, J.; Lin, K.; Yang, Z.; Wang, L.; Shou, M.Z. COSMO: COntrastive Streamlined MultimOdal Model with Interleaved Pre-Training. *arXiv preprint arXiv:2401.00849* **2024**.
266. Sun, Q.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; Wang, Y.; Gao, H.; Liu, J.; Huang, T.; Wang, X. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222* **2023**.

267. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.
268. Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Authur, R.; Bogin, B.; Chandu, K.; Dumas, J.; Elazar, Y.; others. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159* **2024**.
269. Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; Launay, J. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* **2023**.
270. Yuan, S.; Zhao, H.; Du, Z.; Ding, M.; Liu, X.; Cen, Y.; Zou, X.; Yang, Z.; Tang, J. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open* **2021**, *2*, 65–68.
271. Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C.C.T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; others. Textbooks are all you need. *arXiv preprint arXiv:2306.11644* **2023**.
272. Hernandez, D.; Brown, T.; Conerly, T.; DasSarma, N.; Drain, D.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Henighan, T.; Hume, T.; others. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487* **2022**.
273. Suárez, P.J.O.; Sagot, B.; Romary, L. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache, 2019.
274. Computer, T. RedPajama: an Open Dataset for Training Large Language Models **2023**.
275. Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; Carlini, N. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499* **2021**.
276. Silcock, E.; D'Amico-Wong, L.; Yang, J.; Dell, M. Noise-robust de-duplication at scale. Technical report, National Bureau of Economic Research, 2022.
277. Kaddour, J. The minipile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442* **2023**.
278. Abbas, A.; Tirumala, K.; Simig, D.; Ganguli, S.; Morcos, A.S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540* **2023**.
279. Zauner, C. Implementation and benchmarking of perceptual image hash functions **2010**.
280. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3195–3204.
281. Zhang, P.; Li, C.; Qiao, L.; Cheng, Z.; Pu, S.; Niu, Y.; Wu, F. VSR: a unified framework for document layout analysis combining vision, semantics and relations. Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16. Springer, 2021, pp. 115–130.
282. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. arXiv **2022**.
283. Gurari, D.; Li, Q.; Stangl, A.J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; Bigham, J.P. Vizwiz grand challenge: Answering visual questions from blind people. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3608–3617.
284. Zhu, Y.; Groth, O.; Bernstein, M.; Fei-Fei, L. Visual7w: Grounded question answering in images. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4995–5004.
285. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems* **2020**, *33*, 2611–2624.
286. Acharya, M.; Kafle, K.; Kanan, C. TallyQA: Answering complex counting questions. Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 8076–8084.
287. Xia, H.; Lan, R.; Li, H.; Song, S. ST-VQA: shrinkage transformer with accurate alignment for visual question answering. *Applied Intelligence* **2023**, *53*, 20967–20978.
288. Chang, S.; Palzer, D.; Li, J.; Fosler-Lussier, E.; Xiao, N. MapQA: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545* **2022**.

289. Shah, S.; Mishra, A.; Yadati, N.; Talukdar, P.P. Kvqa: Knowledge-aware visual question answering. Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 8876–8884.
290. Lerner, P.; Ferret, O.; Guinaudeau, C.; Le Borgne, H.; Besançon, R.; Moreno, J.G.; Lovón Melgarejo, J. ViQuAE, a dataset for knowledge-based visual question answering about named entities. 45th ACM SIGIR, 2022, pp. 3108–3120.
291. Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; Tao, D. Activitynet-qa: A dataset for understanding complex web videos via question answering. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 9127–9134.
292. Xiao, J.; Shang, X.; Yao, A.; Chua, T.S. Next-qa: Next phase of question-answering to explaining temporal actions. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9777–9786.
293. Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; Tenenbaum, J.B. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442* **2019**.
294. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Learning to answer visual questions from web videos. *arXiv preprint arXiv:2205.05019* **2022**.
295. Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; Kim, G. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2758–2766.
296. Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J.B.; Gan, C. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711* **2024**.
297. Lei, J.; Yu, L.; Bansal, M.; Berg, T.L. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* **2018**.
298. Jahagirdar, S.; Mathew, M.; Karatzas, D.; Jawahar, C. Watching the news: Towards videoqa models that can read. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4441–4450.
299. Marti, U.V.; Bunke, H. The IAM-database: an English sentence database for offline handwriting recognition. *International journal on document analysis and recognition* **2002**, *5*, 39–46.
300. Mishra, A.; Shekhar, S.; Singh, A.K.; Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. 2019 ICDAR. IEEE, 2019, pp. 947–952.
301. Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; Rohrbach, M. Towards vqa models that can read. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8317–8326.
302. Wendler, C. wendlerc/RenderedText, 2023.
303. Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; Park, S. OCR-Free Document Understanding Transformer. European Conference on Computer Vision (ECCV), 2022.
304. Mathew, M.; Karatzas, D.; Jawahar, C. Docvqa: A dataset for vqa on document images. WACV, 2021, pp. 2200–2209.
305. Kantharaj, S.; Leong, R.T.K.; Lin, X.; Masry, A.; Thakkar, M.; Hoque, E.; Joty, S. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486* **2022**.
306. Kafle, K.; Price, B.; Cohen, S.; Kanan, C. Dvqa: Understanding data visualizations via question answering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5648–5656.
307. Masry, A.; Long, D.X.; Tan, J.Q.; Joty, S.; Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244* **2022**.
308. Methani, N.; Ganguly, P.; Khapra, M.M.; Kumar, P. Plotqa: Reasoning over scientific plots. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1527–1536.
309. Kahou, S.E.; Michalski, V.; Atkinson, A.; Kádár, Á.; Trischler, A.; Bengio, Y. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300* **2017**.
310. Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; Jawahar, C. Infographicvqa. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1697–1706.
311. Lu, P.; Qiu, L.; Chang, K.W.; Wu, Y.N.; Zhu, S.C.; Rajpurohit, T.; Clark, P.; Kalyan, A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610* **2022**.

