# From Specific-MLLMs to Omni-MLLMs: A Survey on MLLMs Aligned with Multi-modalities

**Shixin Jiang[1], Jiafeng Liang[1], Jiyuan Wang[1], Xuan Dong[1], Heng Chang[3]**
**Weijiang Yu[3], Jinhua Du[4], Ming Liu[1,2*], Bing Qin[1,2]**
[1]Harbin Institute of Technology, Harbin, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Huawei Inc., Shenzhen, China
[4]Huawei London Research Centre, London, UK
{sxjiang, jfliang, jywang, mliu, qinb}@ir.hit.edu.cn

## Abstract

To tackle complex tasks in real-world scenarios, more researchers are focusing on Omni-MLLMs, which aim to achieve omni-modal understanding and generation. Beyond the constraints of any specific non-linguistic modality, Omni-MLLMs map various non-linguistic modalities into the embedding space of LLMs and enable the interaction and understanding of arbitrary combinations of modalities within a single model. In this paper, we systematically investigate relevant research and provide a comprehensive survey of Omni-MLLMs. Specifically, we first explain the four core components of Omni-MLLMs for unified multi-modal modeling with a meticulous taxonomy that offers novel perspectives. Then, we introduce the effective integration achieved through two-stage training and discuss the corresponding datasets as well as evaluation. Furthermore, we summarize the main challenges of current Omni-MLLMs and outline future directions. We hope this paper serves as an introduction for beginners and promotes the advancement of related research. Resources have been made publicly available at https://github.com/threegold116/Awesome-Omni-MLLMs.

## 1 Introduction

The remarkable performance of continuously evolving Multi-modal Large Language Models (MLLMs) has pointed to a possible direction for achieving general artificial intelligence (Bubeck et al., 2023; OpenAI, 2023b). MLLMs extend Large Language Models (LLMs) by integrating them with pre-trained models tailored to specific modalities, such as Vision-MLLMs (Liu et al., 2023c; Wang et al., 2024b; Sun et al., 2024c), Audio-MLLMs (Zhang et al., 2023a; Chu et al., 2023), and 3D-MLLMs (Xu et al., 2024b). However, these modality-specific MLLMs (Specific-MLLMs) are insufficient to tackle complex tasks in real-world scenarios that simultaneously involve multiple modalities. Therefore, efforts are being made to expand the range of modalities for understanding and generation, giving rise to the omni-modality MLLMs (Omni-MLLMs).

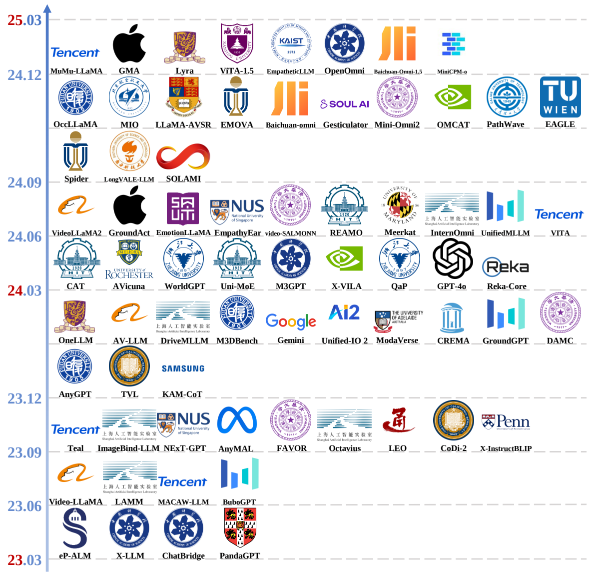


Figure 1: This timeline underscores the swift advancements in the development of Omni-MLLMs.

By integrating multiple pre-trained models of more non-linguistic modalities (Radford et al., 2021, 2023; Xue et al., 2024b; Rombach et al., 2022; Liu et al., 2023b), Omni-MLLMs expand the modalities for understanding and even generation based on Specific-MLLMs. Omni-MLLMs leverage the emergent capabilities of LLMs to treat various non-linguistic modalities as different *foreign languages*, enabling the interaction and understanding of information across different modalities within a unified space (Chen et al., 2023a; Panagopoulou et al., 2024). Compared to Specific-MLLMs, Omni-MLLMs can perform multiple unimodal understanding and generation tasks, as well

*Corresponding author



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 8617–8652
July 27 - August 1, 2025 

as cross-modal[1] tasks across two or more non-linguistic modalities, allowing a single model to handle arbitrary combinations of modalities.

A review of the development of Omni-MLLMs reveals that it has been continuously expanding in three directions. On the one hand, the types of modalities processed by Omni-MLLM have been continuously increasing, from X-LLMs that handle vision and audio to X-InstructBLIP (Panagopoulou et al., 2024) which adds 3D modality capabilities, PandaGPT (Su et al., 2023) that incorporates IMU modality, and finally One-LLM (Han et al., 2024a), which processes eight different modalities simultaneously. On the other hand, the ability to interact across modalities of Omni-MLLMs has also expanded, from the joint 3D-Image and Audio-Image cross-modal reasoning capability in ImageBind-LLM (Han et al., 2023) to the cross-modal generation capability of CoDi-2 that leverages interleaved audio and image contexts to generate both audio and images (Tang et al., 2024b). The Omni-MLLM is thus trending towards an "Any-to-Any" model. Besides, the application scenarios of Omni-MLLMs have been broadened, encompassing real-time multimodal speech interaction like Mini-Omni2 and Lyra (Xie and Wu, 2024; Zhong et al., 2024), world simulation like WordGPT (Ge et al., 2024b), multi-sensor autonomous driving like DriveMLM (Wang et al., 2023b), etc. In addition to the open-source models, there are also some closed-source Omni-MLLMs such as GPT-4o (OpenAI), Gemini (Reid et al., 2024), and Reka (Ormazabal et al., 2024). The timeline of Omni-MLLMs is shown in Figure 1. Despite the emergence of numerous Omni-MLLMs, there is still a lack of systematic evaluation and analysis.

To fill the gap, we propose this work to conduct a comprehensive and detailed analysis of Omni-MLLMs. We first review the architecture of Omni-MLLMs in four parts (§2). Next, we summarize how Omni-MLLMs expand across multiple modalities through the two-stage training process (§3); then present the training data construction and performance evaluation (§4). Furthermore, we highlight some key challenges and future directions (§5). Finally, we provide a brief summary (§6) and discuss related surveys in the Appendix A.

Our contributions can be summarized as follows:

[1]"Uni" and "Cross" refer to the number of non-linguistic modalities involved in the interaction, in contrast to "multi-modal reasoning," traditionally reserved for vision-language tasks (Panagopoulou et al., 2024).

(1) ***Comprehensive Survey***: This is the first comprehensive survey dedicated for Omni-MLLMs; (2) ***Meticulous taxonomy***: We introduce a meticulous taxonomy (shown in Figure 2); (3) ***Challenges and Future***: We outline the challenges of Omni-MLLMs and shed light on future research.

## 2 Omni-MLLM Architecture

As the extension of Specific-MLLMs, Omni-MLLMs inherit the architecture of *encoding, alignment, interaction, and generation* and broaden the types of non-linguistic modalities involved. This section introduces the implementation methods and functions of the four components in Omni-MLLM: Multi-modalities Encoding (§2.1), Multi-modalities Alignment (§2.2), Multi-modalities Interaction (§2.3), and Multi-modalities Generation (§2.4). More details about the architecture of Omni-MLLMs are shown in Appendix B.

### 2.1 Multi-modalities Encoding

Based on the encoding feature spaces of multiple modalities, we categorize the Omni-MLLM encoding methods into three types: 1) continuous encoding, 2) discrete encoding, and 3) hybrid encoding.

#### 2.1.1 Continuous Encoding

Continuous encoding refers to encoding the modality into the continuous feature space. Omni-MLLMs that adopt continuous encoding, such as X-LLM (Chen et al., 2023a) and ChatBridge (Zhao et al., 2023b), often integrate multiple pre-trained uni-modality encoders. These modality-specific encoders encode different modalities $\boldsymbol{X}$ into distinct feature spaces $\mathbb{R}_x$ as $\mathbf{F}_x$, formulated as:

$$\mathbf{F}_x = \text{SpecificEncoder}\left(\boldsymbol{X}\right), \quad \mathbf{F}_x \in \mathbb{R}_x \quad (1)$$

where SpecificEncoder refers to different modality-specific encoders used in Omni-MLLMs, such as InternVit (Chen et al., 2023g) for encoding visual modality, Whisper (Radford et al., 2023) for encoding auditory modality, ULIP-2 (Xue et al., 2024b) for encoding 3D modality, IMU2CLIP (Moon et al., 2022) for encoding IMU modality, etc.

Besides using heterogeneous encoders for continuous encoding, some Omni-MLLMs (Han et al., 2024a, 2023; Su et al., 2023; Fu et al., 2024c) employ pre-aligned encoders for multiple modalities, encoding different modalities $\boldsymbol{X}$ into the same feature space $\mathbb{R}_{uni}$, as shown in Equation 2.

$$\mathbf{F}_x = \text{PreAlignEncoder}\left(\boldsymbol{X}\right), \quad \mathbf{F}_x \in \mathbb{R}_{uni} \quad (2)$$

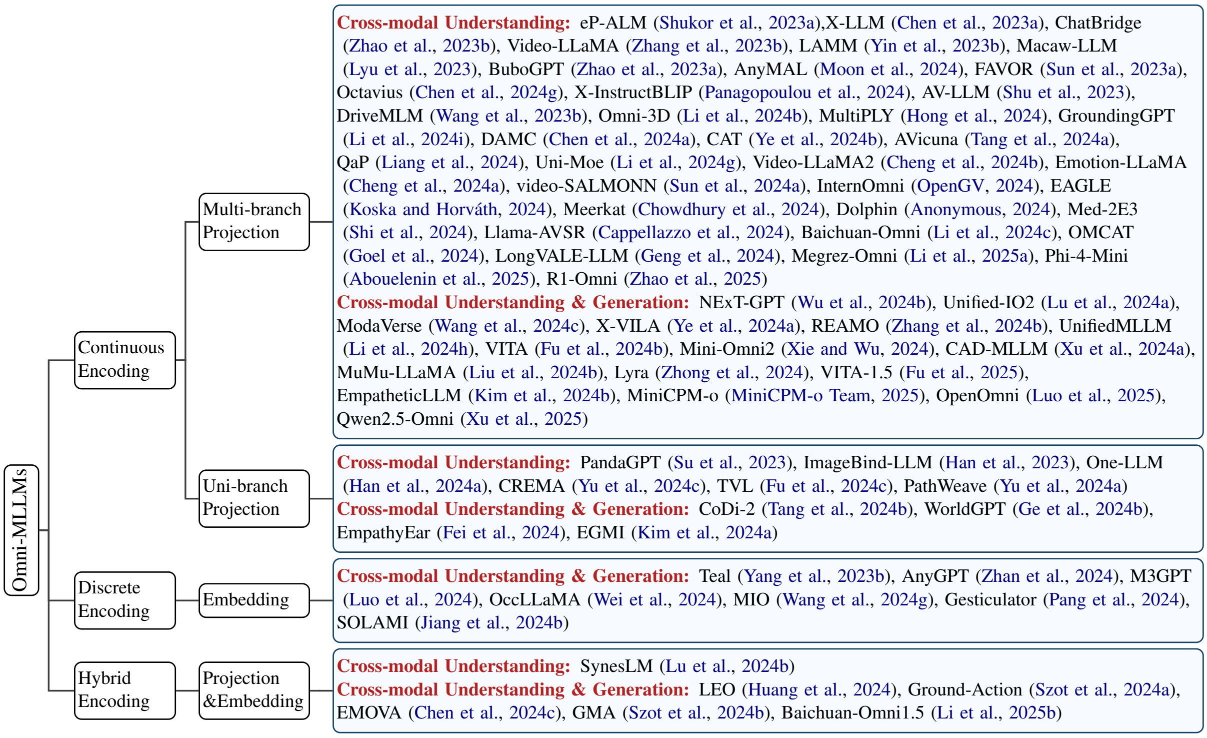


Figure 2: The taxonomy of Omni-MLLMs is organized according to their encoding strategies (§2.1), alignment methods (§2.2), interaction modes (§2.3), and generative capabilities (§2.4).

where $\mathrm{PreAlignEncoder}$ refer to encoders that uniformly encode multiple modalities, such as LanguageBind (Zhu et al., 2024a) which uses text as a bridge to align different modalities, and ImageBind (Girdhar et al., 2023) which uses images as a bridge to align different modalities.

#### 2.1.2 Discrete Encoding

To better facilitate the seamless integration and generation of new non-linguistic modalities, some Omni-MLLMs, such as AnyGPT (Zhan et al., 2024) and Teal (Yang et al., 2023b), adopt a discrete encoding approach. This method encodes different raw modalities $\boldsymbol{X}$ into the same discrete token space $\mathbb{V}_{uni}$ as $\mathbf{T}_x$, formulated as follows:

$$\mathbf{T}_x = \mathrm{SpecificTokenizer}\left(\boldsymbol{X}\right), \quad \mathbf{T}_x \in \mathbb{V}_{uni} \quad (3)$$

where $\mathrm{SpecificTokenizer}$ refers to different modality-specific tokenizers used in Omni-MLLMs, including the SEED tokenizer (Ge et al., 2024a) based on Vector Quantized Tokenization (VQ), the SpeechTokenizer (Zhang et al., 2023c) based on Residual Vector Quantized Tokenization (RVQ), the AudioTokenizer of Teal (Yang et al., 2023b) based on k-means clustering, and so on.

#### 2.1.3 Hybrid Encoding

Although discrete encoding facilitates the unified processing of different non-linguistic modalities and text compared to continuous encoding, discrete modality tokens often struggle to capture the detailed information inherent in raw continuous modalities (Chen et al., 2024c; Xie and Wu, 2024). Therefore, some Omni-MLLMs combine both encoding approaches instead of a fully discretized manner, choosing different encoding methods for different modalities. For instance, EMOVA (Chen et al., 2024c) uses the discrete S2U tokenizer to encode auditory modalities while employing the continuous encoder InternVit for visual modalities to retain more vision semantic information. Similarly, GroundAction (Szot et al., 2024a) encodes visual modalities using the CLIP Vit and action modalities with its trained action tokenizer.

### 2.2 Multi-modalities Alignment

Omni-MLLMs align the encoded features of various non-linguistic modalities with the embedding space of LLMs. The multi-modality alignment can be categorized into two approaches: 1) projection alignment and 2) embedding alignment.

