Advances in Large Multi-Modal Models from the Perspective of Representation Space Extension: A Survey

Anonymous ACL submission

Abstract

The success of large language models (LLMs) has attracted much focus on extending these models to multi-modal domains, giving rise to large multi-modal models (LMMs). Unlike existing reviews that focuses on specific model frameworks or scenarios, this paper aims to provides an encyclopedic survey on LMMs from a general perspective, i.e. **representation space extension**. By systematically analyzing the input-output representations of existing LMMs, this paper summarizes the design of model architectures to align the constructed multi-modal representation space. Lastly, this paper demonstrates the extensibility of LMMs as embodied agents in view of proposed representation space extension. With the insights revealed through surveying the field, this paper discusses several fundamental problems of constructing LMMs and inspires future work at the end.

1 Introduction

The goal of AI research is to build versatile intelligent systems capable of fulfilling tasks across diverse scenarios. Recently, the generalization and interactivity demonstrated by large language models (LLMs) have significantly advanced the progress towards general-purpose AI (OpenAI, 2023; Touvron et al., 2023b; Bai et al., 2023a; AI@Meta, 2024). To adapt these capabilities to multi-modal contexts, research on large multi-modal models (LMMs) is emerging, aiming to extend the input and output representation space of the language-based interface to more modalities. As shown in Figure 1, to extend the input space, existing methods introduce discretely or continuously encoded modality representations into the text input and learn cross-modal alignment from multi-modal intertwined data, enabling LMMs to understand multi-modal information (Li et al., 2023b; Liu et al., 2024b; Bai et al., 2023b; Ma et al., 2024). Similarly, the output space can be divided into multiple subspaces

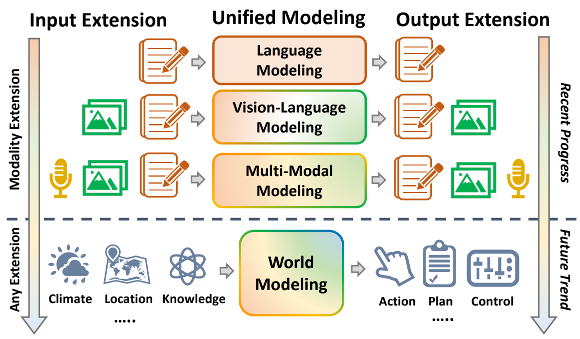

Figure 1: Illustration of the general LMM framework: expanding input and output representation space to more modalities and aligning representations across modalities through unified multi-modal modeling.

of different modalities, which are further aligned with corresponding modality decoders to generate multi-modal content (Koh et al., 2024; Zhang et al., 2024a; Wu et al., 2023d; Zhan et al., 2024a).

Although there are several surveys that detail the current progress in constructing LMMs, most of these works are limited to specific sub-problems in the construction of LMMs, such as applications in specific modalities (Tang et al., 2023b; Latif et al., 2023) and scenarios (Xiao et al., 2024; Cui et al., 2024). Meanwhile, most existing reviews focus on a specific type of model framework: encoding information from other modalities in a continuous manner and aligning them with text embeddings through connection modules (Wu et al., 2023c; Caffagni et al., 2024), neglecting related research on other architectures, such as unified discretely represented LMMs (Team, 2024; Zhan et al., 2024a). These limitations prevent existing reviews from adequately covering research problems in LMM construction and limit their applicability.

To this end, this survey aims to summarize related works from a more general perspective: **the extension of input-output representation space**. As illustrated in Figure 1, existing LMMs can be systematically summarized from this view, encompassing various modalities, scenarios, and model architectures, while also leaving room for further exploration to more modalities and scenarios. We omit the details about data and evaluations that have been sufficiently reviewed by previous works (Li and Lu, 2024; Huang and Zhang, 2024; Bai et al., 2024), but keep a tight focus on the architectures.

To conduct a holistic survey, we follow a top-down logic to break down the construction of LMMs into several sub-problems, providing detailed discussions to offer insights to readers. Particularly, we try to answer the following questions. (i) How can modality signals be encoded using discrete or continuous representations, and how to construct multi-modal representation spaces? (§2) (ii) How to design model architectures to align the constructed multi-modal representation space? (§3) (iii) How to extend the representation space to real scenarios, i.e. embodied agents? (§4) This further demonstrates the extensibility of LMMs from the perspective discussed in this paper. Finally, in §5, we summarize the discussion on the questions raised above, providing readers with key take-home messages and an outlook on future research.