312. Hsiao, Y.C.; Zubach, F.; Wang, M.; others. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199* **2022**.
313. Tanaka, R.; Nishida, K.; Yoshida, S. Visualmrc: Machine reading comprehension on document images. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 13878–13888.
314. Van Landeghem, J.; Tito, R.; Borchmann, Ł.; Pietruszka, M.; Joziak, P.; Powalski, R.; Jurkiewicz, D.; Coustaty, M.; Anckaert, B.; Valveny, E.; others. Document understanding dataset and evaluation (dude). Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19528–19540.
315. Tito, R.; Karatzas, D.; Valveny, E. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition* **2023**, *144*, 109834.
316. Gao, J.; Pi, R.; Zhang, J.; Ye, J.; Zhong, W.; Wang, Y.; Hong, L.; Han, J.; Xu, H.; Li, Z.; others. G-llava A diagram is worth a dozen images.: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370* **2023**.
317. Cao, J.; Xiao, J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 1511–1520.
318. Kazemi, M.; Alvari, H.; Anand, A.; Wu, J.; Chen, X.; Soricut, R. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241* **2023**.
319. Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; Zhu, S.C. Raven: A dataset for relational and analogical visual reasoning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5317–5327.
320. Saikh, T.; Ghosal, T.; Mittal, A.; Ekbal, A.; Bhattacharyya, P. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries* **2022**, *23*, 289–301.
321. Lu, P.; Gong, R.; Jiang, S.; Qiu, L.; Huang, S.; Liang, X.; Zhu, S.C. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165* **2021**.
322. Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; Farhadi, A. A diagram is worth a dozen images. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, 2016, pp. 235–251.
323. Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; Zhu, S.C. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214* **2021**.
324. Kembhavi, A.; Seo, M.; Schwenk, D.; Choi, J.; Farhadi, A.; Hajishirzi, H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, 2017, pp. 4999–5007.
325. Laurençon, H.; Tronchon, L.; Sanh, V. Unlocking the conversion of Web Screenshots into HTML Code with the WebSight Dataset. *arXiv preprint arXiv:2403.09029* **2024**.
326. Belouadi, J.; Lauscher, A.; Eger, S. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367* **2023**.
327. Si, C.; Zhang, Y.; Yang, Z.; Liu, R.; Yang, D. Design2Code: How Far Are We From Automating Front-End Engineering? *arXiv preprint arXiv:2403.03163* **2024**.
328. Lindström, A.D.; Abraham, S.S. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358* **2022**.
329. Gupta, T.; Marten, R.; Kembhavi, A.; Hoiem, D. Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653* **2022**.
330. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; others. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **2017**, *123*, 32–73.
331. Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; Sun, J. Objects365: A large-scale, high-quality dataset for object detection. Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8430–8439.
332. Xu, Z.; Shen, Y.; Huang, L. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv:2212.10773* **2022**.
333. Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; Chen, W. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483* **2024**.
334. Chen, F.; Han, M.; Zhao, H.; Zhang, Q.; Shi, J.; Xu, S.; Xu, B. X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. *arXiv:2305.04160* **2023**.

335. Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; others. M3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv:2306.04387* **2023**.
336. Li, Y.; Zhang, G.; Ma, Y.; Yuan, R.; Zhu, K.; Guo, H.; Liang, Y.; Liu, J.; Yang, J.; Wu, S.; others. OmniBench: Towards The Future of Universal Omni-Language Models. *arXiv preprint arXiv:2409.15272* **2024**.
337. Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; Wang, L.; Qiao, Y. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark, 2023, [arXiv:cs.CV/2311.17005].
338. Ren, S.; Yao, L.; Li, S.; Sun, X.; Hou, L. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. *ArXiv* **2023**, *abs/2312.02051*.
339. Xu, Z.; Feng, C.; Shao, R.; Ashby, T.; Shen, Y.; Jin, D.; Cheng, Y.; Wang, Q.; Huang, L. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690* **2024**.
340. Liu, J.; Wang, Z.; Ye, Q.; Chong, D.; Zhou, P.; Hua, Y. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956* **2023**.
341. Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790* **2023**.
342. Zhao, H.; Cai, Z.; Si, S.; Ma, X.; An, K.; Chen, L.; Liu, Z.; Wang, S.; Han, W.; Chang, B. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915* **2023**.
343. Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; Tian, Y. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems* **2024**, *36*.
344. Lai, Z.; Zhang, H.; Zhang, B.; Wu, W.; Bai, H.; Timofeev, A.; Du, X.; Gan, Z.; Shan, J.; Chuah, C.N.; others. VeCLIP: Improving CLIP Training via Visual-Enriched Captions. European Conference on Computer Vision. Springer, 2025, pp. 111–127.
345. Yu, Q.; Sun, Q.; Zhang, X.; Cui, Y.; Zhang, F.; Cao, Y.; Wang, X.; Liu, J. Capsfusion: Rethinking image-text data at scale. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14022–14032.
346. Pi, R.; Gao, J.; Diao, S.; Pan, R.; Dong, H.; Zhang, J.; Yao, L.; Han, J.; Xu, H.; Zhang, L.K.T. DetGPT: Detect What You Need via Reasoning. *arXiv:2305.14167* **2023**.
347. Zhao, L.; Yu, E.; Ge, Z.; Yang, J.; Wei, H.; Zhou, H.; Sun, J.; Peng, Y.; Dong, R.; Han, C.; others. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474* **2023**.
348. Liu, Z.; Chu, T.; Zang, Y.; Wei, X.; Dong, X.; Zhang, P.; Liang, Z.; Xiong, Y.; Qiao, Y.; Lin, D.; others. MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs. *arXiv preprint arXiv:2406.11833* **2024**.
349. Pi, R.; Zhang, J.; Han, T.; Zhang, J.; Pan, R.; Zhang, T. Personalized Visual Instruction Tuning. *arXiv preprint arXiv:2410.07113* **2024**.
350. Zhang, R.; Wei, X.; Jiang, D.; Zhang, Y.; Guo, Z.; Tong, C.; Liu, J.; Zhou, A.; Wei, B.; Zhang, S.; others. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739* **2024**.
351. Chen, G.H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; Wang, B. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. *arXiv:2402.11684* **2024**.
352. Wang, W.; Ren, Y.; Luo, H.; Li, T.; Yan, C.; Chen, Z.; Wang, W.; Li, Q.; Lu, L.; Zhu, X.; others. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474* **2024**.
353. Yang, R.; Song, L.; Li, Y.; Zhao, S.; Ge, Y.; Li, X.; Shan, Y. GPT4Tools: Teaching Large Language Model to Use Tools via Self-instruction. *arXiv:2305.18752* **2023**.
354. Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; Li, C. Video Instruction Tuning With Synthetic Data. *arXiv preprint arXiv:2410.02713* **2024**.
355. Zhang, R.; Gui, L.; Sun, Z.; Feng, Y.; Xu, K.; Zhang, Y.; Fu, D.; Li, C.; Hauptmann, A.; Bisk, Y.; others. Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward. *arXiv preprint arXiv:2404.01258* **2024**.
356. Tang, J.; Lin, C.; Zhao, Z.; Wei, S.; Wu, B.; Liu, Q.; Feng, H.; Li, Y.; Wang, S.; Liao, L.; others. TextSquare: Scaling up Text-Centric Visual Instruction Tuning. *arXiv preprint arXiv:2404.12803* **2024**.
357. Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Yang, J.; Li, C.; Liu, Z. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425* **2023**.

358. Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; Sun, T. LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding, 2024, [arXiv:cs.CV/2306.17107].
359. Wang, J.; Meng, L.; Weng, Z.; He, B.; Wu, Z.; Jiang, Y.G. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574* **2023**.
360. Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. The Twelfth International Conference on Learning Representations, 2023.
361. Liu, J.; Huang, X.; Zheng, J.; Liu, B.; Wang, J.; Yoshie, O.; Liu, Y.; Li, H. MM-Instruct: Generated Visual Instructions for Large Multimodal Model Alignment. *arXiv preprint arXiv:2406.19736* **2024**.
362. Zhao, B.; Wu, B.; He, M.; Huang, T. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087* **2023**.
363. Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; Li, C. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895* **2024**.
364. Wang, B.; Wu, F.; Han, X.; Peng, J.; Zhong, H.; Zhang, P.; Dong, X.; Li, W.; Li, W.; Wang, J.; others. Vigc: Visual instruction generation and correction. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 5309–5317.
365. Gao, W.; Deng, Z.; Niu, Z.; Rong, F.; Chen, C.; Gong, Z.; Zhang, W.; Xiao, D.; Li, F.; Cao, Z.; others. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174* **2023**.
366. Li, H.; Li, S.; Cai, D.; Wang, L.; Liu, L.; Watanabe, T.; Yang, Y.; Shi, S. Textbind: Multi-turn interleaved multimodal instruction-following. *arXiv preprint arXiv:2309.08637* **2023**.
367. Pan, J.; Wu, J.; Gaur, Y.; Sivasankaran, S.; Chen, Z.; Liu, S.; Li, J. Cosmic: Data efficient instruction-tuning for speech in-context learning. *arXiv preprint arXiv:2311.02248* **2023**.
368. Huang, Y.; Meng, Z.; Liu, F.; Su, Y.; Collier, N.; Lu, Y. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463* **2023**.
369. Li, Y.; Zhang, C.; Yu, G.; Wang, Z.; Fu, B.; Lin, G.; Shen, C.; Chen, L.; Wei, Y. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253* **2023**.
370. Zhao, Y.; Lin, Z.; Zhou, D.; Huang, Z.; Feng, J.; Kang, B. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581* **2023**.
371. Luo, R.; Zhang, H.; Chen, L.; Lin, T.E.; Liu, X.; Wu, Y.; Yang, M.; Wang, M.; Zeng, P.; Gao, L.; others. MMEvol: Empowering Multimodal Large Language Models with Evol-Instruct. *arXiv preprint arXiv:2409.05840* **2024**.
372. Liu, Y.; Cao, Y.; Gao, Z.; Wang, W.; Chen, Z.; Wang, W.; Tian, H.; Lu, L.; Zhu, X.; Lu, T.; others. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838* **2024**.
373. Maaz, M.; Rasheed, H.; Khan, S.; Khan, F. VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding. *arXiv preprint arXiv:2406.09418* **2024**.
374. Zhang, H.; Gao, M.; Gan, Z.; Dufter, P.; Wenzel, N.; Huang, F.; Shah, D.; Du, X.; Zhang, B.; Li, Y.; others. MM1. 5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning. *arXiv preprint arXiv:2409.20566* **2024**.
375. Zhao, Z.; Guo, L.; Yue, T.; Chen, S.; Shao, S.; Zhu, X.; Yuan, Z.; Liu, J. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *arXiv:2305.16103* **2023**.
376. Panagopoulou, A.; Xue, L.; Yu, N.; Li, J.; Li, D.; Joty, S.; Xu, R.; Savarese, S.; Xiong, C.; Niebles, J.C. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799* **2023**.
377. Jia, M.; Yu, W.; Ma, K.; Fang, T.; Zhang, Z.; Ouyang, S.; Zhang, H.; Jiang, M.; Yu, D. LEOPARD: A Vision Language Model For Text-Rich Multi-Image Tasks. *arXiv preprint arXiv:2410.01744* **2024**.
378. Yin, Z.; Wang, J.; Cao, J.; Shi, Z.; Liu, D.; Li, M.; Sheng, L.; Bai, L.; Huang, X.; Wang, Z.; others. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. *arXiv:2306.06687* **2023**.
379. Li, Z.; Luo, R.; Zhang, J.; Qiu, M.; Wei, Z. VoCoT: Unleashing Visually Grounded Multi-Step Reasoning in Large Multi-Modal Models. *arXiv preprint arXiv:2405.16919* **2024**.
380. Gong, Y.; Liu, A.H.; Luo, H.; Karlinsky, L.; Glass, J. Joint audio and speech understanding. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023, pp. 1–8.

381. Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; Li, H. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999* **2024**.
382. Yun, S.; Lin, H.; Thushara, R.; Bhat, M.Q.; Wang, Y.; Jiang, Z.; Deng, M.; Wang, J.; Tao, T.; Li, J.; others. Web2Code: A Large-scale Webpage-to-Code Dataset and Evaluation Framework for Multimodal LLMs. *arXiv preprint arXiv:2406.20098* **2024**.
383. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; others. Scaling instruction-finetuned language models. *arXiv:2210.11416* **2022**.
384. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* **2023**, *3*, 7.
385. Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; others. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems* **2024**, *36*.
386. Chen, Y.; Liu, L.; Ding, C. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843* **2023**.
387. Zhang, X.; Kuang, H.; Mou, X.; Lyu, H.; Wu, K.; Chen, S.; Luo, J.; Huang, X.; Wei, Z. SoMeLVLM: A Large Vision Language Model for Social Media Processing. *arXiv preprint arXiv:2402.13022* **2024**.
388. Liu, J.; Wang, Z.; Ye, Q.; Chong, D.; Zhou, P.; Hua, Y. Qilin-Med-VL: Towards Chinese Large Vision-Language Model for General Healthcare, 2023, [arXiv:cs.CV/2310.17956].
389. Zhao, Z.; Guo, L.; Yue, T.; Chen, S.; Shao, S.; Zhu, X.; Yuan, Z.; Liu, J. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103* **2023**.
390. Ren, S.; Yao, L.; Li, S.; Sun, X.; Hou, L. Timechat: A time-sensitive multimodal large language model for long video understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14313–14323.
391. Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; others. Mvbench: A comprehensive multi-modal video understanding benchmark. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22195–22206.
392. Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; Kong, L.; Liu, Q. M^3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv:2306.04387* **2023**.
393. Fei, J.; Li, D.; Deng, Z.; Wang, Z.; Liu, G.; Wang, H. Video-CCAM: Enhancing Video-Language Understanding with Causal Cross-Attention Masks for Short and Long Videos. *arXiv preprint arXiv:2408.14023* **2024**.
394. Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N.A.; Khashabi, D.; Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560* **2022**.
395. Pi, R.; Gao, J.; Diao, S.; Pan, R.; Dong, H.; Zhang, J.; Yao, L.; Han, J.; Xu, H.; Kong, L.; others. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167* **2023**.
396. Pan, J.; Wu, J.; Gaur, Y.; Sivasankaran, S.; Chen, Z.; Liu, S.; Li, J. COSMIC: Data Efficient Instruction-tuning For Speech In-Context Learning, 2024, [arXiv:cs.CL/2311.02248].
397. Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; others. Internlm2 technical report. *arXiv preprint arXiv:2403.17297* **2024**.
398. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, 2021, [arXiv:cs.CL/2106.09685].
399. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.
400. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
401. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. Text summarization branches out, 2004, pp. 74–81.
402. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575.

403. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
404. Xu, P.; Shao, W.; Zhang, K.; Gao, P.; Liu, S.; Lei, M.; Meng, F.; Huang, S.; Qiao, Y.; Luo, P. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv:2306.09265* **2023**.
405. Li, Z.; Wang, Y.; Du, M.; Liu, Q.; Wu, B.; Zhang, J.; Zhou, C.; Fan, Z.; Fu, J.; Chen, J.; others. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569* **2023**.
406. Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; others. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv:2306.13394* **2023**.
407. Zhang, W.; Aljunied, M.; Gao, C.; Chia, Y.K.; Bing, L. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems* **2023**, *36*, 5484–5505.
408. Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; Zhang, W.; Lin, Y.; Liu, S.; others. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006* **2024**.
409. Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; Shan, Y. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv:2307.16125* **2023**.
410. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; others. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281* **2023**.
411. Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; others. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330* **2024**.
412. Liu, Y.; Li, Z.; Li, H.; Yu, W.; Huang, M.; Peng, D.; Liu, M.; Chen, M.; Li, C.; Jin, L.; others. On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895* **2023**.
413. Du, M.; Wu, B.; Li, Z.; Huang, X.; Wei, Z. EmbSpatial-Bench: Benchmarking Spatial Understanding for Embodied Tasks with Large Vision-Language Models. *arXiv preprint arXiv:2406.05756* **2024**.
414. Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.W.; Gao, P.; others. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624* **2024**.
415. Chen, P.; Ye, J.; Wang, G.; Li, Y.; Deng, Z.; Li, W.; Li, T.; Duan, H.; Huang, Z.; Su, Y.; others. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361* **2024**.
416. Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; others. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv:2311.16502* **2023**.
417. Liu, Y.; Li, Z.; Li, H.; Yu, W.; Huang, M.; Peng, D.; Liu, M.; Chen, M.; Li, C.; Jin, L.; Bai, X. On the Hidden Mystery of OCR in Large Multimodal Models. *ArXiv* **2023**, *abs/2305.07895*.
418. Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.W.; Galley, M.; Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* **2023**.
419. Li, S.; Tajbakhsh, N. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349* **2023**.
420. Bitton, Y.; Bansal, H.; Hessel, J.; Shao, R.; Zhu, W.; Awadalla, A.; Gardner, J.; Taori, R.; Schmidt, L. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595* **2023**.
421. Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; Wang, L. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities, 2023, [arXiv:cs.AI/2308.02490].
422. Bai, S.; Yang, S.; Bai, J.; Wang, P.; Zhang, X.; Lin, J.; Wang, X.; Zhou, C.; Zhou, J. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890* **2023**.
423. Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1645–1653.