### 2.2.1 Projection Alignment

The continuous encoding Omni-MLLMs insert adapters, referred to as *projectors*, between the encoders and the LLMs. These projectors map the continuously encoded modality features $\mathbf{F}_x$ into the text embedding space as $\mathbf{F}_p$. As discussed in Section 2.1.1, $\mathbf{F}_x$ may either reside in distinct feature spaces $\mathbb{R}_x$ or share the same feature space $\mathbb{R}_{uni}$. For the former, multiple projectors are typically employed to align the $\mathbf{F}_x$ of each modality into $\mathbb{R}_t$ as $\mathbf{F}_p$ independently, addressing dimensional mismatch and feature misalignment across modalities (Ye et al., 2024a; Lyu et al., 2023; Moon et al., 2024), formulated as follows:

$$\mathbf{F}_p = \mathrm{SpecificProjector}(\mathbf{F}_x),\ \mathbf{F}_p \in \mathbb{R}_t \quad (4)$$

where SpecificProjector refers to the modality-specific projector corresponding to different modalities, called *multi-branch projection*.

For the latter case, besides the multi-branch approach, Omni-MLLMs like PandaGPT (Su et al., 2023) and WorldGPT (Ge et al., 2024b) adopt a shared projector to achieve unified alignment across modalities to reduce the parameters of multiple projectors, as shown in Equation 5.

$$\mathbf{F}_p = \mathrm{UnifiedProjector}(\mathbf{F}_x),\ \mathbf{F}_p \in \mathbb{R}_t \quad (5)$$

where UnifiedProjector refers to the unified projector used to align multiple modalities, a design known as the *uni-branch projection*. A comparison of the two approaches is illustrated in Figure 3.

In terms of the ***specific implementation*** of the projector, the most straightforward approach is to use a multi-layer perceptron (MLP) or a single linear layer (Wu et al., 2024b; Cheng et al., 2024b; OpenGV, 2024). Alternatively, attention mechanisms can be employed to compress the encoded information of non-linguistic modalities. This includes cross-attention-based methods like Q-Former (Panagopoulou et al., 2024; Chen et al., 2023a) and Perceiver (Zhao et al., 2023b; Liang et al., 2024), as well as self-attention-based methods such as UPM in OneLLM (Han et al., 2024a). Additionally, BaiChuan-Omni (Li et al., 2024c) and EMOVA (Chen et al., 2024c) incorporate CNNs to compress the projected features, thereby achieving locality preservation (Cha et al., 2024).

It is also worth noting that in multi-branch Omni-MLLMs, different branches may utilize distinct implementations to better accommodate the unique characteristics of each modality (Li et al., 2024i).

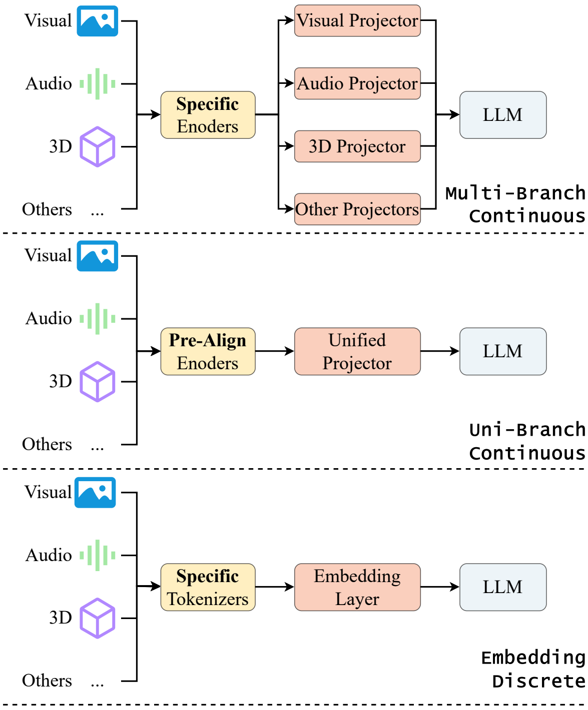


Figure 3: Three representative structures[2] for encoding and alignment in Omni-MLLM, differing in their encoding spaces and alignment mechanisms.

For example, Uni-MoE (Li et al., 2024g) uses a linear projection for the visual modality and a Q-Former for the auditory modality. Meanwhile, uni-branch Omni-MLLMs, when using an attention-based projector, typically design multiple modality-specific learnable vectors to extract key information from various non-linguistic modalities (Yu et al., 2024a; Han et al., 2024a; Yu et al., 2024c).

### 2.2.2 Embedding Alignment

As for discrete encoding Omni-MLLMs, the features of non-linguistic modalities are represented as quantized codes, which reside in the same discrete space $\mathbb{V}_{uni}$ as text tokens. Therefore, new modality-specific discrete tokens $\mathbf{T}_x$ are embedded into the continuous feature space $\mathbb{R}_t$ by modifying the vocabulary of LLMs and the corresponding embeddings layer, as shown in Equation 6.

$$\mathbf{F}_p = \mathrm{Embedding}(\mathbf{T}_x),\ \mathbf{F}_p \in \mathbb{R}_t \quad (6)$$

where Embedding refers to the unified embedding layer corresponding to different modalities, which is typically achieved by adding discrete codebooks from various modalities to the vocabulary and expanding the embedding layer of LLMs (Zhan et al.,

[2]Multi-branch structures can be integrated with unified encoders as well (Wu et al., 2024b; Ye et al., 2024a).

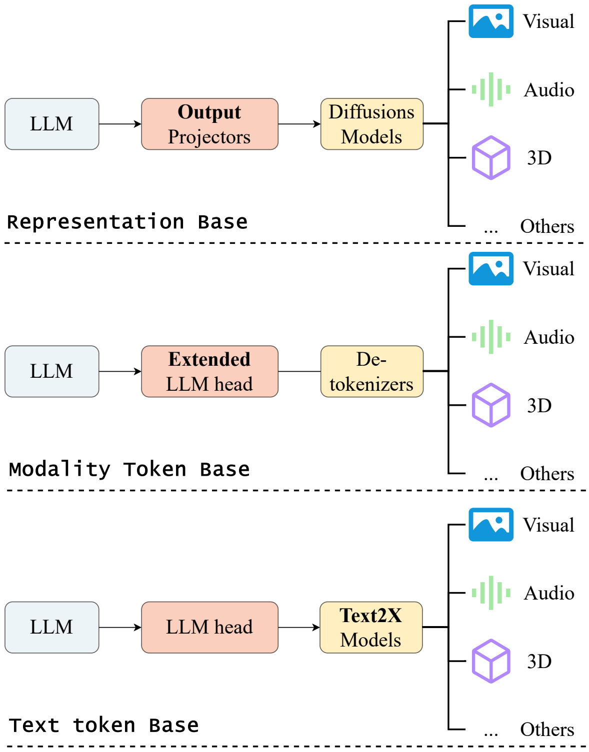


Figure 4: Three generation methods in Omni-MLLMs are implemented based on different output spaces of LLM and the corresponding generative models.

2024; Yang et al., 2023b; Wei et al., 2024). For instance, AnyGPT extends the vocabulary of the LLaMA-2 by incorporating 17,408 codes across three modalities—image, speech, and music (Zhan et al., 2024). Besides, some works like Ground-Action (Szot et al., 2024a) and LEO (Huang et al., 2024) overwrite infrequently used tokens in the original vocabulary for alignment, as they extend a smaller set of modality-specific discrete tokens.

Additionally, for hybrid encoding models, alignment is achieved by simultaneously employing both the projection method and the embedding method (Chen et al., 2024c; Szot et al., 2024b).

## 2.3 Multi-modalities Interaction

Omni-MLLMs utilize transformer-based LLMs to facilitate information interaction between different modalities within a unified feature space $\mathbb{R}_t$. Commonly used LLMs include the LLaMA series (Touvron et al., 2023), the Qwen series (Bai et al., 2023), and others (Cai et al., 2024; Zeng et al., 2024).

For interaction, most Omni-MLLMs (Chen et al., 2023a; Ye et al., 2024a; OpenGV, 2024; Li et al., 2025b) concatenate aligned non-linguistic modality features $\mathbf{F}_p$ with textual features $\mathbf{F}_t$ at the input level, enabling interaction in a progressive and layer-by-layer manner. Meanwhile, some works, such as ImageBind-LLM (Han et al., 2023) and TVL-LLaMA (Fu et al., 2024c), insert $\mathbf{F}_p$ into specific layers or all layers of the LLMs to mitigate the loss of modality information (Shukor et al., 2023a).

In terms of the number of modalities involved in interactions, compared to Specific-MLLMs that are limited to dual-modal interactions between a single non-linguistic modality and text (Liu et al., 2023c; Xu et al., 2024b), Omni-MLLMs not only support multiple dual-modal interactions but also enable omni-multimodal interactions involving more than two non-linguistic modalities (Zhao et al., 2023b; Wang et al., 2024g). For example, X-InstructBLIP (Panagopoulou et al., 2024) enables dual-modal interactions such as vision-text, audio-text, and 3D-text, as well as omni-modal interactions like vision-audio-text and 3D-vision-text, showcasing the ability of Omni-MLLMs to handle arbitrary combinations of modalities.

## 2.4 Multi-modalities Generation

Omni-MLLMs can output text while also generating non-linguistic modalities by integrating different generation models. As shown in Figure 4, we categorize multi-modalities generation into three types: text-based generation, representation-based generation, and modality-token-based generation.

**Text-based** This approach directly utilizes the discrete text output from the LLM to invoke Text-to-X generation models (Liu et al., 2023b; Luo et al., 2023c; Brooks et al., 2023) based on the content of the text. For example, VITA (Fu et al., 2024b) employs TTS tools (RVC-Boss) to convert the output text into corresponding speech, while ModelVerse (Wang et al., 2024c) and UnifiedM-LLM (Li et al., 2024h) use the text to specify the generation model and utilize the corresponding descriptions to generate different modalities.

**Modality-Token-based** Works like MiniOmni-2 (Xie and Wu, 2024) and AnyGPT (Zhan et al., 2024) extend the corresponding LLM head with codebooks from different modality tokenizers to generate modality-specific discrete tokens. These tokens are then decoded using the corresponding de-tokenizers (Esser et al., 2021; Yu et al., 2024b; Zeghidour et al., 2022; Dhariwal et al., 2020) to produce various modalities.

**Representation-based** To alleviate the potential noise introduced by discrete tokens, works like X-

VILA (Ye et al., 2024a) and NextGPT (Wu et al., 2024b) incorporate modality-specific signal tokens into the vocabulary. They then use transformers or MLPs to map the signal token representations into the ones that are understandable to the multimodal decoders, typically off-the-shelf latent-conditioned diffusion models (Rombach et al., 2022; Tang et al., 2023; Xue et al., 2024a; Blattmann et al., 2023), enabling effective generation capabilities.

## 3 Omni-MLLM Training

To achieve alignment across different vector spaces and improve instruction-following ability under arbitrary modality settings, Omni-MLLMs extend the standard two-stage training pipeline of Specific-MLLMs: *multi-modalities alignment pre-training* and *multi-modalities instruction fine-tuning*.

### 3.1 Multi-modalities Alignment Pre-training

Multi-modalities alignment pre-training involves *input alignment* training between the feature spaces of different modalities and the embedding space of LLMs on the encoding side, as well as *output alignment* training between the embedding space and the input spaces of various modality decoders on the decoding side. Input alignment and output alignment can be carried out separately (Wu et al., 2024b) or simultaneously (Ye et al., 2024a).

#### 3.1.1 Input Alignment

Input alignment mainly uses X-Text paired datasets of different modalities and minimizes the text generation loss of the corresponding description text to optimize. In this phase, continuous encoding Omni-MLLMs normally update parameters of projectors, while discrete encoding Omni-MLLMs adjust the parameters of the embedding layer.

In terms of ***training order*** of different modalities alignment, multi-branch Omni-MLLM performs separate alignment training for each modality-specific projector, directly aligning each non-linguistic modality with text and using text as a bridge to align different non-linguistic modalities (Zhao et al., 2023b; Panagopoulou et al., 2024). The uni-branch Omni-MLLM, on the other hand, uses the unified projector for different modalities, which may lead to interference in the alignment performance between different modalities. Thus, Han et al. (2024a) employ a progressive alignment strategy to align multiple modalities in a specific order. In contrast, discrete encoding Omni-MLLMs, like AnyGPT (Zhan et al., 2024) and M3GPT (Luo et al., 2024), mix the alignment data from different modalities and perform alignment simultaneously.

Besides, in addition to directly leveraging X-Text paired datasets from different modalities for direct alignment, PandaGPT (Su et al., 2023), ImageBind-LLM (Han et al., 2023), and VideoL-LaMA (Zhang et al., 2023b) utilize the pre-aligned modality feature space $\mathbb{V}_{uni}$ to achieve indirect alignment between other non-linguistic modalities and text by training solely on Image-Text data.

#### 3.1.2 Output Alignment

The training of output alignment typically utilizes the same X-text paired dataset as input alignment and adheres to the identical training sequence. Meanwhile, the training objectives for output alignment vary depending on the multi-modality generation methods in Section 2.4. Token-based generative Omni-MLLMs optimize the extended LLM head by minimizing the text generation loss associated with modality-specific discrete tokens (Lu et al., 2024a; Wei et al., 2024). Representation-based generative Omni-MLLMs generally optimize their output projectors by minimizing the composite loss comprising three components (Xie and Wu, 2024; Yang et al., 2023b): 1) the text generation loss of signal tokens; 2) the L2 distance between the output representation and the condition vector of the corresponding decoder, i.e. MSE loss; and 3) the conditional latent denoising loss (Rombach et al., 2022). For text-based generative Omni-MLLMs, as there is no additional output structure, the output alignment training is generally not required (Wang et al., 2024c; Li et al., 2024h).

### 3.2 Multi-modalities Instruction Fine-tuning

The instruction fine-tuning phase aims to enhance generalization capability under arbitrary modalities of Omni-MLLMs (Panagopoulou et al., 2024; Wu et al., 2024b; Ye et al., 2024a). Instruction fine-tuning primarily utilizes instruction-following datasets and computes the text generation loss for the corresponding responses to optimize. For models with generation capabilities, the loss mentioned in section 3.1.2 may also be incorporated. During this phase, Omni-MLLMs further perform full-scale tuning of the LLM parameters (Han et al., 2024a; Cheng et al., 2024b) or use PEFT techniques (Han et al., 2024b), such as LoRA (Hu et al., 2022), for partial tuning (Wang et al., 2024e).

Compared to Specific-MLLMs, Omni-MLLMs not only leverage multiple uni-modal instruction data of different modalities for training but also use cross-modal instruction data to enhance their cross-modal ability (Ye et al., 2024a; Zhan et al., 2024; Li et al., 2024c). In addition to directly mixing different instruction data for training (Panagopoulou et al., 2024; Ye et al., 2024a), some works like Uni-Moe (Li et al., 2024g) and Lyra (Zhong et al., 2024) adopt a multi-step fine-tuning approach, introducing different uni-modal and cross-modal instruction data in a specific order for training to gradually enhance their uni-modal and cross-modal ability.