In summary, our contributions are threefold:

- Going beyond specific scenarios and model framework, we review the current LMMs from a general perspective of input-output representation space extension.
- Based on the structure of input-output spaces, we systematically review the existing models, including mainstream models based on discrete-continuous hybrid spaces and models with unified multi-modal discrete representations. Furthermore, we summarize the design of model architectures to align the constructed multi-modal representation space.
- We elaborate on how to extend LMMs to embodied scenarios to highlight the extensibility of LMMs from the input-output extension perspective. To the best of our knowledge, this is the first survey to include embodied LMMs.

2 Representation Space Extension

In this section, we introduce prevalent solutions to construct multi-modal representation space. As illustrated in Figure 2, existing methods can be categorized based on different input-output space structures, and the extension to other modalities can be summarized in a similar manner.

2.1 Encode Input Representation

Regarding the input, the core research problems involve how to code the representations of each modality and how to integrate them into a multi-modal input space (see lower part in Figure 2).

2.1.1 Textual Representation

As a discrete signal, text is a sequence composed of characters. Following the practice of LLMs, LMMs typically utilize tokenizers, such as BPE (Sennrich et al., 2015; Radford et al., 2019), WordPiece (Wu, 2016), and Unigram (Kudo, 2018), to merge characters into sub-word tokens. Ultimately, texts are represented as sequences of discrete tokens.

2.1.2 Visual Representation

For visual signals with spatial-temporal information, LMMs mainly employ pre-trained visual encoders for representing images (videos) into continuous features or discrete codes. Figure 5 in Appendix A shows the evolution of visual encoders.

Commonly adopted architectures of visual encoders can be divided into two categories: convolution-based (He et al., 2016; Liu et al., 2022b) and vision-Transformer-based models (Dosovitskiy et al., 2020; Liu et al., 2021). Both methods encode images into continuous 2D feature maps. These continuous features can be further compressed into discrete visual codes through vector quantization (VQ) by learning a fixed-size visual codebook (Van Den Oord et al., 2017; Esser et al., 2021). In addition, models like Fuyu (Bavishi et al., 2023) do not rely on visual encoders and directly use pixel values of image patches as the visual representations.

Based on the sequence modeling framework of current LMMs, multiple images can be intuitively arranged in the input sequence (Luo et al., 2023b; Zhang et al., 2023b; Li et al., 2023a; Yu et al., 2024b). For videos, where images (frames) are temporally related, spatial-temporal encoders such as TimeSformer (Bertasius et al., 2021) and VideoSwin (Liu et al., 2022a) can be further used for encoding (Li et al., 2023c; Xu et al., 2023).

2.1.3 Multi-Modal Representation

As illustrated in the lower part Figure 2, there exist two mainstream types of multi-modal input space.

Type A: Hybrid Input Space Text are represented in a discrete form, while visual signals are encoded in continuous representations, preseving the complete visual information. However, due to