424. Mangalam, K.; Akshulakov, R.; Malik, J. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* **2024**, *36*.
425. Patraucean, V.; Smaira, L.; Gupta, A.; Recasens, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Malinowski, M.; Yang, Y.; Doersch, C.; others. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems* **2024**, *36*.
426. Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; others. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075* **2024**.
427. Li, Y.; Chen, X.; Hu, B.; Wang, L.; Shi, H.; Zhang, M. VideoVista: A Versatile Benchmark for Video Understanding and Reasoning. *arXiv preprint arXiv:2406.11303* **2024**.
428. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5288–5296.
429. Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; others. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476* **2024**.
430. Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; Liu, Z. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264* **2024**.
431. Du, J.; Na, X.; Liu, X.; Bu, H. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583* **2018**.
432. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* **2019**.
433. Wang, C.; Wu, A.; Pino, J. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310* **2020**.
434. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* **2018**.
435. Lipping, S.; Sudarsanam, P.; Drossos, K.; Virtanen, T. Clotho-aqa: A crowdsourced dataset for audio question answering. 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022, pp. 1140–1144.
436. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, 2016, pp. 382–398.
437. Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; Murphy, K. Improved image captioning via policy gradient optimization of spider. Proceedings of the IEEE international conference on computer vision, 2017, pp. 873–881.
438. Yang, Q.; Xu, J.; Liu, W.; Chu, Y.; Jiang, Z.; Zhou, X.; Leng, Y.; Lv, Y.; Zhao, Z.; Zhou, C.; others. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. *arXiv preprint arXiv:2402.07729* **2024**.
439. Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; Chen, N.F. AudioBench: A Universal Benchmark for Audio Large Language Models. *arXiv preprint arXiv:2406.16020* **2024**.
440. Huang, C.y.; Lu, K.H.; Wang, S.H.; Hsiao, C.Y.; Kuan, C.Y.; Wu, H.; Arora, S.; Chang, K.W.; Shi, J.; Peng, Y.; others. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 12136–12140.
441. Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; others. Emu3: Next-Token Prediction is All You Need. *arXiv preprint arXiv:2409.18869* **2024**.
442. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; others. Scaling rectified flow transformers for high-resolution image synthesis. Forty-first International Conference on Machine Learning, 2024.
443. Xue, Z.; Song, G.; Guo, Q.; Liu, B.; Zong, Z.; Liu, Y.; Luo, P. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems* **2024**, *36*.

444. Zheng, Q.; Zheng, L.; Guo, Y.; Li, Y.; Xu, S.; Deng, J.; Xu, H. Self-Adaptive Reality-Guided Diffusion for Artifact-Free Super-Resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25806–25816.
445. Zheng, D.; Wu, X.M.; Yang, S.; Zhang, J.; Hu, J.F.; Zheng, W.S. Selective Hourglass Mapping for Universal Image Restoration Based on Diffusion Model. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25445–25455.
446. Mou, C.; Wang, X.; Song, J.; Shan, Y.; Zhang, J. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8488–8497.
447. Shi, J.; Xiong, W.; Lin, Z.; Jung, H.J. Instantbooth: Personalized text-to-image generation without test-time finetuning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8543–8552.
448. Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; Huang, T. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284* **2023**.
449. Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; Lee, Y.J. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems* **2024**, *36*.
450. Geng, Z.; Yang, B.; Hang, T.; Li, C.; Gu, S.; Zhang, T.; Bao, J.; Zhang, Z.; Li, H.; Hu, H.; others. Instructdiffusion: A generalist modeling interface for vision tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12709–12720.
451. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Ieee, 2009, pp. 248–255.
452. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **2017**, *30*.
453. Bińkowski, M.; Sutherland, D.J.; Arbel, M.; Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401* **2018**.
454. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.
455. Li, D.; Kamko, A.; Akhgari, E.; Sabet, A.; Xu, L.; Doshi, S. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245* **2024**.
456. Ku, M.; Jiang, D.; Wei, C.; Yue, X.; Chen, W. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867* **2023**.
457. Peng, Y.; Cui, Y.; Tang, H.; Qi, Z.; Dong, R.; Bai, J.; Han, C.; Ge, Z.; Zhang, X.; Xia, S.T. DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation. *arXiv preprint arXiv:2406.16855* **2024**.
458. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* **2021**.
459. Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; Ramanan, D. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291* **2024**.
460. Huang, K.; Sun, K.; Xie, E.; Li, Z.; Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* **2023**, *36*, 78723–78747.
461. Ghosh, D.; Hajishirzi, H.; Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems* **2024**, *36*.
462. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; others. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **2022**, *35*, 36479–36494.
463. Petsiuk, V.; Siemenn, A.E.; Surbehera, S.; Chin, Z.; Tyser, K.; Hunter, G.; Raghavan, A.; Hicke, Y.; Plummer, B.A.; Kerret, O.; others. Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark. arXiv 2022. *arXiv preprint cs.CV/2211.12112* **2022**.
464. Cho, J.; Zala, A.; Bansal, M. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3043–3054.
465. Barratt, S.; Sharma, R. A note on the inception score. *arXiv preprint arXiv:1801.01973* **2018**.