### 3.3 Other Train Recipes

In addition to the general training paradigms mentioned in Section 3, some other useful training recipes are also used. (1) **Prior knowledge from Specific-MLLMs**: Since Specific-MLLMs have already achieved effective alignment in single-modal scenarios, some Omni-MLLMs directly leverage their well-trained projectors to reduce the training overhead during the alignment phase. For example, InstructBLIP (Panagopoulou et al., 2024) and X-LLM (Chen et al., 2023a) use the Q-former trained by BLIP2 to align the visual modality, while NaviveMC and DAMC (Chen et al., 2024a) further leverage projectors from multiple models to handle alignment for visual, audio, and 3D modalities separately; (2) **Additional human preference training**: Szot et al. (2024b) and Ye et al. (2024b) adopt HF training methods like PPO and ADPO to better align with human preferences; (3) **Modalities Blending**: During progressive alignment pre-training or multi-step instruction fine-tuning, some works (Han et al., 2024a; Li et al., 2024c; Chen et al., 2024c) mix previously trained modality data with the current new modality data for training to prevent catastrophic forgetting.

## 4 Data Construction and Evaluation

This section summarizes the construction of modality alignment data and instruction data used in the Omni-MLLM training process (§4.1), as well as the evaluation across four different capabilities (§4.2).

### 4.1 Training Data

**Alignment Data** Omni-MLLMs leverage caption datasets from various modalities to construct X-Text paired data for alignment pre-training, such as the WebVid (Bain et al., 2021) for visual modality and the AudioCaps (Kim et al., 2019) for auditory modality. However, for data-scarce modalities like depth maps and thermal maps, large-scale text-paired data is lacking (Zhu et al., 2024a; Girdhar et al., 2023). To address this, synthetic methods that use DPT models (Ranftl et al., 2021; Bhat et al., 2023; Xu et al., 2023) or image translation models (Lee et al., 2023) to convert image-text pairs into other modality text pairs are widely employed (Han et al., 2024a; Chen et al., 2024c; Zhu et al., 2024a). Moreover, interleaved datasets (Zhu et al., 2023a) are used for alignment pre-training in some works (Tang et al., 2024b) to enhance the contextual understanding capability.

**Instruction Data** Omni-MLLMs not only leverage uni-modal instruction datasets from Specific-MLLMs, but also construct cross-modal instruction data through diverse methods as follows.
(1) **Template-based Construction**: Most works (Sun et al., 2023a; Zhao et al., 2023b; Zhang et al., 2024b) utilize cross-modal downstream datasets (Sanabria et al., 2018; Chen et al., 2020b) combined with predefined templates to construct cross-modal instructions; (2) **GPT Generation**: Following the paradigm of LLaVA (Liu et al., 2023c), some Omni-MLLMs (Lyu et al., 2023; Zhao et al., 2023b) leverage the labels from the annotated dataset (Lin et al., 2014; Bain et al., 2021) or use pre-trained models like SAM (Chen et al., 2023c) and GRIT (Wu et al., 2024a) to extract meta-information (e.g., captions and object categories) of different modalities. Then they employ powerful LLMs (OpenAI, 2023b,a) to generate cross-modal instructions based on the obtained meta-information; (3) **T2X Generation**: Li et al. (2024g) use TTS tools to convert the Image-Text2Text uni-modal instructions from LLaVA-v1.5 (Liu et al., 2024a) into Image-Speech-Text2Text cross-modal instructions. AnyGPT (Zhan et al., 2024) and NextGPT (Wu et al., 2024b) leverage Text2X models such as DALL-E-3 (Shi et al., 2020) and MusicGen (Copet et al., 2023) to convert the GPT-generated pure text instructions into Xs2Xs cross-modal instructions. Details about training data are shown in Appendix C.1

### 4.2 Benchmark

We provide a brief overview of the benchmarks used to evaluate Omni-MLLMs. The statistics of the benchmarks are shown in Appendix C.2.

**Uni-modal Understanding** Uni-modal understanding assesses the ability of Omni-MLLMs to comprehend and reason on different non-linguistic modalities, including downstream X-Text2Text datasets such as X-Caption (Plummer et al., 2015; Xu et al., 2016), X-QA (Goyal et al., 2017; Xu et al., 2017a), and X-Classification (Deitke et al., 2023), as well as comprehensive multi-task benchmarks (Liu et al., 2024d; Fu et al., 2024a, 2023).

**Uni-modal Generation** Uni-modal generation aims to evaluate the ability of Omni-MLLMs to generate a single non-linguistic modality, including the Text2X generation task (Kim et al., 2019; Ruiz et al., 2023) and the Text-X2Text editing task (Veaux et al., 2017; Perazzi et al., 2016).

**Cross-modal Understanding** Cross-modal understanding evaluates the ability of Omni-MLLMs to jointly comprehend and reason across multiple non-linguistic modalities like Image-Speech-Text2Text (Li et al., 2024g; OpenGV, 2024), Video-Audio-Text2Text (Li et al., 2022a,b), and Image-3D-Text2Text (Panagopoulou et al., 2024).

**Cross-modal Generation** Cross-modal generation further evaluates the ability of Omni-MLLMs to generate non-linguistic modalities in conjunction with other non-linguistic modality inputs. For example, the Xs-Text2X benchmark proposed by X-VILA (Ye et al., 2024a) includes tasks such as Image-Text2Audio and Image-Audio-Text2Video.

## 5 Challenges and Future Directions

Despite Omni-MLLMs having showcased remarkable performance on numerous tasks, there are still some challenges that necessitate further research.

### 5.1 Expansion of modalities

Most Omni-MLLMs can only process 2-3 types of non-linguistic modalities, and they still face several challenges when expanding more modalities.

**Training efficiency** The common method that introduces new modalities through additional alignment pre-training and instruction fine-tuning can lead to significant training cost. Leveraging prior knowledge from Specific-MLLMs (Panagopoulou et al., 2024; Chen et al., 2024a) or using pre-aligned encoders for indirect alignment (Han et al., 2023; Su et al., 2023) can help reduce training overhead but may impact cross-modal performance.

**Catastrophic forgetting** Expanding new modalities may adjust the shared parameters, potentially causing catastrophic forgetting of previously trained modalities knowledge (Yu et al., 2024a). This issue can be partially mitigated by mixing trained modality data (Han et al., 2024a; Li et al., 2024g) or fine-tuning only the modality-specific parameters (Yu et al., 2024a,c), but both approaches make the training process more complex.

**Low-resource modalities** Although the data synthesis method in Section 4.1 can help alleviate the lack of text-paired data and instruction data for low-resource modalities (Han et al., 2024a; Jiang et al., 2024c), the absence of real modality may lead to biases in understanding of that modality.

### 5.2 Cross-modal capabilities

The Omni-MLLMs have achieved promising performance in cross-modal understanding and generation tasks, but there are still some challenges.

**Long Context** When the input contains multiple sequence modalities (video, speech...), the length of the multi-modalities token sequence may exceed the context window of LLMs and lead to memory overflow. While methods such as token compressing (Yu et al., 2024c; Li et al., 2024c) or token sampling (Zhan et al., 2024; Zhong et al., 2024) can reduce the number of input tokens, they also result in a decline in cross-modal performance.

**Modality Bias** Due to the imbalance in training data volume and the performance disparity among different modality encoders, Omni-MLLMs may tend to pay attention to the dominant modality while neglecting information from other modalities during cross-modal inference. Balancing the data volume across modalities or enhancing the corresponding modality-specific modules could potentially help mitigate this issue (Leng et al., 2024).

**Temporal Alignment** When dealing with different modalities that have temporal dependencies, retaining their temporal alignment information is crucial for subsequent cross-modal understanding. Some attempts have been made to preserve the temporal alignment information between audio and video, such as interleaved modality-specific tokens of video and audio (Tang et al., 2024a) and inserting the time-related special tokens into the multi-modalies tokens (Goel et al., 2024).

**Data and Benchmark** Although Omni-MLLMs employ various methods in Section 4.1 to generate cross-modal instruction data, there is still significant room for improvement and expansion, including enhancing the diversity of instructions, incorporating longer contextual dialogues, and exploring more diverse modality interaction paradigms. Similarly, cross-modal benchmarks such as OmniBench (Li et al., 2024f) and OmniR (Chen et al., 2024d) still fall short in terms of task richness and instruction diversity when compared to uni-modal benchmarks like MMMU (Yue et al., 2024) and MME (Fu et al., 2023). And the variety of modalities they cover is also relatively limited.

### 5.3 Application scenarios

The emergence of Omni-MLLM brings new opportunities and possibilities for various applications. **(1) Real-time Multi-modalities Interaction**: Fu et al. (2025) and Xie and Wu (2024) achieve robust capabilities in both vision and speech understanding, enabling efficient speech-to-speech interactions with vision in real-time. **(2) Comprehensive Planning**: Wang et al. (2023b) and Szot et al. (2024a) leverage the complementarity across multiple modalities to achieve better path planning and action planning capabilities than planning with vision information only. **(3) World Simulator**: Ge et al. (2024b) not only understands and generates different modalities but also predicts state transitions for any combination of modalities.

## 6 Conclusion

In this paper, we provide a comprehensive survey report on Omni-MLLM, offering a comprehensive review of the field. Specifically, we break down Omni-MLLM into four key components and categorize them based on modal encoding and alignment methods. Subsequently, we provide a detailed summary of the training process of Omni-MLLM and the related resources used. We also summarize the current challenges and the future development directions. This paper is the first systematic survey dedicated to Omni-MLLMs. We hope this survey will facilitate further research in this area.

## Limitations

This study provides the first comprehensive survey of Omni-MLLMs. Related work, architecture statistics, more details of training and evaluation, as well as other training recipes, can be found in Appendix A,B,C.

We have made our best effort, but there may still be some limitations. On one hand, due to page limitations, we can only provide a concise overview of the core contributions of mainstream Omni-MLLMs, rather than exhaustive technical details. On the other hand, our review primarily covers research from *ACL, NeurIPS, ICLR, ICML, COLING, CVPR, IJCAI, ECCV, and arXiv, and there is a chance that we may have missed some important work published in other venues. We will stay updated with ongoing discussions in the research community and plan to revise our work in the future to include overlooked contributions.


## Acknowledgments

The research in this article is supported by the National Science Foundation of China (U22B2059, 62276083). We also appreciate the support from China Mobile Group Heilongjiang Co., Ltd. @ on our research,the research is jointly completed by both parties.


## References


Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *CoRR*, abs/2503.01743.

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 422–440. Springer.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. 2023. Musiclm: Generating music from text. *CoRR*, abs/2301.11325.

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Huda AlAmri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2018. Audio visual scene-aware dialog (AVSD) challenge at DSTC7. *CoRR*, abs/1806.00525.

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

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Anonymous. 2024. Aligned better, listen better for audio-visual large language models. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19107–19117. IEEE.

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2021. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13117–13126. IEEE.

Fan Bai, Yuxin Du, Tiejun Huang, Max Qinghu Meng, and Bo Zhao. 2024a. M3D: advancing 3d medical image analysis with multi-modal large language models. *CoRR*, abs/2404.00578.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1708–1718. IEEE.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *CoRR*, abs/2302.12288.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Minesh Mathew, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019.

ICDAR 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1563–1570. IEEE.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017*, pages 1–5. IEEE.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13590–13618. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Xiaomeng Zhao, and et al. 2024. Internlm2 technical report. *CoRR*, abs/2403.17297.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2024. Large language models are strong audio-visual speech recognition learners. *CoRR*, abs/2409.12319.

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

Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. 2023. Pix2video: Video editing using image diffusion. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23149–23160. IEEE.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal LLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13817–13827. IEEE.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE.

Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. 2024a. Model composition for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11246–11262. Association for Computational Linguistics.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. 2020a. Scanrefer: 3d object localization in RGB-D scans using natural language. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, volume 12365 of *Lecture Notes in Computer Science*, pages 202–221. Springer.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. *CoRR*, abs/2305.04160.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 3670–3674. ISCA.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. 2023b. Videocrafter1: Open diffusion models for high-quality video generation. *CoRR*, abs/2310.19512.

Hong Chen, Xin Wang, Yuwei Zhou, Bin Huang, Yipeng Zhang, Wei Feng, Houlun Chen, Zeyang Zhang, Siao Tang, and Wenwu Zhu. 2024b. Multi-modal generative AI: multi-modal llm, diffusion and beyond. *CoRR*, abs/2409.14993.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020b. Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 721–725. IEEE.

Jiaqi Chen, Zeyu Yang, and Li Zhang. 2023c. Semantic segment anything. `https://github.com/fudan-zvg/Semantic-Segment-Anything`.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, Dingdong Wang, Kun Xiang, Haoyuan Li, Haoli Bai, Jianhua Han, Xiaohui Li, Weike Jin, Nian Xie, Yu Zhang, James T. Kwok, Hengshuang Zhao, Xiaodan Liang, Dit-Yan Yeung, Xiao Chen, Zhenguo Li, Wei Zhang, Qun Liu, Jun Yao, Lanqing Hong, Lu Hou, and Hang Xu. 2024c. EMOVA: empowering language models to see, hear and speak with vivid emotions. *CoRR*, abs/2409.18042.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, and Boqing Gong. 2024d. Omnixr: Evaluating omni-modality language models on reasoning across modalities. *CoRR*, abs/2410.12219.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024e. Sharegpt4v: Improving large multi-modal models with better captions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, volume 15075 of *Lecture Notes in Computer Science*, pages 370–387. Springer.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023d. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023e. VALOR: vision-audio-language omni-perception pre-training model and dataset. *CoRR*, abs/2304.08345.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023f. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024f. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13320–13331. IEEE.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, and Jing Shao. 2024g. Octavius: Mitigating task interference in mllms via lora-moe. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023g. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024a. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *CoRR*, abs/2406.11161.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476.

Chee Kheng Chng and Chee Seng Chan. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. In *14th IAPR International Conference on Document Analysis and Recognition,* *ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 935–942. IEEE.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. 2024. MEERKAT: audio-visual large language model for grounding in space and time. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV*, volume 15122 of *Lecture Notes in Computer Science*, pages 52–70. Springer.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Enis Berk Çoban, Michael I. Mandel, and Johanna Devaney. 2024. What do mllms hear? examining reasoning with text and sound components in multimodal large language models. *CoRR*, abs/2406.04615.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2024. A survey on multimodal large language models for autonomous driving. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2024 - Workshops, Waikoloa, HI, USA, January 1-6, 2024*, pages 958–979. IEEE.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13142–13153. IEEE.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *CoRR*, abs/2005.00341.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 736–740. IEEE.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: transforming mandarin ASR research into industrial scale. *CoRR*, abs/1808.10583.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA - multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2416–2428. Association for Computational Linguistics.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE.