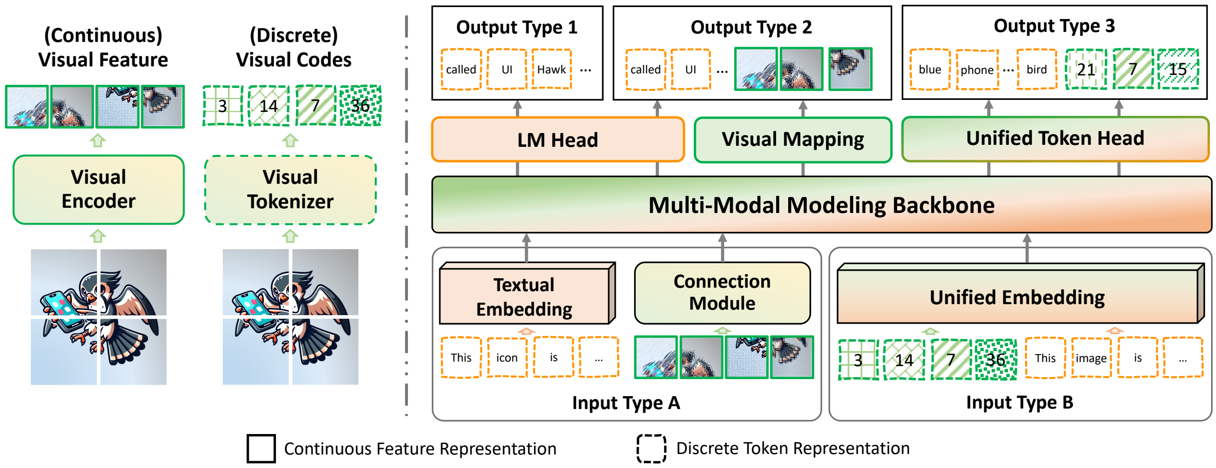

Figure 2: Summary and illustration of different input-output space structures for extension to vision modality.

the gap in the input space, connection modules are required to perform input-level cross-modal alignment, which is discussed in Section 3.

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

2.1.4 Extension to More Input Modalities

Beyond the vision modality, signals from other modalities can be encoded and introduced into the input space following a similar paradigm. For example, various encoders can help encode audio into continuous (Hsu et al., 2021; Elizalde et al., 2023; Girdhar et al., 2023) or discrete (Zhang et al., 2023e) representations. As a step further, an arbitrary-modality input space can be represented in either hybrid (Wu et al., 2023d; Han et al., 2023; Tang et al., 2023c; Lu et al., 2023a) or unified discrete forms (Zhan et al., 2024a).

2.2 Decode Output Representation

Based on the input, backbones of LMMs present continuous multi-modal output representations which can be used to decode the output signals of different modalities. For example, with the commonly used causal modeling framework, the output representation can be leveraged to predict the signal at the next position in the sequence. Predicted token sequence can be converted to text with the tokenizer while different image generator can be adopted to decode images from output in different forms. In this section, we discuss the commonly adopted paradigms to partition the output space of different modalities and perform corresponding decoding, as shown in the upper part of Figure 2.

2.2.1 Type 1: Text-Only Output Space

If only text output is required, similar to LLMs, discrete tokens can be generated from the ouput representations through a classification-based language modeling (LM) head and specific decoding strategies (Li et al., 2023b; Liu et al., 2024b). Please note that models that first generate text descriptions and then use external tools like Stable Diffusion and CLIP to generate or retrieve content in other modalities, such as Visual ChatGPT (Wu et al., 2023a), InternLM-XComposer series (Zhang et al., 2023c; Dong et al., 2024b), and Mini-Gemini (Li et al., 2024d), are also classified as text-only output models because they are not in an end-to-end manner.

2.2.2 Type 2: Hybrid Output Space

The hybrid output space includes the discrete text tokens and continuous visual features. Such output space is initially proposed to support image generation. A series of methods first introduce special tokens, such as the start and end tokens for images, or a series of consecutive placeholder tokens to indicate where images should be generated. The continuous visual representations at the corresponding positions are then connected to visual decoders (mainly Diffusion models (Rombach et al., 2022)) through visual mapping modules (Koh et al., 2024; Dong et al., 2024a; Zheng et al., 2024b; Sun et al., 2024b). Similar to the hybrid input space, visual mapping modules perform output-level alignment and typically requires further training.

2.2.3 Type 3: Unified Discrete Output Space

The unified discrete output space contains discrete text tokens and discrete visual codes. Based on the joint vocabulary constructed within Type B input space described in Section 2.1.3, image generation is naturally integrated into the token decoding process. The predicted visual codes are fed to the corresponding codebook detokenizer to generate the image (Ge et al., 2023b; Team, 2024).

2.2.4 Extension to More Output Modalities

Type 2 and Type 3 output spaces can be expanded to support arbitrary-modality output. For example, Next-GPT (Wu et al., 2023d) and Codi-2 (Tang et al., 2023c) further extends the hybrid output space, while AnyGPT (Zhan et al., 2024a) and UnifiedIO-2 (Lu et al., 2023a) construct unified discrete spaces for all modalities.