466. Guo, J.; Chai, W.; Deng, J.; Huang, H.W.; Ye, T.; Xu, Y.; Zhang, J.; Hwang, J.N.; Wang, G. Versat2i: Improving text-to-image models with versatile reward. *arXiv preprint arXiv:2403.18493* **2024**.
467. Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; others. Rich human feedback for text-to-image generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19401–19411.
468. Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* **2024**, *36*.
469. Cho, J.; Zala, A.; Bansal, M. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems* **2024**, *36*.
470. Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; Smith, N.A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20406–20417.
471. Wu, T.; Yang, G.; Li, Z.; Zhang, K.; Liu, Z.; Guibas, L.; Lin, D.; Wetzstein, G. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22227–22238.
472. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; others. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005* **2022**.
473. Cho, J.; Hu, Y.; Garg, R.; Anderson, P.; Krishna, R.; Baldridge, J.; Bansal, M.; Pont-Tuset, J.; Wang, S. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235* **2023**.
474. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.
475. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics* **2021**, *52*, 8574–8586.
476. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
477. Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; Jia, J. Lisa: Reasoning segmentation via large language model. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9579–9589.
478. Gan, Y.; Park, S.; Schubert, A.; Philippakis, A.; Alaa, A.M. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390* **2023**.
479. Wu, J.; Zhong, M.; Xing, S.; Lai, Z.; Liu, Z.; Wang, W.; Chen, Z.; Zhu, X.; Lu, L.; Lu, T.; others. VisionLLM v2: An End-to-End Generalist Multimodal Large Language Model for Hundreds of Vision-Language Tasks. *arXiv preprint arXiv:2406.08394* **2024**.
480. Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; Chen, C. ControlNet

++

: Improving Conditional Controls with Efficient Consistency Feedback. European Conference on Computer Vision. Springer, 2025, pp. 129–147.
481. Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Wang, S.; Huang, T.; Liu, Z. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340* **2024**.
482. Soomro, K. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* **2012**.
483. Liu, Y.; Li, L.; Ren, S.; Gao, R.; Li, S.; Chen, S.; Sun, X.; Hou, L. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems* **2024**, *36*.
484. Fan, F.; Luo, C.; Zhan, J.; Gao, W. AIGCBench: Comprehensive Evaluation of Image-to-Video Content Generated by AI. *arXiv preprint arXiv:2401.01651* **2024**.

485. Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; Zhou, J. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145* **2023**.
486. Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* **2017**.
487. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
488. Saito, M.; Saito, S.; Koyama, M.; Kobayashi, S. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision* **2020**, *128*, 2586–2606.
489. Liu, Y.; Cun, X.; Liu, X.; Wang, X.; Zhang, Y.; Chen, H.; Liu, Y.; Zeng, T.; Chan, R.; Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22139–22149.
490. Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; others. Vbench: Comprehensive benchmark suite for video generative models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21807–21818.
491. Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; Lin, W. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20144–20154.
492. Wu, J.Z.; Fang, G.; Wu, H.; Wang, X.; Ge, Y.; Cun, X.; Zhang, D.J.; Liu, J.W.; Gu, Y.; Zhao, R.; others. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781* **2024**.
493. Lai, W.S.; Huang, J.B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.H. Learning blind video temporal consistency. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 170–185.
494. Lei, C.; Xing, Y.; Chen, Q. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems* **2020**, *33*, 1083–1093.
495. Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; Germanidis, A. Structure and content-guided video synthesis with diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7346–7356.
496. Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15932–15942.
497. Liao, M.; Lu, H.; Zhang, X.; Wan, F.; Wang, T.; Zhao, Y.; Zuo, W.; Ye, Q.; Wang, J. Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094* **2024**.
498. Yuan, S.; Huang, J.; Xu, Y.; Liu, Y.; Zhang, S.; Shi, Y.; Zhu, R.; Cheng, X.; Luo, J.; Yuan, L. ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation. *arXiv preprint arXiv:2406.18522* **2024**.
499. Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. FVD: A new metric for video generation **2019**.
500. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
501. Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* **2018**.
502. Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; Wong, T.T. Dynamicrafter: Animating open-domain images with video diffusion priors. European Conference on Computer Vision. Springer, 2025, pp. 399–417.
503. Yang, D.; Guo, H.; Wang, Y.; Huang, R.; Li, X.; Tan, X.; Wu, X.; Meng, H. UniAudio 1.5: Large Language Model-driven Audio Codec is A Few-shot Audio Task Learner. *arXiv preprint arXiv:2406.10056* **2024**.
504. Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; others. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407* **2024**.
505. Liu, W.; Guo, Z.; Xu, J.; Lv, Y.; Chu, Y.; Zhao, Z.; Lin, J. Analyzing and Mitigating Inconsistency in Discrete Audio Tokens for Neural Codec Language Models. *arXiv preprint arXiv:2409.19283* **2024**.

506. Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; others. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**.
507. Reddy, C.K.; Gopal, V.; Cutler, R. DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 886–890.
508. Kilgour, K.; Zuluaga, M.; Roblek, D.; Sharifi, M. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466* **2018**.
509. Shlens, J. Notes on kullback-leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000* **2014**.
510. Yuan, Y.; Liu, H.; Liang, J.; Liu, X.; Plumbley, M.D.; Wang, W. Leveraging pre-trained AudioLDM for sound generation: A benchmark study. 2023 31st European Signal Processing Conference (EUSIPCO). IEEE, 2023, pp. 765–769.
511. Agostinelli, A.; Denk, T.I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; others. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* **2023**.
512. Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; Défossez, A. Simple and controllable music generation. *Advances in Neural Information Processing Systems* **2024**, *36*.
513. Wu, S.L.; Donahue, C.; Watanabe, S.; Bryan, N.J. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2024**, *32*, 2692–2703.
514. Meng, L.; Zhou, L.; Liu, S.; Chen, S.; Han, B.; Hu, S.; Liu, Y.; Li, J.; Zhao, S.; Wu, X.; others. Autoregressive Speech Synthesis without Vector Quantization. *arXiv preprint arXiv:2407.08551* **2024**.
515. Chen, S.; Liu, S.; Zhou, L.; Liu, Y.; Tan, X.; Li, J.; Zhao, S.; Qian, Y.; Wei, F. VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2406.05370* **2024**.
516. Sun, P.; Cheng, S.; Li, X.; Ye, Z.; Liu, H.; Zhang, H.; Xue, W.; Guo, Y. Both Ears Wide Open: Towards Language-Driven Spatial Audio Generation. *arXiv preprint arXiv:2410.10676* **2024**.
517. Anastassiou, P.; Chen, J.; Chen, J.; Chen, Y.; Chen, Z.; Chen, Z.; Cong, J.; Deng, L.; Ding, C.; Gao, L.; others. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. *arXiv preprint arXiv:2406.02430* **2024**.
518. SpeechTeam, T. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs. *arXiv preprint arXiv:2407.04051* **2024**.
519. Chen, K.; Gou, Y.; Huang, R.; Liu, Z.; Tan, D.; Xu, J.; Wang, C.; Zhu, Y.; Zeng, Y.; Yang, K.; others. EMOVA: Empowering Language Models to See, Hear and Speak with Vivid Emotions. *arXiv preprint arXiv:2409.18042* **2024**.
520. Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2023**, *31*, 1720–1733.
521. Mei, X.; Liu, X.; Huang, Q.; Plumbley, M.D.; Wang, W. Audio captioning transformer. *arXiv preprint arXiv:2107.09817* **2021**.
522. Liu, X.; Iqbal, T.; Zhao, J.; Huang, Q.; Plumbley, M.D.; Wang, W. Conditional sound generation using neural discrete time-frequency representation learning. 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2021, pp. 1–6.
523. Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; Saruwatari, H. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152* **2022**.
524. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W.X.; Wen, J.R. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355* **2023**.
525. Hu, H.; Zhang, J.; Zhao, M.; Sun, Z. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301* **2023**.
526. Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; Fung, P. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338* **2023**.
527. Rohrbach, A.; Hendricks, L.A.; Burns, K.; Darrell, T.; Saenko, K. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156* **2018**.

528. Ding, Y.; Wang, Z.; Ahmad, W.; Ding, H.; Tan, M.; Jain, N.; Ramanathan, M.K.; Nallapati, R.; Bhatia, P.; Roth, D.; others. Crosscodeeval: A diverse and multilingual benchmark for cross-file code completion. *Advances in Neural Information Processing Systems* **2024**, *36*.
529. Jing, L.; Li, R.; Chen, Y.; Jia, M.; Du, X. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477* **2023**.
530. Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; Sang, J. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation, 2024, [arXiv:cs.CL/2311.07397].
531. Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.Y.; Wang, Y.X.; Yang, Y.; others. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525* **2023**.
532. Gunjal, A.; Yin, J.; Bas, E. Detecting and preventing hallucinations in large vision language models. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 18135–18143.
533. Wang, J.; Zhou, Y.; Xu, G.; Shi, P.; Zhao, C.; Xu, H.; Ye, Q.; Yan, M.; Zhang, J.; Zhu, J.; others. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126* **2023**.
534. Duan, J.; Yu, S.; Tan, H.L.; Zhu, H.; Tan, C. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence* **2022**, *6*, 230–244.
535. Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; Batra, D. Embodied question answering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1–10.
536. Gordon, D.; Kembhavi, A.; Rastegari, M.; Redmon, J.; Fox, D.; Farhadi, A. Iqa: Visual question answering in interactive environments. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4089–4098.
537. Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3674–3683.
538. Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; Lee, S. Beyond the nav-graph: Vision-and-language navigation in continuous environments. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. Springer, 2020, pp. 104–120.
539. Shi, T.; Karpathy, A.; Fan, L.; Hernandez, J.; Liang, P. World of bits: An open-domain platform for web-based agents. International Conference on Machine Learning. PMLR, 2017, pp. 3135–3144.
540. Rawles, C.; Li, A.; Rodriguez, D.; Riva, O.; Lillicrap, T. Androidinthewild: A large-scale dataset for android device control. *Advances in Neural Information Processing Systems* **2024**, *36*.
541. Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; Fox, D. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10740–10749.
542. Padmakumar, A.; Thomason, J.; Shrivastava, A.; Lange, P.; Narayan-Chen, A.; Gella, S.; Piramuthu, R.; Tur, G.; Hakkani-Tur, D. Teach: Task-driven embodied agents that chat. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 2017–2025.
543. Yenamandra, S.; Ramachandran, A.; Yadav, K.; Wang, A.; Khanna, M.; Gervet, T.; Yang, T.Y.; Jain, V.; Clegg, A.W.; Turner, J.; others. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565* **2023**.
544. Gupta, A.; Kumar, V.; Lynch, C.; Levine, S.; Hausman, K. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956* **2019**.
545. Mees, O.; Hermann, L.; Rosete-Beas, E.; Burgard, W. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters* **2022**, *7*, 7327–7334.
546. Padalkar, A.; Pooley, A.; Jain, A.; Bewley, A.; Herzog, A.; Irpan, A.; Khazatsky, A.; Rai, A.; Singh, A.; Brohan, A.; others. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864* **2023**.
547. Chen, H.; Suhr, A.; Misra, D.; Snavely, N.; Artzi, Y. Touchdown: Natural language navigation and spatial reasoning in visual street environments. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12538–12547.

548. Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W.Y.; Shen, C.; Hengel, A.v.d. Reverie: Remote embodied visual referring expression in real indoor environments. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9982–9991.
549. Fried, D.; Hu, R.; Cirik, V.; Rohrbach, A.; Andreas, J.; Morency, L.P.; Berg-Kirkpatrick, T.; Saenko, K.; Klein, D.; Darrell, T. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems* **2018**, *31*.
550. Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; others. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* **2022**.
551. Driess, D.; Xia, F.; Sajjadi, M.S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; others. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* **2023**.
552. Chaplot, D.S.; Gandhi, D.; Gupta, S.; Gupta, A.; Salakhutdinov, R. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155* **2020**.
553. Chaplot, D.S.; Salakhutdinov, R.; Gupta, A.; Gupta, S. Neural topological slam for visual navigation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 12875–12884.
554. Cartillier, V.; Ren, Z.; Jain, N.; Lee, S.; Essa, I.; Batra, D. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 964–972.
555. Cartillier, V.; Jain, N.; Essa, I. 3D Semantic MapNet: Building Maps for Multi-Object Re-Identification in 3D. *arXiv preprint arXiv:2403.13190* **2024**.
556. Hong, Y.; Zhou, Y.; Zhang, R.; Dernoncourt, F.; Bui, T.; Gould, S.; Tan, H. Learning navigational visual representations with semantic map supervision. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3055–3067.
557. Zhan, Z.; Yu, L.; Yu, S.; Tan, G. MC-GPT: Empowering Vision-and-Language Navigation with Memory Map and Reasoning Chains. *arXiv preprint arXiv:2405.10620* **2024**.
558. Chen, C.; Zhang, J.; Yang, K.; Peng, K.; Stiefelhagen, R. Trans4Map: Revisiting Holistic Bird's-Eye-View Mapping From Egocentric Images to Allocentric Semantics With Vision Transformers. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4013–4022.
559. Xiong, X.; Liu, Y.; Yuan, T.; Wang, Y.; Wang, Y.; Zhao, H. Neural map prior for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17535–17544.
560. Wang, Z.; Li, X.; Yang, J.; Liu, Y.; Jiang, S. Gridmm: Grid memory map for vision-and-language navigation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15625–15636.
561. Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; Fei-Fei, L. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973* **2023**.
562. Szot, A.; Schwarzer, M.; Agrawal, H.; Mazoure, B.; Metcalf, R.; Talbott, W.; Mackraz, N.; Hjelm, R.D.; Toshev, A.T. Large language models as generalizable policies for embodied tasks. The Twelfth International Conference on Learning Representations, 2023.
563. Zhou, G.; Hong, Y.; Wu, Q. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 7641–7649.
564. Zheng, D.; Huang, S.; Zhao, L.; Zhong, Y.; Wang, L. Towards learning a generalist model for embodied navigation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13624–13634.
565. Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; others. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* **2023**.
566. Li, X.; Liu, M.; Zhang, H.; Yu, C.; Xu, J.; Wu, H.; Cheang, C.; Jing, Y.; Zhang, W.; Liu, H.; others. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378* **2023**.
567. Team, O.M.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Kreiman, T.; Xu, C.; others. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213* **2024**.

568. Huang, J.; Yong, S.; Ma, X.; Linghu, X.; Li, P.; Wang, Y.; Li, Q.; Zhu, S.C.; Jia, B.; Huang, S. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871* **2023**.
569. Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; others. Do as i can, not as i say: Grounding language in robotic affordances. Conference on robot learning. PMLR, 2023, pp. 287–318.
570. Ma, X.; Yong, S.; Zheng, Z.; Li, Q.; Liang, Y.; Zhu, S.C.; Huang, S. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474* **2022**.
571. Wani, S.; Patel, S.; Jain, U.; Chang, A.; Savva, M. Multion: Benchmarking semantic map memory using multi-object navigation. *Advances in Neural Information Processing Systems* **2020**, *33*, 9700–9712.
572. Zhu, F.; Liang, X.; Zhu, Y.; Yu, Q.; Chang, X.; Liang, X. Soon: Scenario oriented object navigation with graph-based exploration. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12689–12699.
573. Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; Su, Y. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems* **2024**, *36*.
574. Zhang, J.; Wu, J.; Teng, Y.; Liao, M.; Xu, N.; Xiao, X.; Wei, Z.; Tang, D. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713* **2024**.
575. Lu, Q.; Shao, W.; Liu, Z.; Meng, F.; Li, B.; Chen, B.; Huang, S.; Zhang, K.; Qiao, Y.; Luo, P. GUI Odyssey: A Comprehensive Dataset for Cross-App GUI Navigation on Mobile Devices. *arXiv preprint arXiv:2406.08451* **2024**.
576. Zhang, J.; Yu, Y.; Liao, M.; Li, W.; Wu, J.; Wei, Z. UI-Hawk: Unleashing the Screen Stream Understanding for GUI Agents. *Preprints* **2024**, *manuscript/202408.2137*.
577. Liu, X.; Zhang, T.; Gu, Y.; Iong, I.L.; Xu, Y.; Song, X.; Zhang, S.; Lai, H.; Liu, X.; Zhao, H.; Sun, J.; Yang, X.; Yang, Y.; Qi, Z.; Yao, S.; Sun, X.; Cheng, S.; Zheng, Q.; Yu, H.; Zhang, H.; Hong, W.; Ding, M.; Pan, L.; Gu, X.; Zeng, A.; Du, Z.; Song, C.H.; Su, Y.; Dong, Y.; Tang, J. VisualAgentBench: Towards Large Multimodal Models as Visual Foundation Agents, 2024, [arXiv:cs.AI/2408.06327].