Yihe Fan, Yuxin Cao, Ziyu Zhao, Ziyao Liu, and Shaofeng Li. 2024. Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security. *CoRR*, abs/2404.05264.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T. Toshev, and Vaishaal Shankar. 2024a. Data filtering networks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024b. Llama-omni: Seamless speech interaction with large language models. *CoRR*, abs/2409.06666.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. Empathyear: An open-source avatar multimodal empathetic chatbot. *CoRR*, abs/2406.15177.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024b. VITA: towards open-source interactive omni multimodal LLM. *CoRR*, abs/2408.05211.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Preprint*, arXiv:2501.01957.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. 2024c. A touch, vision, and language dataset for multimodal alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a SEED of vision in large language model. *CoRR*, abs/2307.08041.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024a. Making llama SEE and draw with SEED tokenizer. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024b. Worldgpt: Empowering LLM as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 7346–7355. ACM.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22942–22951. IEEE.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2024. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. *CoRR*, abs/2411.19772.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. 2024. OMCAT: omni context aware transformer. *CoRR*, abs/2410.12109.

Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: audio spectrogram transformer. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 571–575. ISCA.

Yuan Gong, Jin Yu, and James R. Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*

*2022, Virtual and Singapore, 23-27 May 2022*, pages 151–155. IEEE.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18973–18990. IEEE.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 5036–5040. ISCA.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024a. Onellm: One framework to align all modalities with language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26574–26585. IEEE.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Imagebind-llm: Multi-modality instruction tuning. *CoRR*, abs/2309.03905.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024b. Parameter-efficient fine-tuning for large models: A comprehensive survey. *CoRR*, abs/2403.14608.

Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, Jifeng Dai, Yong Zhang, Wei Xue, Qifeng Liu, Yike Guo, and Qifeng Chen. 2024. Llms meet multimodal generation and editing: A survey. *CoRR*, abs/2405.19334.

Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. 2024. Multiply: A multisensory object-centric embodied large language model in 3d world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26396–26406. IEEE.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2024. An embodied generalist agent in 3d world. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *CoRR*, abs/2408.15769.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Xiaoshui Huang, Sheng Li, Wentao Qu, Tong He, Yifan Zuo, and Wanli Ouyang. 2022. Frozen CLIP model is an efficient point cloud backbone. *CoRR*, abs/2212.04098.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd. Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, K. Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. 2023. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *CoRR*, abs/2401.04088.

Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, and Ziwei Liu. 2024b. SOLAMI: social vision-language-action modeling for immersive interaction with 3d autonomous characters. *CoRR*, abs/2412.00174.

Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. 2024c. Infrared-llava: Enhancing understanding of infrared images in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8573–8591. Association for Computational Linguistics.

Yang Jin, Zhicheng Sun, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. 2024a. Video-lavit: Unified video-language pretraining with decoupled visual-motional tokenization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024b. Efficient multimodal large language models: A survey. *CoRR*, abs/2405.10739.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukás Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. 2015. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 1156–1160. IEEE Computer Society.

Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. 2013. ICDAR 2013 robust reading competition. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*, pages 1484–1493. IEEE Computer Society.

Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. 2023. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information*

*Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132. Association for Computational Linguistics.

Taemin Kim, WOOYEOL BAEK, and Heeseok Oh. 2024a. Efficient generative multimodal integration (EGMI): Enabling audio generation from text-image pairs through alignment with large language models. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Yeonju Kim, Se Jin Park, and Yong Man Ro. 2024b. Empathetic response in audio-visual conversations using emotion preference optimization and mamba-compressor. *CoRR*, abs/2412.17572.

Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. 2023. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. 2024. Videopoet: A large language model for zero-shot video generation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2880–2894.

Ben Koska and Mojmír Horváth. 2024. Towards multimodal mastery: A 4.5b parameter truly multi-modal small language model. *CoRR*, abs/2411.05903.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Jinxiang Lai, Jie Zhang, Jun Liu, Jian Li, Xiaocheng Lu, and Song Guo. 2024. Spider: Any-to-many multimodal LLM. *CoRR*, abs/2411.09439.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. 2023. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 8291–8298. IEEE.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *CoRR*, abs/2410.12787.

Boxun Li, Yadong Li, Zhiyuan Li, Congyi Liu, Weilin Liu, Guowei Niu, Zheyue Tan, Haiyang Xu, Zhuyu Yao, Tao Yuan, Dong Zhou, Yueqing Zhuang, Shengen Yan, Guohao Dai, and Yu Wang. 2025a. Megrez-omni technical report. *CoRR*, abs/2502.15803.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022a. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19086–19096. IEEE.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022b. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19086–19096. IEEE.

Jian Li and Weiheng Lu. 2024. A survey on benchmarks of multimodal large language models. *CoRR*, abs/2408.08632.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022c. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2046–2065. Association for Computational Linguistics.

Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Zhuoyuan Li, Gang Yu, and Tao Chen. 2024b. M3dbench: Towards omni 3d assistant with interleaved multi-modal instructions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVIII*, volume 15116 of *Lecture Notes in Computer Science*, pages 41–59. Springer.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. 2025b. Baichuan-omni-1.5 technical report. *Preprint*, arXiv:2501.15368.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. 2024c. Baichuan-omni technical report. *CoRR*, abs/2410.08565.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024d. Mini-gemini: Mining the potential of multi-modality vision language models. *CoRR*, abs/2403.18814.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2022d. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *CoRR*, abs/2205.15439.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. 2024e. MERT: acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. 2024f. Omnibench: Towards the future of universal omni-language models. *CoRR*, abs/2409.15272.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2024g. Uni-moe: Scaling unified multimodal llms with mixture of experts. *CoRR*, abs/2405.11273.

Zhaowei Li, Wei Wang, Yiqing Cai, Qi Xu, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. 2024h. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model. *CoRR*, abs/2408.02503.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. 2024i. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6657–6678. Association for Computational Linguistics.

Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, and Qiang Zhu. 2024. Querying as prompt: Parameter-efficient learning for multimodal language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26845–26855. IEEE.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024b. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022*, pages 1140–1144. IEEE.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Trans. Assoc. Comput. Linguistics*, 11:635–651.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. 2023b. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, and Ying Shan. 2024b. Mumu-llama: Multi-modal music understanding and generation via large language models. *CoRR*, abs/2412.06660.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024c. Evalcrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22139–22149. IEEE.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024d. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022a. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3192–3201. IEEE.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024a. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26429–26445. IEEE.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yichen Lu, Jiaqi Song, Xuankai Chang, Hengwei Bian, Soumi Maiti, and Shinji Watanabe. 2024b. Syneslm: A unified approach for audio-visual speech recognition and translation via language model and synthetic data. *CoRR*, abs/2408.00624.

Mingshuang Luo, Ruibing Hou, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. 2024. $M^3$gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *CoRR*, abs/2405.16273.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023a. Valley: Video assistant with large language model enhanced ability. *CoRR*, abs/2306.07207.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Yangyi Chen, Hamid Alinejad-Rokny, and Fei Huang. 2025. Openomni: Large language models pivot zero-shot omnimodal alignment across language with real-time self-aware emotional speech synthesis. *Preprint*, arXiv:2501.04561.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. 2023b. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023c. Videofusion: Decomposed diffusion models for high-quality video generation. *CoRR*, abs/2303.08320.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093.

Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, Philip H. S. Torr, Marc Pollefeys, Matthias Nießner, Ian D. Reid, Angel X. Chang, Iro Laina, and Victor Adrian Prisacariu. 2024. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *CoRR*, abs/2405.10255.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. SQA3D: situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:3339–3354.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *24th European Signal Processing Conference, EUSIPCO 2016, Budapest, Hungary, August 29 - September 2, 2016*, pages 1128–1132. IEEE.

Microsoft. Text to speech an ai speech feature that converts text to lifelike speech.

OpenBMB MiniCPM-o Team. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone. https://github.com/OpenBMB/MiniCPM-o. Accessed: 2025-02-10.

Anand Mishra, Karteek Alahari, and C. V. Jawahar. 2012. Scene text recognition using higher order language priors. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–11. BMVA Press.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952. IEEE.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kamcot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18798–18806. AAAI Press.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. IMU2CLIP: multimodal contrastive learning for IMU motion sensors from egocentric videos and text. *CoRR*, abs/2210.14395.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. 2024. Anymal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pages 1314–1332. Association for Computational Linguistics.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: January 3, 2025.

OpenAI. 2023a. Chatgpt. Technical report, OpenAI.

OpenAI. 2023b. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenGV. 2024. Interomni. Accessed: 2024-07-27.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.

Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. 2024. Reka core, flash, and edge: A series of powerful multimodal language models. *CoRR*, abs/2404.12387.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2024. X-instructblip: A framework for aligning image, 3d, audio, video to llms and its emergent cross-modal reasoning. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLV*, volume 15103 of *Lecture Notes in Computer Science*, pages 177–197. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Haozhou Pang, Tianwei Ding, Lanshan He, Ming Tao, Lu Zhang, and Qi Gan. 2024. LLM gesticulator: Leveraging large language models for scalable and controllable co-speech gesture synthesis. *CoRR*, abs/2410.10851.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *CoRR*, abs/2208.06366.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 724–732. IEEE Computer Society.

Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. 2013. Recognizing text with perspective distortion in natural scenes. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 569–576. IEEE Computer Society.

Karol J. Piczak. 2015. ESC: dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 1015–1018. ACM.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 2757–2761. ISCA.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6967–6977. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12159–12168. IEEE.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chu Yuan Zhang, and Junzuo Zhou. 2024. Fewer-token neural speech codec with time-invariant codes. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12737–12741. IEEE.

Anhar Risnumawan, Palaiahnakote Shivakumara, Chee Seng Chan, and Chew Lim Tan. 2014. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.*, 41(18):8027–8048.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510. IEEE.

RVC-Boss. Gpt-sovits. `https://github.com/RVC-Boss/GPT-SoVITS`. Accessed: January 3, 2025.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. 2022b(a). Laion-aesthetics.

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

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021a. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021b. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Share. 2024. Sharegemini: Scaling up video caption data for multimodal large language models.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yiming Shi, Xun Zhu, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. 2024. Med-2e3: A 2d-enhanced 3d medical multimodal large language model. *CoRR*, abs/2411.12783.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving image captioning with better use of captions. *CoRR*, abs/2006.11807.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual LLM for video understanding. *CoRR*, abs/2312.06720.

Mustafa Shukor, Corentin Dancette, and Matthieu Cord. 2023a. ep-alm: Efficient perceptual augmentation of language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21999–22012. IEEE.

Mustafa Shukor, Corentin Dancette, Alexandre Ramé, and Matthieu Cord. 2023b. Unival: Unified model for image, video, audio and language tasks. *Trans. Mach. Learn. Res.*, 2023.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 510–526. Springer.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. 2024. SNAC: multi-scale neural audio codec. *CoRR*, abs/2410.14411.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 567–576. IEEE Computer Society.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024a. video-salmonn: Speech-enhanced audio-visual large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023a. Fine-grained audio-visual joint representations for multimodal large language models. *CoRR*, abs/2310.05863.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. 2023b. Journeydb: A benchmark for generative image understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023c. MAE-DFER: efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 6110–6121. ACM.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023d. MAE-DFER: efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 6110–6121. ACM.

Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. 2024b. Auto-acd: A large-scale dataset for audio-language representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 5025–5034. ACM.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023e. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024c. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024d. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Andrew Szot, Bogdan Mazoure, Harsh Agrawal, R. Devon Hjelm, Zsolt Kira, and Alexander Toshev. 2024a. Grounding multimodal large language models in actions. *CoRR*, abs/2406.07904.

Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. 2024b. From multimodal llms to generalist embodied agents: Methods and lessons. *Preprint*, arXiv:2412.08442.

Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. 2024a. Avicuna: Audio-visual LLM with interleaver and context-boundary alignment for temporal referential dialogue. *CoRR*, abs/2403.16276.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024b. Codi-2: In-context, interleaved, and interactive any-to-any generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27415–27424. IEEE.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 436–454. Springer.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022,* *New Orleans, LA, USA, November 28 - December 9, 2022*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26690–26699. IEEE.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *CSTR*, 6:15.

Andreas Veit, Tomas Matera, Lukás Neumann, Jiri Matas, and Serge J. Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *CoRR*, abs/1601.07140.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. 2019. RIO: 3d object instance re-localization in changing indoor environments. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7657–7666. IEEE.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024a. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *CoRR*, abs/2408.01319.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023a. Modelscope text-to-video technical report. *CoRR*, abs/2308.06571.

Kai Wang, Boris Babenko, and Serge J. Belongie. 2011. End-to-end scene text recognition. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 1457–1464. IEEE Computer Society.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Wenhai Wang, Jiangwei Xie, Chuanyang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, Hao Tian, Lewei Lu, Xizhou Zhu, Xiaogang Wang, Yu Qiao, and Jifeng Dai. 2023b. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *CoRR*, abs/2312.09245.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Xinyu Wang, Bohan Zhuang, and Qi Wu. 2024c. Modaverse: Efficiently transforming modalities with llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26596–26606. IEEE.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024d. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024e. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Zekun Wang, Jingchang Chen, Wangchunshu Zhou, Haichao Zhu, Jiafeng Liang, Liping Shan, Ming Liu, Dongliang Xu, Qing Yang, and Bing Qin. 2024f. Smarttrim: Adaptive tokens and attention pruning for efficient vision-language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 14937–14953. ELRA and ICCL.

Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang

Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. 2024g. Mio: A foundation model on multimodal tokens. *Preprint*, arXiv:2409.17692.

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. 2024. Occllama: An occupancy-language-action generative world model for autonomous driving. *CoRR*, abs/2409.03272.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. 2017. AI challenger : A large-scale dataset for going deeper in image understanding. *CoRR*, abs/1711.06475.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2024a. Grit: A generative region-to-text transformer for object understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pages 207–224. Springer.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, pages 2247–2256. IEEE.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024b. Next-gpt: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920. IEEE Computer Society.

Xunzhi Xiang, Haiwei Xue, Zonghong Dai, Di Wang, Minglei Li, Ye Yue, Fei Ma, Weijiang Yu, Heng Chang, and Fei Richard Yu. 2025. Remask-animate: Refined character image animation using mask-guided adapters. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 8628–8636. AAAI Press.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2024. A comprehensive survey of large language models and multimodal large language models in medicine. *CoRR*, abs/2405.08603.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017a. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653. ACM.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017b. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653. ACM.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. 2023. Unifying flow, stereo and depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13941–13958.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Jingwei Xu, Chenyu Wang, Zibo Zhao, Wen Liu, Yi Ma, and Shenghua Gao. 2024a. CAD-MLLM: unifying multimodality-conditioned CAD generation with MLLM. *CoRR*, abs/2411.04954.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024b. Pointllm: Empowering large language models to understand point clouds. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXV*, volume 15083 of *Lecture Notes in Computer Science*, pages 131–147. Springer.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024c. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. 2024a. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:4700–4712.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2024b. ULIP-2: towards scalable multimodal pre-training for 3d understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27081–27091. IEEE.

Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. 2023a. GD-MAE: generative decoder for MAE pre-training on lidar point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9403–9414. IEEE.