2.3 Prevalent Representation Paradigms

Considering the representation space structures introduced above, most existing LMMs can be categorized to three types: (1) **Multi-modal understanding models** that rely on Type A input and Type 1 output, these models are mainly designed for understanding tasks that can be fully expressed in language (Dai et al., 2023; Bai et al., 2023b; Lu et al., 2024a; Chen et al., 2023f); (2) **Multi-modal generation models** which comprise of Type A input and Type 2 output, such models excel in generating multi-modal interleaved responses based on the context (Koh et al., 2024; Wu et al., 2023d; Sun et al., 2024a); (3) **Unified multi-modal models** that represent and generate multiple modalities in a unified discrete form (Ge et al., 2023b; Zhan et al., 2024a; Team, 2024). Table 1 and Table 2 in appendix list the design paradigms of contemporary LMMs, grouped according to the aforementioned classification criteria. The alignment architectures discussed in Section 3 are also included.

3 Multi-Modal Alignment Architecture

Based on the multi-modal representation spaces introduced in Section 2, the design of LMMs needs to consider how to align representations across different modalities. Mainstream architectures take an LLM-centric paradigm: aligning inputs from all modalities to a **unified multi-modal backbone** for interaction and generating multi-modal responses. To facilitate the unified modeling, additional modules are required, as summarized in Figure 3. We detail the architecture as follows.

3.1 Multi-Modal Modeling Backbone

Typically, the backbone is based on a decoder-only architecture composed of multiple transformer blocks (Vaswani et al., 2017). To better understand language, the backbone is primarily initialized with a pre-trained LLM, such as LLaMA (Touvron et al., 2023a,b; Dubey et al., 2024), Vicuna (Chiang et al., 2023), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023a; Yang et al., 2024), and so on (Bi et al., 2024; Cai et al., 2024; Young et al., 2024). In addition, LMMs for edge devices are usually initialized with smaller language models, such as MobileLLaMA (Kan et al., 2024), Phi (Abdin et al., 2024), etc. (Team et al., 2024a; Hu et al., 2024a). The backbone can also inherit MoE-based language models like Mixtral 8x7B (MistralAITeam, 2023).

Apart from the commonly used architecture mentioned above, some LMMs adopt encoder-decoder backbones (Chung et al., 2024; Chen et al., 2023e; Lu et al., 2023a; Bachmann et al., 2024; Mizrahi et al., 2023). Additionally, native LMMs like Chameleon (Team, 2024) are not initialized with pre-trained LLMs and trained from scratch.

3.2 Input-level Alignment

To enable the backbone to process multi-modal information uniformly, it is necessary to align the form and space of inputs across modalities at the input level. Specifically, for Type B input space, since all modalities are represented in a unified discrete token form, input-level alignment can be achieved by directly merging the vocabularies of multiple modalities and learning the token representations through subsequent alignment training (Ge et al., 2023b; Team, 2024; Zhan et al., 2024a).

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

MLP Based A typical connection module is multi-layer perceptron (MLP). This module directly aligns the dimension of representations from other modalities with text (Liu et al., 2024b, 2023) by flattening the 2D or 3D features into 1D in a specific order (Maaz et al., 2023; Wu et al., 2023d; Liu et al., 2024a). The advantage of MLP-based module lies in the simplicity and fast convergence during alignment training. However, MLP-based module cannot compress redundant information, which

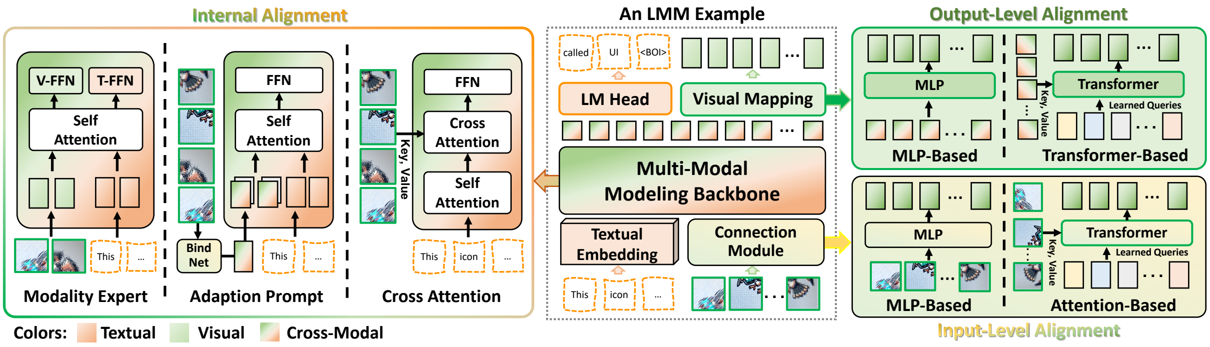

Figure 3: **Summarization of multi-modal alignment modules[1]:** (1) **input-level alignment** to unify the multi-modal inputs into a consistent form and space; (2) **internal alignment** of the backbone for complex cross-modal interactions; and (3) **output-level alignment** to map the outputs to different modality decoders.

could result in excessively long representation sequences (e.g., for high-resolution images). Reducing the computational efficiency requires additional designs to compress the information (Zhu et al., 2023a; Chen et al., 2024c; Dong et al., 2024c).