Zhen Yang, Yingxue Zhang, Fandong Meng, and Jie Zhou. 2023b. TEAL: tokenize and embed ALL for multi-modal large language models. *CoRR*, abs/2311.04589.

Ziyi Yang, Mahmoud Khademi, Yichong Xu, Reid Pryzant, Yuwei Fang, Chenguang Zhu, Dongdong Chen, Yao Qian, Xuemei Gao, Yi-Ling Chen, Robert Gmyr, Naoyuki Kanda, Noel Codella, Bin Xiao, Yu Shi, Lu Yuan, Takuya Yoshioka, Michael Zeng, and Xuedong Huang. 2024. i-code V2: an autoregressive generation framework over vision, language, and speech data. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1615–1627. Association for Computational Linguistics.

Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, and Hongxu Yin. 2024a. X-VILA: cross-modality alignment for large language model. *CoRR*, abs/2405.19335.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2024b. CAT: enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part X*, volume 15068 of *Lecture Notes in Computer Science*, pages 146–164. Springer.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023a. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, Jing Shao, and Wanli Ouyang. 2023b. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652.

Jiazuo Yu, Haomiao Xiong, Lu Zhang, Haiwen Diao, Yunzhi Zhuge, Lanqing Hong, Dong Wang, Huchuan Lu, You He, and Long Chen. 2024a. Llms can evolve continually on modality for x-modal reasoning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2024b. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. 2024c. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. *Preprint*, arXiv:2402.05889.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024d. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Weijiang Yu, Haofan Wang, Guohao Li, Nong Xiao, and Bernard Ghanem. 2023. Knowledge-aware global reasoning for situation recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8621–8633.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. MERLOT RESERVE: neural script knowledge through vision and language and sound. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16354–16366. IEEE.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9637–9662. Association for Computational Linguistics.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022a. WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6182–6186. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12401–12430. Association for Computational Linguistics.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics.

Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. 2024b. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14498–14511. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023c. Speechtokenizer: Unified speech tokenizer for speech large language models. *CoRR*, abs/2308.16692.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023d. Meta-transformer: A unified framework for multimodal learning. *CoRR*, abs/2307.10802.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *CoRR*, abs/2503.05379.

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

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023b. Chatbridge: Bridging modalities with large language model as a language catalyst. *CoRR*, abs/2305.16103.

Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, Shaozuo Yu, Sitong Wu, Eric Lo, Shu Liu, and Jiaya Jia. 2024. Lyra: An efficient and speech-centric framework for omni-cognition. *Preprint*, arXiv:2412.09501.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022. Audio-visual segmentation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697 of *Lecture Notes in Computer Science*, pages 386–403. Springer.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024b. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *CoRR*, abs/2409.18125.

Hongyan Zhu, Shuai Qin, Min Su, Chengzhi Lin, Anjie Li, and Junfeng Gao. 2024c. Harnessing large vision and language models in agriculture: A review. *CoRR*, abs/2407.19679.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023a. Multimodal C4: an open, billion-scale corpus of images interleaved with text. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023b. Multimodal C4: an open, billion-scale corpus of images interleaved with text. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

# A Related Survey

With the advent of MLLMs, there are several surveys detailing the current progress of MLLMs. Yin et al. (2023a); Wu et al. (2023); Caffagni et al. (2024) focus on the early Vision-MLLMs, while Çoban et al. (2024) and Ma et al. (2024) respectively summarize the Audio-MLLMs and 3D-MLLMs. Zhang et al. (2024a); Wang et al. (2024a) conduct an investigation into various Specific-MLLMs of different modalities. He et al. (2024); Chen et al. (2024b) discuss the expansion of MLLM's generative capabilities. Some works discuss MLLMs in specific domains, such as medicine (Xiao et al., 2024), agriculture (Zhu et al., 2024c), and autonomous driving (Cui et al., 2024). Some works highlight some specific tasks such as safety (Fan et al., 2024), hallucination (Bai et al., 2024b), and acceleration (Zhu et al., 2024c). And Li and Lu (2024); Huang and Zhang (2024) focus on the evaluation of MLLM performance.

Distinct from the above-mentioned surveys, this paper focuses on MLLMs that align multiple non-linguistic modalities[3] with LLMs (Omni-MLLMs), enabling cross-modal understanding or cross-modal generation. As the first systematic survey on Omni-MLLMs, we hope our work will serve as an overview of this emerging direction, fostering future research in the field.

# B Details about Omni-MLLMs architures

Table 1 presents the details of the structure of mainstream Omni-MLLMs. We will list some of the pre-trained models used.

## B.1 Modality Encoder

**Visual Specific-Encoder** Vit (Dosovitskiy et al., 2021), SigCLIP Vit (Zhai et al., 2023), CLIP Vit (Radford et al., 2021), EVA CLIP Vit (Sun et al., 2023e), InternVit (Chen et al., 2023g), DINOv2 Vit (Oquab et al., 2024), DFNCLIP Vit (Fang et al., 2024a), and OpenCLIP ConvNext (Liu et al., 2022b) encode images to obtain continuous features. TimeSformer (Bertasius et al., 2021), VideoMAE (Tong et al., 2022), MAE-DFer (Sun et al., 2023c), Omni-VL (Sun et al., 2023d), Video-Swin (Liu et al., 2022a), and Vivit (Arnab et al., 2021) encode videos to obtain continuous features.

[3]Since MLLMs capable of comprehending both video and imagery generally process video as multiple frames and employ a single vision encoder, we categorize them as Specific-MLLMs, i.e. Vision-MLLMs.

**Audio Specific-Encoder** AST (Gong et al., 2021), Beats (Chen et al., 2023d), Whisper (Radford et al., 2023), HuBert (Hsu et al., 2021), CLAP (Elizalde et al., 2023), Conformer (Gulati et al., 2020), MERT (Li et al., 2024e), and PANN (Kong et al., 2020) encode the audio modality to obtain continuous features.

**3D Specific-Encoder** ULIP2 (Xue et al., 2024b), GD-MAE (Yang et al., 2023a), PointEncoder (Xu et al., 2024b), FrozenCLIP (Huang et al., 2022), and M3D-CLIP (Bai et al., 2024a) encode the 3D modality to obtain continuous features.

**Pre-align Uni-Encoder** LanguageBind (Zhu et al., 2024a), ImageBind (Girdhar et al., 2023), Meta-Transformers (Zhang et al., 2023d), TVL (Fu et al., 2024c), and SSVTP (Kerr et al., 2023) encode multiple non-linguistic modalities into a unified feature space and obtain continuous features. TVL, LanguageBind, ImageBind, and SSVTP construct modality-specific encoders for different modalities and achieve multi-modalities alignment through indirect alignment. Meta-Transformers design distinct modality-specific patch embeddings and use a shared encoder to encode multiple modalities.

**Other Specific-Encoder** IMU2CLIP (Moon et al., 2022) encodes the IMU modality to obtain continuous features. Individual modality-specific encoders from LanguageBind or ImageBind are often used independently as specific encoders.

## B.2 Modality Tokenizer

**Visual Tokenizer** VQ-GAN (Esser et al., 2021), DALL-E (Ramesh et al., 2021), BEiT-V2 (Peng et al., 2022), MAGVIT-v2 (Yu et al., 2024b), and SEED (Ge et al., 2023) encode the visual modality into discrete visual tokens, which can be decoded back into the original image using the de-tokenizer.

**Audio Tokenizer** Jukebox (Dhariwal et al., 2020), SoundStream (Zeghidour et al., 2022), SpeechTokenizer (Zhang et al., 2023c), Encodec (Défossez et al., 2023), and S2U (Chen et al., 2024c) encode the audio modality into discrete audio tokens, which can be decoded back into the audio using the corresponding de-tokenizer.

**Other Tokenizer** Scene Tokenizer (Wei et al., 2024) encodes the 3D modality into discrete 3D tokens. LEO (Huang et al., 2024), Ground-Action (Szot et al., 2024a), OccLLaMA (Wei et al., 2024), and GMA (Szot et al., 2024b) perform discrete encoding of the action modality to obtain corresponding action tokens, which can be decoded back into the original action using the corresponding de-tokenizer. M3GPT (Luo et al., 2024), Gesticulator (Pang et al., 2024), and SOLAMI (Jiang et al., 2024b) perform discrete encoding of the motion modality to obtain corresponding motion tokens, which can be decoded back into the original motion using the corresponding de-tokenizer.

### B.3 Modality Generation Model

For image generation, Stable Diffusion (Rombach et al., 2022) and Instruct-Pix2Pix (Brooks et al., 2023) are used. Video generation models include Zeroscope (Cerspense, 2023), VideoFusion (Luo et al., 2023c), VideoCrafter (Chen et al., 2023b), and ModelScope (Wang et al., 2023a). For audio generation, models such as AudioLDM (Liu et al., 2023b), SNAC (Siuzdak et al., 2024), LLaMA-Omni's audio decoder (Fang et al., 2024b), MusicGen (Copet et al., 2023), and TiCodec (Ren et al., 2024) are utilized. Meanwhile, StyleTTS (Li et al., 2022d) and GPT-SoVITS (RVC-Boss) are employed for speech generation.

### B.4 LLM Backbone

Commonly used LLMs include the T5 series (Raffel et al., 2020), LLaMA series (Touvron et al., 2023), Qwen series (Bai et al., 2023), Internlm series (Cai et al., 2024), Chatglm series (Zeng et al., 2024), OPT series (Zhang et al., 2022b), Mixtral series (Jiang et al., 2024a), Mistral series (Jiang et al., 2023), Phi series (Gunasekar et al., 2023), and Yi series (Young et al., 2024).

## C Details of Training and evaluation

### C.1 Details of Training Data

The statistical results of some commonly used alignment datasets and the instruction data of mainstream Omni-MLLMs are shown in Table 2 and Table 3. There is still a lack of alignment data for data-scarcity modalities and cross-modal instruction data.

### C.2 Details of Benchmark

The statistical data of some commonly used benchmarks are shown in Table 4. Existing benchmarks still require improvements in terms of the number of modalities and the forms of modality interaction.

### C.3 Performance of Omni-MLLMs

We statistic the performance of various mainstream Omni-MLLMs in uni-modal understanding, cross-modal understanding, and cross-modal, as shown in Table 5. We also show the performance of several Specific-MLLMs (Lin et al., 2024b; Li et al., 2024a; Chu et al., 2023; Xu et al., 2024b; Sun et al., 2024c; Jin et al., 2024a) on selected tasks for comparison. It is worth noting that due to differences in the size and performance of the pre-trained models, Omni-MLLMs with the same backbone LLM may still not be fairly comparable. Therefore, this table only provides a rough trend of performance.

The results are mainly from corresponding papers (some results are used as baselines in other papers). although Omni-MLLMs generally possess cross-modal understanding capabilities, some of their original papers do not provide systematic evaluations or do not report results on our chosen benchmarks.

It can be seen from the table that most Omni-MLLMs still exhibit a significant performance gap in uni-modal understanding tasks compared to Specific-MLLMs. Meanwhile, in uni-modal generation tasks, models like AnyGPT and CoDi-2 have achieved performance close to or even surpassing Specific-MLLMs. Additionally, Omni-MLLMs are capable of performing cross-modal tasks that Specific-MLLMs cannot handle.

| Model | Capabilities | Multi-Modalities Encoding | | | Multi-Modalities Alignment | | | Multi-Modalities Interaction | | Multi-Modalities Generation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Modalities | Method | Encoding Model | Method | Projector | Vocabulary | Method | LLM | Modalities | Method | Generation model |
| eP-ALM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | Vit/TimeSformer/AST | multi-branch | linear | – | injection | OPT | – | – | – |
| VALOR | Corss-modal Understanding | Visual/Audio | Continuous Encoding | Vit/VideoSwin/AST | multi-branch | MLP | – | injection | Bert | – | – | – |
| X-LLM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | Vit/Conformer | multi-branch | Q-former+Linear | – | concatenate | ChatGLM | – | – | – |
| ChatBridge | Corss-modal Understanding | Visual/Audio | Continuous Encoding | EVAL CLIP Vit/Beats | multi-branch | Preceiver | – | concatenate | Vicuna | – | – | – |
| PandaGPT | Corss-modal Understanding | Visual/Audio/3D/ IMU/thermal | Continuous Encoding | ImageBind | uni-branch | linear | – | concatenate | Vicuna | – | – | – |
| VideoLLama | Corss-modal Understanding | Visual/Audio | Continuous Encoding | EVA CLIP Vit/ ImageBind-Audio | multi-branch | Q-former+Linear | – | concatenate | Vicuna | – | – | – |
| LAMM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/ FrozenCLIP | multi-branch | MLP | – | concatenate | Vicuna | – | – | – |
| Macaw-LLM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | Vit/Whisper | multi-branch | Cross-Attention | – | concatenate | LLaMA | – | – | – |
| BuboGPT | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/ImageBind-Audio | multi-branch | Q-former+Linear | – | concatenate | LLaMA | – | – | – |
| Teal | Corss-modal Understanding && Generation | Visual/Audio | Discrete Encoding | VQ-GAN/ Whisper+K-means | embedding | – | Added Vocabulary | concatenate | LLaMA | Image | Modality-Token-based | VQGAN detokenizer |
| ImageBind-LLM | Corss-modal Understanding | Visual/Audio/3D | Continuous Encoding | ImageBind | uni-branch | MLP | – | injection | LLaMA | – | – | – |
| Next-GPT | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | Linear | – | concatenate | Vicuna | Image/Video/Audio | Representation-based | StableDiffusion1.5/ Zeroscope/AudioLDM |
| Any-MAL | Corss-modal Understanding | Visual/Audio/IMU | Continuous Encoding | CLIP Vit/CLAP/ IMU2CLIP | multi-branch | Preceiver | – | injection | LLaMA-2 | – | – | – |
| FAVOR | Corss-modal Understanding | Visual/Audio | Continuous Encoding | EVA CLIP Vit/Whisper | multi-branch | Q-former+Linear | – | concatenate | Vicuna | – | – | – |
| Octavius | Corss-modal Understanding | Visual/3D | Continuous Encoding | CLIP Vit/Object-As-Scene | multi-branch | MLP | – | concatenate | Vicuna | – | – | – |
| LEO | Corss-modal Understanding && Generation | Visual/3D/Action | Hybrid Encoding | OpenCLIP Convnext/PointNet++/ LEO's Action tokenizer | multi-branch | MLP/ Spatial Transformer | Overwrite Vocabulary | concatenate | Vicuna | Action | Modality-Token-based | LEO's Action detokenizer |
| CoDi-2 | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | ImageBind | uni-branch | MLP | – | concatenate | LLaMA-2 | Image/Video/Audio | Representation-based | StableDiffusion2.1/ Zeroscope/AudioLDM2 |
| X-InstructBLIP | Corss-modal Understanding | Visual/Audio/3D | Continuous Encoding | EVA CLIP Vit/Beats/ ULIP2 | multi-branch | Q-former+Linear | – | concatenate | Vicuna | – | – | – |
| One-LLM | Corss-modal Understanding | Visual/Audio/3D/IMU/fMRI/ Normal Map | Continuous Encoding | Meta-transformer | uni-branch | UPM(self-attention) | – | concatenate | LLaMA-2 | – | – | – |
| AV-LLM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/CLAP | multi-branch | Linear | – | concatenate | Vicuna | – | – | – |
| DriveMLM | Corss-modal Understanding | Visual/3D | Continuous Encoding | EVA CLIP Vit/GD-MAE | multi-branch | Q-former+Linear | – | concatenate | LLaMA | – | – | – |
| Omni-3D | Corss-modal Understanding | Visual/3D | Continuous Encoding | CLIP Vit/PointNet++ | multi-branch | MLP | – | concatenate | LLaMA-2 | – | – | – |
| ModaVerse | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | Linear | – | concatenate | Vicuna | Image/Video/Audio | Text-based | StabelDiffusion /AudioLDM/VideoFusion |
| MultiPLY | Corss-modal Understanding | Visual/Audio/3D /thermal/touch | Continuous Encoding | CLIP Vit/CLAP /ConceptGraph | multi-branch | MLP/Linear | – | concatenate | Vicuna | – | – | – |
| CREMA | Corss-modal Understanding | Visual/Audio/3D /thermal/touch/optical | Continuous Encoding | EVA CLIP Vit+Linear/ Beats+Linear/ConceptFusion+Linear | uni-branch | Q-former+Linear | – | concatenate | Flan-T5 | – | – | – |
| GroundingGPT | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/ImageBind-Audio | multi-branch | Q-former+Linear/MLP | – | concatenate | Vicuna | – | – | – |
| DAMC | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/Beats/PointBert(PointLLM) | multi-branch | Q-former+Linear/MLP | – | concatenate | Vicuna | – | – | – |
| AnyGPT | Corss-modal Understanding && Generation | Visual/Audio | Diserect Encoding | SEED tokenizer/ Speech tokenizer/Encodec | embedding | – | Extend Vocabulary | concatenate | LLaMA-2 | Image/Speech/Music | – | SEED de-tokenizer/Speech de-tokenizer/Encodec de-tokenizer |
| TVL-LLaMA | Corss-modal Understanding | Visual/Touch | Continuous Encoding | TVL Encoders | uni-branch | MLP | – | injection | LLaMA | – | – | – |
| SSVTP-LLaMA | Corss-modal Understanding | Visual/Touch | Continuous Encoding | SSVTP Encoders | uni-branch | MLP | – | injection | LLaMA | – | – | – |
| CAT | Corss-modal Understanding | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | Linear | – | concatenate | LLaMA-2 | – | – | – |
| AVicuna | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/CLAP | multi-branch | MLP | – | concatenate | Vicuna | – | – | – |
| WorldGPT | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | LanguageBind | uni-branch | Linear | – | concatenate | Vicuna | Image/Video/Audio | Representation-based | Stable Diffusion /AudioLDM/Zeroscope |
| QaP | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/CLAP | multi-branch | dot attention+Linear | – | injection | DeBERTa-V2-XLarge | – | – | – |
| Uni-Moe | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/Beats | multi-branch | Q-former+Linear/MLP | – | concatenate | LLaMA | – | – | – |
| M3GPT | Corss-modal Understanding | Audio/Motion | Diserect Encoding | Jukebox tokenizer /M3GPT's Motion tokenier | embedding | Extend Vocabulary | – | concatenate | T5 | Music/Motion | Modality-Token-based | Jukebox de-tokenizer /M3GPT's Motion de-tokenizer |
| X-VILA | Corss-modal Understanding | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | MLP | – | concatenate | Vicuna | Image/Video/Audio | Representation-based | Stable Diffusion /AudioLDM/VideoCrafter |
| REAMO | Corss-modal Understanding | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | Linear | – | concatenate | Vicuna | – | – | – |
| VideoLLaMA2 | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/Beats | multi-branch | MLP | – | concatenate | Mistral | – | – | – |
| Ground-Action | Corss-modal Understanding | Visual/Action | Hybrid Encoding | CLIP Vit /Ground-Action action tokenizer | multi-branch | Preceiver | Overwrite Vocabulary | concatenate | Vicuna | Action | Modality-Token-based | Ground-Action action de-tokenizer |
| Emotion-LLaMA | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | CLIP Vit /HuBert-Chinese | multi-branch | MLP | – | concatenate | LLaMA-2 | – | – | – |
| EmpathyEar | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | ImageBind | uni-branch | Linear | – | concatenate | ChatGLM | Video/Audio | Text-based | StyleTTS2/EAT |
| video-SALMONN | Corss-modal Understanding | Visual/Audio | Continuous Encoding | Vit/Beats | multi-branch | Q-former+Linear | – | concatenate | Vicuna | – | – | – |
| Meerkat | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/CLAP | multi-branch | MLP | – | concatenate | LLaMA-2 | – | – | – |
| InternOmni | Corss-modal Understanding | Visual/Audio | Continuous Encoding | Intern Vit/Whisper | multi-branch | MLP | – | concatenate | InternLM-2.5 | – | – | – |
| SynesLM | Corss-modal Understanding | Visual/Audio | Hybird Encoding | SigCLIP Vit /XLSR+K-means | multi-branch | MLP | Extend Vocabulary | concatenate | OPT | – | – | – |
| UnifiedMLLM | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | CLIP Vit /ImageBind-Audio | multi-branch | Q-former+Linear | – | concatenate | OPT | Image/Video/Audio | Text-based | Instruct-pix2pix/ Auffusion/ModelScope |
| VITA | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | CLIP Vit /VITA's Audio Encoder | multi-branch | MLP | – | concatenate | Mixtral | Speech | Text-based | GPT-SoVITS |
| OccLLaMA | Corss-modal Understanding && Generation | 3D/Action | Discrete Encoding | OccLLaMA's 3D tokenizer /OccLLaMA's Action tokenizer | embedding | – | Extend Vocabulary | concatenate | LLaMA-3.1 | Action/3D | Modality-Token-based | OccLLaMA's 3D de-tokenizer /OccLLaMA's Action de-tokenizer |
| Llama-AVSR | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | AV-HuBert /Whisper | multi-branch | MLP | – | concatenate | LLaMA-3.1 | – | – | – |
| MIO | Corss-modal Understanding && Generation | Visual/Audio | Discrete Encoding | SEED tokenizer /Speech tokenizer | embedding | – | Extend Vocabulary | concatenate | Yi | Image/Speech | Modality-Token-based | SEED de-tokenizer /Speech de-tokenizer |
| EMOVA | Corss-modal Understanding && Generation | Visual/Audio | Hybird Encoding | Intern Vit /EMOVA's S2U tokenizer | multi-branch | C-Abstractor | Extend Vocabulary | concatenate | LLaMA-3.1 | Speech | Modality-Token-based | EMOVA's S2U de-tokenizer |
| LLM Gesticulator | Corss-modal Understanding && Generation | Audio/Motion | Diserect Encoding | MotionRVQ tokenizer /Encodec tokenizer | embedding | – | Extend Vocabulary | concatenate | Qwen-1.5 | Audio/Motion | Modality-Token-based | MotionRVQ de-tokenizer /Encodec de-tokenizer |
| Baichuan-Omni | Corss-modal Understanding | Audio/Motion | Continuous Encoding | SigCLIP /Whisper | multi-branch | CNN+MLP/Conv-GMLP | – | concatenate | – | – | – | – |
| EGMI | Corss-modal Understanding && Generation | Audio/Motion | Continuous Encoding | ImageBind | uni-branch | Linear | – | concatenate | Vicuna | Image/Audio | Representation-based | StableDiffusion /AudioLDM |
| Dolphin | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | CLIP Vit/ImageBind-Audio | multi-branch | MLP | – | concatenate | Vicuna | – | – | – |
| Mini-Omni2 | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | CLIP Vit/Whisper | multi-branch | MLP | – | concatenate | Qwen2 | Audio | Modality-Token-based | SNAC de-tokenizer |
| OMCAT | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/ImageBind-Audio | multi-branch | Q-former+transformer | – | concatenate | Vicuna | – | – | – |
| PathWeave | Corss-modal Understanding | Visual/Whisper | Continuous Encoding | EVA CLIP Vit/Beats /ULIP2 | uni-branch | Q-former+Linear | – | concatenate | Vicuna | – | – | – |
| CAD-MLLM | Corss-modal Understanding | Visual/3D | Continuous Encoding | DINO v2 /Michelangelo | multi-branch | Preceiver+Linear/Linear | – | concatenate | Vicuna | – | – | – |
| EAGLE | Corss-modal Understanding | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | MLP | – | concatenate | LLaMA-2 | Image/Video/Audio | Text-based | StableDiffusion /AudioLDM/Zeroscope |
| Spider | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | ImageBind | multi-branch | MLP | – | concatenate | LLaMA-2 | Image/Video/Audio | Text-based | StableDiffusion /AudioLDM/Zeroscope |
| Med-2E3 | Corss-modal Understanding | Visual/3D | Continuous Encoding | SigCLIP Vit/M3D-CLIP | multi-branch | Q-former+Linear/MLP | – | concatenate | Phi | – | – | – |
| LongVALE-LLM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/Beats/Whisper | multi-branch | MLP | – | concatenate | Vicuna | – | – | – |
| SOLAMI | Corss-modal Understanding | Audio/Motion | Diserect Encoding | SpeechTokenizer /SOLAMI's MotionTokenizer | embedding | – | Extend Vocabulary | concatenate | Vicuna | Speech/Motion | Modality-Token-based | Speech de-tokenizer /SOLAMI's Motion de-tokenizer |
| MuMu-LLaMA | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | Vit/ViViT/MERT | multi-branch | Conv+MLP/Conv+Rnn+MLP | – | injection | LLaMA-2 | Music | Representation-based | MusicGen |
| GMA | Corss-modal Understanding && Generation | Visual/Action | Hybird Encoding | SigCLIP Vit /GMA's Action tokenizer | multi-branch | MLP | Overwrite Vocabulary | concatenate | Qwen-2 | Action | Modality-Token-based | GMA's Action de-tokenizer |
| Lyra | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | DFNCLIP Vit/Whipser | multi-branch | MLP | – | concatenate | Qwen-2 | Audio | Representation-based | LLaMA-Omni's audio decoder |
| VITA-1.5 | Corss-modal Understanding && Generation | Visual/Audio | Continuous Encoding | InternVit/VITA's Audio Encoder | multi-branch | MLP/CNN+MLP | – | concatenate | Mixtral | Speech | Representation-based | TiCodec decoder |
| EmpatheticLLM | Corss-modal Understanding | Visual/Audio | Continuous Encoding | CLIP Vit/Whisper | multi-branch | Q-former+Linear | – | concatenate | Qwen2.5 | – | – | – |

Table 1: **The architectures of mainstream OmniMLLMs.** The architectures of 70 Omni-MLLMs are displayed by encoding, alignment, interaction, and generation.

| Name | Type | Modality | #Sample |
|---|---|---|---|
| MSCOCO (Lin et al., 2014) | X-Text | Image,Text | 620K |
| Visual Genome (Krishna et al., 2017b) | X-Text | Image,Text | 4.5M |
| Flickr30k (Plummer et al., 2015) | X-Text | Image,Text | 158K |
| SBU (Ordonez et al., 2011) | X-Text | Image,Text | 1M |
| DCI (Urbanek et al., 2024) | X-Text | Image,Text | 7.8K |
| BLIP-Capfilt (Li et al., 2022c) | X-Text | Image,Text | 129M |
| AI Challenger captions (Wu et al., 2017) | X-Text | Image,Text | 1.5M |
| Wukong Captions (Gu et al., 2022) | X-Text | Image,Text | 101M |
| CC12M (Changpinyo et al., 2021) | X-Text | Image,Text | 12.4M |
| CC3M (Sharma et al., 2018) | X-Text | Image,Text | 3.3M |
| LAION-5B (Schuhmann et al., 2022) | X-Text | Image,Text | 5.9B |
| Redcaps (Desai et al., 2021) | X-Text | Image,Text | 12M |
| LAION-COCO (Schuhmann et al., 2022b(b)) | X-Text | Image,Text | 600M |
| LAION-CAT (Radenovic et al., 2023) | X-Text | Image,Text | 440M |
| LAION-AESTHETICS (Schuhmann et al., 2022b(a)) | X-Text | Image,Text | 120M |
| ShareGPT4V (Chen et al., 2024e) | X-Text | Image,Text | 1.2M |
| LAION-115M (Schuhmann et al., 2021a) | X-Text | Image,Text | 115M |
| Journeydb (Sun et al., 2023b) | X-Text | Image,Text | 4.4M |
| Multimodal c4 (Zhu et al., 2023b) | X-Text-X | Image,Text | 43.3M |
| OBELICS (Laurençon et al., 2023) | X-Text-X | Image,Text | 141M |
| Panda-70M (Chen et al., 2024f) | X-Text | Video,Text | 70M |
| Webvid2M (Bain et al., 2021) | X-Text | Video,Text | 2M |
| Valley-Pretrain-703k (Luo et al., 2023a) | X-Text | Video,Text | 703K |
| Webvid10M (Bain et al., 2021) | X-Text | Video,Text | 10M |
| YT-Temporal (Zellers et al., 2022) | X-Text | Video,Text | 180M |
| ActivityNet Captions (Krishna et al., 2017a) | X-Text | Video,Text | 100K |
| InterVid (Wang et al., 2024d) | X-Text | Video,Text | 10M |
| MSRVTT (Xu et al., 2016) | X-Text | Video,Text | 200K |
| ShareGemini (Share, 2024) | X-Text | Video,Text | 530K |
| AudioSet (Gemmeke et al., 2017) | X-Text | Audio,Text | 2.1M |
| Clotho (Drossos et al., 2020) | X-Text | Audio,Text | 5k |
| Auto-ACD (Sun et al., 2024b) | X-Text | Audio,Text | 1.5M |
| AudioCap (Kim et al., 2019) | X-Text | Audio,Text | 46k |
| WavCaps (Mei et al., 2024) | X-Text | Audio,Text | 403K |
| AISHELL-1 (Bu et al., 2017) | X-Text | Audio,Text | 128K |
| AISHELL-2 (Du et al., 2018) | X-Text | Audio,Text | 1M |
| Gigaspeech (Chen et al., 2021) | X-Text | Speech,Text | – |
| Common Voice (Ardila et al., 2020) | X-Text | Speech,Text | – |
| MLS (Pratap et al., 2020) | X-Text | Speech,Text | – |
| Music caption (Zhan et al., 2024) | X-Text | Music,Text | 100M |
| Cap3D (Luo et al., 2023b) | X-Text | 3D,Text | 1M |
| Objaverse (Deitke et al., 2023) | X-Text | 3D,Text | 800K |
| ScanRefer (Chen et al., 2020a) | X-Text | 3D,Text | 51.5K |
| Normal Caption (Han et al., 2024a) | X-Text | Normal,Text | 0.5M |
| Depth Caption (Han et al., 2024a) | X-Text | Depth,Text | 0.5M |
| NSD (Allen et al., 2022) | X-Text | fMRI,Text | 9K |
| Ego4d (Grauman et al., 2022) | X-Text | Video,IMU,Text | 528k |
| PU-VALOR (Tang et al., 2024a) | X-Y-Text | Video,Audio,Text | 114K |
| VALOR (Chen et al., 2023e) | X-Y-Text | Video,Audio,Text | 16k |
| VAST (Chen et al., 2023f) | X-Y-Text | Video,Audio,Text | 414k |
| VIDAL (Zhu et al., 2024a) | X-Y-Text | Video, Thermal, Depth, Audio | 10M |
| TVL (Fu et al., 2024c) | X-Y-Text | Image,Touch,Text | 44K |
| M3D-Cap (Bai et al., 2024a) | X-Y-Text | Image,3D,Text | 115K |

Table 2: **The statistics for alignment datasets in Omni-MLLMs**, including single non-linguistic modality text pairing data (X-Text), multiple non-linguistic modalities text pairing data (X-Text-Y), and single non-linguistic modality text interleaved data (X-Text-X).

| Name | Source | Task | Modality | Construction Method | #Sample |
|---|---|---|---|---|---|
| XLLM's SFT (Chen et al., 2023a) | MiniGPT-4, AISHELL-2, VSDial-CN, ActivityNet Caps | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | Template Instructionalization, T2X generation | 10k |
| ChatBridge's SFT (Zhao et al., 2023b) | MSRVTT, AudioCaps, VQAv2, VG-QA... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | Template Instructionalization, GPT generation | 4.4M+209k |
| Macaw-LLM (Lyu et al., 2023) | MSCOCO, Charades, AVSD, VG-QA... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | GPT generation | 69K+50K |
| BuboGPT's SFT | LLaVA, Clotho, VGGSS | Uni-Modal Understanding,Cross-Modal Understanding | Image,Audio,Text | Template Instructionalization, GPT generation | 196K |
| NextGPT's SFT (Wu et al., 2024b) | WebVid, CC3M, AudioCap, Youtube... | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation,Cross-Modal Generation | Image,Video,Audio,Text | Template Instructionalization, GPT generation+retrieval, T2X generation | 20K |
| AnyMal's SFT (Moon et al., 2024) | – | Uni-Modal Understanding | Image,Video,Audio,Text | Manual Annotation, GPT generation | 210K |
| FAVOR's SFT (Sun et al., 2023a) | LLaVA, MSCOCO, Ego4D, LibriSpeech... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | Template Instructionalization, GPT generation | – |
| LEO's SFT (Huang et al., 2024) | ScanQA, SQA3D, 3RScan, CLIPort... | Uni-Modal Understanding,Cross-Modal Understanding | Image,3D,Text,Action | Template Instructionalization, GPT generation | 220k |
| CoDi-2's SFT (Tang et al., 2024b) | MIMIC-IT, LAION-400M, AudioSet, Webvid... | Uni-Modal Understanding,Uni-Modal Generation | Image,Audio,Text | Template Instructionalization | – |
| X-InstructBLIP's SFT (Panagopoulou et al., 2024) | MSCOCO, Clotho, MSVD, Cap3D... | Uni-Modal Understanding | Image,Video,Audio,3D,Text | Template Instructionalization, GPT generation | 1.6M |
| OneLLM's SFT (Han et al., 2024a) | LLaVA-150K, Clotho, Ego4D, NSD... | Uni-Modal Understanding | Image,Video,Audio,3D,ImU, Depth,fMRI,Normal,Text | Template Instructionalization, T2X generation | 2M |
| AVLLM's SFT (Shu et al., 2023) | ACAV100M, VGGSound, WebVid2M, WavCaps... | Uni-Modal Understanding,Cross-Modal Understanding | Video,Audio,Text | GPT generation | 1.4M |
| Uni-IO2's SFT (Lu et al., 2024a) | CC3M, AudioSet, Webvid3m, Omni3D... | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation,Cross-Modal Generation | Image,Video,Audio,Text | Template Instructionalization, GPT generation | 775m |
| ModaVerse's SFT (Wang et al., 2024c) | – | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation | Image,Video,Audio,Text | GPT generation | 2M |
| REAMO's SFT (Zhang et al., 2024b) | – | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | Template Instructionalization | 10K |
| GroundingGPT's SFT (Li et al., 2024i) | Flickr30K, VCR, Activitynet Captions, Clotho... | Uni-Modal Understanding | Image,Video,Audio,Text | GPT Instructionalization | 1M |
| AnyGPT's SFT (Du et al., 2018) | – | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation,Cross-Modal Generation | Image,Audio,Text | GPT Instructionalization, T2X generation | 208K |
| CAT's SFT (Ye et al., 2024b) | VGGSound, AVQA, VideoInstruct100K... | Cross-Modal Understanding | Video,Audio,Text | GPT Instructionalization | 100K |
| AVicuna's SFT (Tang et al., 2024a) | UnAV-100, VideoInstruct100K, ActivityNet Captions, DiDeMo | Uni-Modal Understanding,Cross-Modal Understanding | Video,Audio,Text | Template Instructionalization | 49K |
| M3DBench'SFT (Li et al., 2024b) | Scannet, ScanRefer, ShapeNet... | Uni-Modal Understanding,Cross-Modal Understanding | Image,3D,Text | Template Instructionalization,//GPT Instructionalization | 320k |
| Uni-Moe's SFT (Li et al., 2024g) | LLaVA-Instruct-150K, LibriSpeech, VideoInstruct100K... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | Template Instructionalization, T2X generation | 874K |
| X-VILA's SFT (Ye et al., 2024a) | WebVid, ActivityNetCaption, LLaVA-Instruct-150K... | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation,Cross-Modal Generation | Image,Video,Audio,Text | Template Instructionalization | – |
| EMOVA's SFT (Chen et al., 2024c) | ShareGPT-4o, MSCOCO, LLaVA-Instruct-150K... | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation,Cross-Modal Generation | Image,Audio,Text | Template Instructionalization, GPT Instructionalization, T2X generation | 4.4M |
| VideoLLaMA2's SFT (Cheng et al., 2024b) | AVQA, AVSD, MusicCaps... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Text | Template Instructionalization | 1.5M |
| PathWeave's SFT (Yu et al., 2024a) | VQAV2, MSRVTT, Cap3D... | Uni-Modal Understanding | Image,Video,Audio,3D,Depth | Template Instructionalization, T2X generation | 23.2M |
| Spider's SFT (Lai et al., 2024) | AudioCap, CC3M, Webvid... | Cross-Modal Understanding,Cross-Modal Generation | Image,Video,Audio,Text | Template Instructionalization, GPT Generation | – |
| GMA's SFT (Szot et al., 2024b) | Meta-World,CALVIN, Maniskill... | Uni-Modal Understanding,Cross-Modal Understanding, Cross-Modal Generation | Image,Text,Action | Template Instructionalization | 2.2M |
| OCTAVIUS's SFT (Chen et al., 2024g) | MSCOCO,Bamboo, ScanNet... | Uni-Modal Understanding | Image,3D,Text | Template Instructionalization, GPT Generation | |
| Lyra's SFT (Zhong et al., 2024) | Mini-Gemini, Collected Youtube's Audio | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation | Image,Audio,Text | GPT Generation, T2X Generation | 1.5M |
| video-SALMONN's SFT (Sun et al., 2024a) | LibriSpeech,AudioCaps, LLaVA-Instruct-150K... | Uni-Modal Understanding,Cross-Modal Understanding | Video,Audio,Text | Template Instructionalization, T2X Generation | – |
| Meerkat's SFT (Chowdhury et al., 2024) | VGG-SS,AVSBench, AVQA,MUSIC-AVQA... | Cross-Modal Understanding | Video,Audio,Text | Template Instructionalization, GPT Generation | 3M |
| VITA's SFT (Fu et al., 2024b) | ShareGPT4V,LLaVA-Instruct-150K, ShareGTP4o,ShareGemini... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | T2X Generation | – |
| Baichuan-omni's SFT (Li et al., 2024c) | vFLAN,VideoInstruct100K... | Uni-Modal Understanding,Cross-Modal Understanding | Image,Video,Audio,Text | T2X Generation | – |
| LongVALE-LLM's SFT (Geng et al., 2024) | LongVALE | Uni-Modal Understanding,Cross-Modal Understanding | Video,Audio,Text | GPT Generation | 25.4K |
| UnifiedMLLM's SFT (Li et al., 2024h) | LISA,SmartEdit... | Uni-Modal Understanding,Cross-Modal Understanding, Uni-Modal Generation,Cross-Modal Generation | Image,Video,Audio,Text | Template Instructionalization, GPT Generation | 100K |
| Dolphin's SFT (Anonymous, 2024) | AVQA,Flickr-SoundNet, VGGSound,LLP... | Uni-Modal Understanding,Cross-Modal Understanding | Video,Audio,Text | Template Instructionalization, GPT Generation | – |

Table 3: **The statistics for OmniMLLM's Instruction Data,** including the data sources, interaction forms, involved modalities, and construction methods.

| Name | Capability Category | Modality | Specific-Task | Metrics |
|---|---|---|---|---|
| VQA v2 (Goyal et al., 2017) | Unimodal Understanding | Image,Text | QA | Acc |
| GQA (Hudson and Manning, 2019) | Unimodal Understanding | Image,Text | QA | Acc |
| DocVQA (Mathew et al., 2021) | Unimodal Understanding | Image,Text | QA | Acc |
| IconQA (Lu et al., 2021) | Unimodal Understanding | Image,Text | QA | Acc |
| OCR-VQA (Mishra et al., 2019) | Unimodal Understanding | Image,Text | QA | Acc |
| STVQA (Biten et al., 2019) | Unimodal Understanding | Image,Text | QA | Acc |
| VSR (Liu et al., 2023a) | Unimodal Understanding | Image,Text | QA | Acc |
| Hateful Meme (Kiela et al., 2020) | Unimodal Understanding | Image,Text | QA | AUC |
| OKVQA (Marino et al., 2019) | Unimodal Understanding | Image,Text | QA | Acc |
| VizWiz (Gurari et al., 2018) | Unimodal Understanding | Image,Text | QA | Acc |
| TextVQA (Singh et al., 2019) | Unimodal Understanding | Image,Text | QA | Acc |
| nocap (Agrawal et al., 2019) | Unimodal Understanding | Image,Text | Caption | CIDER |
| ScienceQA (Lu et al., 2022) | Unimodal Understanding | Image,Text | QA | Acc |
| MSCOCO Caption (Lin et al., 2014) | Unimodal Understanding | Image,Text | Caption | CIDER,BLEU |
| Flickr Caption (Plummer et al., 2015) | Unimodal Understanding | Image,Text | Caption | CIDER |
| Visual Dialog (Das et al., 2017) | Unimodal Understanding | Image,Text | Dialogue | MRR |
| RefCOCO (Yu et al., 2016) | Unimodal Understanding | Image,Text | Grounding | Acc |
| RefCOCO+ (Yu et al., 2016) | Unimodal Understanding | Image,Text | Grounding | Acc |
| RefCOCOg (Mao et al., 2016) | Unimodal Understanding | Image,Text | Grounding | Acc |
| A-okvqa (Schwenk et al., 2022) | Unimodal Understanding | Image,Text | QA | Acc |
| POPE (Li et al., 2023b) | Unimodal Understanding | Image,Text | Hallucination | Acc |
| IIIT5K (Mishra et al., 2012) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| IC13 (Karatzas et al., 2013) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| IC15 (Karatzas et al., 2015) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| Total-Text (Chng and Chan, 2017) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| CUTE80 (Risnumawan et al., 2014) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| SVT (Wang et al., 2011) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| SVTP (Phan et al., 2013) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| COCO-Text (Veit et al., 2016) | Unimodal Understanding | Image,Text | OCR | WAC(word ACC) |
| MMB (Liu et al., 2024d) | Unimodal Understanding | Image,Text | Comprehensive Benchmark | GPT ACC |
| MME (Fu et al., 2023) | Unimodal Understanding | Image,Text | Comprehensive Benchmark | GPT ACC |
| LLaVA-Bench (Liu et al., 2023c) | Unimodal Understanding | Image,Text | Comprehensive Benchmark | GPT ACC |
| Mmmu (Yue et al., 2024) | Unimodal Understanding | Image,Text | Comprehensive Benchmark | GPT ACC |
| SEED (Ge et al., 2023) | Unimodal Understanding | Image,Text | Comprehensive Benchmark | GPT ACC |
| MM-Vet (Yu et al., 2024d) | Unimodal Understanding | Image,Text | Comprehensive Benchmark | GPT ACC |
| ActivityNet-QA (Yu et al., 2019) | Unimodal Understanding | Video,Text | QA | Acc |
| MSRVTT-QA (Xu et al., 2016) | Unimodal Understanding | Video,Text | QA | Acc |
| MSVD-QA (Xu et al., 2017a) | Unimodal Understanding | Video,Text | QA | Acc |
| How2QA (Li et al., 2020) | Unimodal Understanding | Video,Text | QA | Acc |
| NExTQA (Xiao et al., 2021) | Unimodal Understanding | Video,Text | QA | ACC |
| STAR (Wu et al., 2021) | Unimodal Understanding | Video,Text | QA | Acc |
| MSVD-Caption (Xu et al., 2017a) | Unimodal Understanding | Video,Text | QA | CIDER |
| VATEX (Wang et al., 2019) | Unimodal Understanding | Video,Text | Caption | CIDER |
| MSRVTT-Caption (Xu et al., 2016) | Unimodal Understanding | Video,Text | Caption | CIDER,BLEU |
| Video-ChatGPT Benchmark (Maaz et al., 2024) | Unimodal Understanding | Video,Text | Comprehensive Benchmark | GPT ACC,GPT Score |
| Kinetics-400 | Unimodal Understanding | Video,Text | Classification | Acc |
| Perception test (Patraucean et al., 2023) | Unimodal Understanding | Video,Text | Comprehensive Benchmark | GPT ACC |
| EgoSchema (Mangalam et al., 2023) | Unimodal Understanding | Video,Text | Comprehensive Benchmark | GPT ACC |
| Mvbench (Li et al., 2024a) | Unimodal Understanding | Video,Text | Comprehensive Benchmark | GPT ACC |
| VideoMME (Fu et al., 2024a) | Unimodal Understanding | Video,Text | Comprehensive Benchmark | GPT ACC |
| Charades-STA (Sigurdsson et al., 2016) | Unimodal Understanding | Video,Text | Grounding | IoU |
| AudioCaps (Kim et al., 2019) | Unimodal Understanding | Audio,Text | Caption | CIDER,SPICE,METEOR,BLEU,SPIDER |
| ClothoAQA (Lipping et al., 2022) | Unimodal Understanding | Audio,Text | QA | Acc |
| Vocalsound (Gong et al., 2022) | Unimodal Understanding | Audio,Text | QA | Acc |
| Clotho v1 (Drossos et al., 2020) | Unimodal Understanding | Audio,Text | Caption | CIDER |
| Clotho v2 (Drossos et al., 2020) | Unimodal Understanding | Audio,Text | Caption | CIDER |
| ESC50 (Piczak, 2015) | Unimodal Understanding | Audio,Text | Classification | Acc |
| LibriSpeech (Panayotov et al., 2015) | Unimodal Understanding | Audio,Text | ASR | WER |
| AISHELL-2 (Du et al., 2018) | Unimodal Understanding | Audio,Text | ASR | WER |
| Wenetspeech (Zhang et al., 2022a) | Unimodal Understanding | Audio,Text | ASR | WER |
| MusicCap (Agostinelli et al., 2023) | Unimodal Understanding | Audio,Text | Caption | CLAP Score |
| TUT2017 (Mesaros et al., 2016) | Unimodal Understanding | Audio,Text | Classification | Acc |
| EHSL (Li et al., 2024g) | Unimodal Understanding | Audio,Text | QA | Acc |
| Cap3D Caption (Luo et al., 2023b) | Unimodal Understanding | 3D,Text | Caption | CIDER |
| Objaverse Caption (Deitke et al., 2023) | Unimodal Understanding | 3D,Text | Caption | METEOR,ROUGE,BLEU |
| Cap3D QA (Luo et al., 2023b) | Unimodal Understanding | 3D,Text | QA | Acc |
| Objaverse Classification (Deitke et al., 2023) | Unimodal Understanding | 3D,Text | Classification | GPT ACC |
| Modelnet40 (Wu et al., 2015) | Unimodal Understanding | 3D,Text | Classification | Acc |
| ScanRefer (Chen et al., 2020a) | Unimodal Understanding | 3D,Text | Grounding | mAP |
| Nr3D (Achlioptas et al., 2020) | Unimodal Understanding | 3D,Text | Caption | BLEU,CIDER,METEOR,ROUGE-L |
| SQA3D (Ma et al., 2023) | Unimodal Understanding | 3D,Text | QA | Acc |
| ScanQA (Azuma et al., 2022) | Unimodal Understanding | 3D,Text | QA | Acc |
| SUN RGB-D (Song et al., 2015) | Unimodal Understanding | Depth,Text | Classification | Acc |
| NYUv2 (Silberman et al., 2012) | Unimodal Understanding | Depth,Text | Classification | Acc |
| SUN RGB-D_generated Nomral (Han et al., 2024a) | Unimodal Understanding | Normal,Text | Classification | Acc |
| NYUv2_generated Nomral (Han et al., 2024a) | Unimodal Understanding | Normal,Text | Classification | Acc |
| ThermalQA (Yu et al., 2024c) | Unimodal Understanding | Thermal, Text | QA | Acc |
| TochQA (Yu et al., 2024c) | Unimodal Understanding | Touch, Text | QA | Acc |
| Ego4D (Grauman et al., 2022) | Unimodal Understanding | IMU,Text | Caption | CIDER,ROUGE |
| NSD (Allen et al., 2022) | Unimodal Understanding | fMRI,Depth Map,Text | Caption | CIDER,ROUGE |
| MSCOCO (Lin et al., 2014) | Unimodal Generation | Image,Text | TX2X Edit | FID,CLIPSIM |
| MSRVTT (Xu et al., 2016) | Unimodal Generation | Video,Text | T2X Generate | CLIPSIM |
| AudioCaps (Kim et al., 2019) | Unimodal Generation | Audio,Text | T2X Generate,TX2X Edit | FAD |
| DAVIS (Perazzi et al., 2016) | Unimodal Generation | Video,Text | TX2X Edit | CLIPSIM |
| UCF-101 (Soomro et al., 2012) | Unimodal Generation | Video,Text | T2X Generate | FID,FVD,IS,CLIPSIM |
| Evalcrafter (Liu et al., 2024c) | Unimodal Generation | Video,Text | T2X Generate | FVD,CLIPSIM |
| VCTK (Veaux et al., 2017) | Unimodal Generation | Audio,Text | T2X Generate,TX2X Edit | WER,MCD |
| MusicCap (Agostinelli et al., 2023) | Unimodal Generation | Audio,Text | T2X Generate | FAD |
| Dreambench (Ruiz et al., 2023) | Unimodal Generation | Image,Text | T2X Generate | CLIP-I,CLIP-T,DINO |
| MUSIC-AVQA (Li et al., 2022b) | Crossmodal Understanding | Video,Audio,Text | QA | Acc |
| AVSD (AlAmri et al., 2018) | Crossmodal Understanding | Video,Audio,Text | Dialogue | CIDER,BLEU |
| RACE-Audio (Li et al., 2024g) | Crossmodal Understanding | Image,Audio,Text | Comprehensive Benchmark | Acc |
| VALOR Caption (Chen et al., 2023e) | Crossmodal Understanding | Video,Audio,Text | Caption | CIDER,BLEU |
| MMBench-Audio (Li et al., 2024g) | Crossmodal Understanding | Image,Audio,Text | Comprehensive Benchmark | Acc |
| AVQA (Li et al., 2022a) | Crossmodal Understanding | Video,Audio,Text | QA | Acc |
| MCUB (Chen et al., 2024a) | Crossmodal Understanding | Image,Video,Audio,3D,Text | Comprehensive Benchmark | Acc |
| DisCRn (Panagopoulou et al., 2024) | Crossmodal Understanding | Image,Video,Audio,3D | Comprehensive Benchmark | Acc |
| OmniXR (Chen et al., 2024d) | Crossmodal Understanding | Image,Video,Audio,Text | Comprehensive Benchmark | Acc |
| Curse (Leng et al., 2024) | Crossmodal Understanding | Image,Video,Audio,Text | Hallucination | Acc |
| ISQA (Sun et al., 2023a) | Crossmodal Understanding | Image,Audio,Text | QA | Acc |
| VGGSound (Chen et al., 2020b) | Crossmodal Understanding | Video,Audio,Text | QA | Acc |
| VATEX (Wang et al., 2019) | Crossmodal Understanding | Video,Audio,Text | Caption | CIDER |
| UnAV-100 (Geng et al., 2023) | Crossmodal Understanding | Video,Audio,Text | Ground | IoU |
| LLP (Tian et al., 2020) | Crossmodal Understanding | Video,Audio,Text | Ground | IoU |
| Presentation-QA (Sun et al., 2024a) | Crossmodal Understanding | Video,Audio,Text | QA | ACC |
| LongVALE Caption | Crossmodal Understanding | Video,Audio,Text | Caption | CIDER |
| TVL Benchmark (Fu et al., 2024c) | Cross-modal Understanding | Touch,Image,Text | QA | ACC |
| AVEB (Sun et al., 2023a) | Crossmodal Understanding,Unimodal Understanding | Image,Video,Audio,Text | Comprehensive Benchmark | ACC,METEOR,SPIDER,WER |
| XtoX Benchmark (Ye et al., 2024a) | Crossmodal Understanding,Crossmodal Generation | Image,Video,Audio,Text | Comprehensive Benchmark | X-to-X Alignment Score |

Table 4: **An overview of benchmarks and tasks of Omni-MLLMs**, including the abilities being evaluated, the involved modalities, specific tasks, and evaluation metrics.

| Model | LLM | Uni-Modal Understanding | | | | | | | | Uni-Modal Generation | | | Cross-Modal Understanding | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MSVD-QA | MSRVTT-QA | $VQA^{v2\ test}$ | Flickr | $MMB^{en}$ | $AudioCaps^{cap\ test}$ | ClothoAQA | Objaverse | $COCO^{gen}$ | $AudioCaps^{gen}$ | $MSRVTT^{gen}$ | VGGSS | AVSD | MUSIC-AVQA | AVQA |
| **Omni-MLLMs** | | | | | | | | | | | | | | | | |
| eP-ALM | OPT-2.7B | 38.4 | 38.51 | 54.47 | – | – | 61.86 | – | – | – | – | – | – | – | – | – |
| ChatBridge 13B | Vicuna-13B | 45.3 | – | – | 82.5 | – | – | – | – | – | – | – | – | | 43 | – |
| PandaGPT | Vicuna-13B | 46.7 | 23.7 | – | – | – | – | – | – | – | – | – | 32.7 | 26.1 | 33.7 | 79.8 |
| Video-LLaMA | Vicuna-7B | 51.6 | 29.6 | – | – | – | – | – | – | – | – | – | 40.8 | 36.7 | 36.6 | 81 |
| Macaw-LLM | LLAMA-7B | 42.1 | 25.5 | – | – | 3.84 | 33.3 | – | – | – | – | – | 36.1 | 34.3 | 31.8 | 78.7 |
| ImageBind-LLM | LLaMA-7B | – | – | – | 23.49 | – | – | 10.3 | 31 | – | – | – | – | – | 39.72 | 54.26 |
| NExT-GPT | Vicuna-7B | 64.5 | 58.4 | 66.7 | 84.5 | 58 | 81.3 | – | – | 10.07 | 8.67 | 31.97 | – | – | – | – |
| AnyMAL 13B | LLaMA2-13B | – | – | 59.6 | – | – | – | – | – | – | – | – | – | – | – | – |
| AnyMAL 70B | LLaMA2-70B | – | – | 64.2 | 95.9 | – | 77.8 | – | – | – | – | – | – | – | – | – |
| X-InstructBLIP 7B | Vicuna-7B | 51.7 | 41.3 | 30.61 | 82.1 | 8.96 | 67.9 | 15.4 | 50 | – | – | – | – | – | 28.1 | – |
| X-InstructBLIP 13B | Vicuna-13B | 49.2 | – | – | 74.7 | – | 53.7 | 21.7 | – | – | – | – | 20.3 | 52.1 | 44.5 | 44.23 |
| OneLLM 7B | LLaMA2-7B | 56.5 | – | 71.6 | 78.6 | 60 | – | 57.9 | 44.5 | – | – | – | – | – | 47.6 | – |
| AV-LLM | Vicuna-7B | 67.3 | 53.7 | – | – | – | 35.5 | – | – | – | – | – | 47.6 | 52.6 | 45.2 | – |
| UIO-2xxl 6.8B | – | 52.2 | 41.5 | 79.4 | – | 71.5 | 48.9 | – | – | 13.39 | 5.89 | – | – | – | – | – |
| ModaVerse | Vicuna-7b | – | 56.5 | – | – | – | 79.2 | – | – | 11.24 | 8.22 | 30.14 | – | – | – | – |
| CREMA 7B | Mistral-7B | – | – | – | – | – | – | – | – | – | – | – | – | – | 52.6 | – |
| GroundingGPT | Vicuna-7B | 67.8 | 51.6 | 78.7 | – | 63.8 | – | – | – | – | – | – | – | – | – | – |
| NaiveMC | Vicuna-7B | – | – | – | – | – | – | – | 55 | – | – | – | – | – | 53.63 | 80.7 |
| DAMC | Vicuna-7B | – | – | – | – | – | – | – | 60.5 | – | – | – | – | – | 57.32 | 81.31 |
| AnyGPT | LLaMA2-7B | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| CAT | LLaMA2-7B | – | 62.7 | – | – | – | – | – | – | – | – | – | – | 48.6 | 92 | |
| AVicuna | Vicuna-7B | 70.2 | 59.7 | – | – | – | – | – | – | – | – | – | – | 53.1 | 49.6 | – |
| Uni-MoE | LLaMA-7B | 55.6 | – | 66.2 | – | 69.82 | – | 32.6 | – | – | – | – | – | – | – | – |
| X-VILA 7B | Vicuna-7B | – | – | 72.9 | – | – | – | – | – | – | – | – | – | – | – | – |
| VideoLLaMA2-7B | Mistral-7B | 71.7 | – | – | – | – | – | – | – | – | – | – | 71.4 | 57.2 | 80.9 | – |
| Meerkat | Llama-2-7B-Chat | – | – | – | – | – | – | – | – | – | – | – | – | – | – | 87.14 |
| InternOmni | InternLM-2-Chat-7B | – | – | – | – | 81.7 | – | – | – | – | – | – | – | – | – | – |
| UnifiedMLLM | Vicuna-7B | – | – | – | – | – | – | – | – | 10.84 | 9.95 | 31.20 | – | – | – | – |
| VITA | Mixtral-8x7B | – | – | – | – | 71.8 | – | – | – | – | – | – | – | – | – | – |
| EMOVA | LLaMA-3.1-8B | – | – | – | – | 82.8 | – | – | – | – | – | – | – | – | – | – |
| BaiChuan-omni-7B | – | 72.2 | – | – | – | 76.2 | – | – | – | – | – | – | – | – | – | – |
| OMCAT | Vicuna-7B | – | – | – | – | – | – | – | – | – | – | – | – | 49.4 | 73.8 | 90.2 |
| PathWeave-7B | Vicuna-7B | 47.8 | 37.4 | – | – | – | 64 | 33.5 | – | – | – | – | – | – | – | – |
| Spider | Llama-2-7B | – | – | – | – | – | 81.7 | – | – | 11.23 | 8.18 | 30.97 | – | – | – | – |
| **Specifc-MLLMs** | | | | | | | | | | | | | | | | |
| VILA-7B | LLaMA-2-7B | – | – | 79.9 | 74.7 | 68.9 | – | – | – | – | – | – | – | – | – | – |
| VideoChat2 | Vicuna-7B | 70 | 54.1 | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Qwen-Audio | Qwen-7B | – | – | – | – | – | – | 57.9 | – | – | – | – | – | – | – | – |
| PointLLM | Vicuna-7B | – | – | – | – | – | – | – | 47.5 | – | – | – | – | – | – | – |
| Emu-13B | – | – | – | 52 | – | – | – | – | – | 11.66 | – | – | – | – | – | – |
| Video-LaVIT | Llama2-7B | 73.2 | – | 80.3 | – | 67.3 | – | – | – | – | – | 30.12 | – | – | – | – |

Table 5: **The performance of Omni-MLLMs on different benchmarks.** The selected uni-modal understanding benchmarks include Video-Text2Text (Xu et al., 2017a, 2016), Image-Text2Text (Goyal et al., 2017; Plummer et al., 2015; Liu et al., 2024d), Audio-Text2Text (Kim et al., 2019; Lipping et al., 2022), and 3D-Text2Text (Deitke et al., 2023). The chosen uni-modal generation benchmarks include Text2Image (Lin et al., 2014), Text2Video (Xu et al., 2016), and Text2Audio (Kim et al., 2019). The selected cross-modal understanding benchmarks are Image-Audio-Text2Text (Chen et al., 2020b; AlAmri et al., 2018) and Video-Audio-Text2Text (Li et al., 2022b,a).