Attention Based Another prevalent connection modules are based on attention mechanisms. This method introduces a fixed number of learnable vectors as queries, which retrieve relevant information from other-modality representations (serving as keys and values) through cross-attention. The output representations of the queries, enriched with information from other modalities, serve as the modality input to the backbone. Representative module architectures include Q-Former (Li et al., 2023b; Dai et al., 2023), abstractor (Ye et al., 2023b,c), resampler (Zeng et al., 2023; Li et al., 2023e), and so on (Bai et al., 2023b; Zhang et al., 2023c). The query-level representations obtained from attention mechanisms effectively compress and aggregate information from other modalities. Additionally, recent works have demonstrated further extensibility, including integrating representations from multiple encoders (Li et al., 2024d; Kar et al., 2024; Tong et al., 2024), incorporating local grounding information (Lu et al., 2023b), and scaling up to an 8B Q-LLaMA (Chen et al., 2023f). However, these modules mainly involve many parameters and typically require additional training (Li et al., 2023b; Lu et al., 2023b). Besides, Yao et al. have found that attention-based modules may result in the loss of important information.

Others In additional to the mainstream structures mentioned above, several other connection modules have been proposed. CNN-based modules utilize the inductive bias of convolutional operations to model local information, further combined with pooling layers, the number of resulted tokens can be effectively reduced (Cha et al., 2024; Chu et al., 2024a; Hong et al., 2024). Adaptive pooling-based modules can compress features using spatial relationships without introducing additional parameters (Yao et al., 2024; Xu et al., 2024a). Furthermore, VL-Mamba explores to use vision selective scanning as connection to integrate representations across different modalities (Qiao et al., 2024).

3.3 Internal Alignment

Researchers have explored introducing extra parametric modules within the backbone to further enhance the alignment between modalities. We summarize the commonly adopted methods as follows:

Cross-Attention Layer Flamingo (Alayrac et al., 2022) is the first to insert cross-attention layers between the original layers of the backbone, allowing text to perceive information from the visual context. And a tanh gating is introduced to control the degree of modality fusion. Subsequently, such design has been widely adopted by recent LMMs (Gong et al., 2023; Awadalla et al., 2023; IDEFICS, 2023; Chen et al., 2024a), CogAgent (Hong et al., 2023a) further utilizes cross-attention to supplement high-resolution image information. Although effective, densely inserted cross-attention layers bring a large number of additional parameters. Ye et al. improve this by introducing sparsely inserted hyper attention, which significantly reduce extra parameters and facilitate model convergence through parallel self-attention and cross-attention calculation.

Adaption Prompt LLaMA-Adapter incorporates visual representations into lightweight learnable adaption prompts and feed the prompts as pre-

[1] The illustration of the input-level and output-level alignment modules is inspired by (Yin et al., 2023).

fix contexts to the backbone (Zhang et al., 2023d). LLaMA-Adapter V2 (Gao et al., 2023) improves this method with an early knowledge fusion strategy. ImageBind-LLM (Han et al., 2023) further extends the prompts to support more modalities.

Visual Expert To distinguish between visual and textual modeling, some LMMs introduce visual expert modules to process visual tokens. Specifically, CogVLM (Wang et al., 2023a) adds additional attention and FFN layers to process visual tokens without compromising the original textual modeling capabilities of backbones. mPLUG-Owl2 (Ye et al., 2023c) only introduces modality-specific parameter blocks in the normalization layers and the K and V mapping layers of the attention modules. InternLM-XComposer2 (Dong et al., 2024b), on the other hand, designs a lightweight Partial LoRA module for additional modeling of visual tokens.

3.4 Output-level Alignment

Regarding the multi-modal output space described in Section 2.2, both Type 1 and Type 3 are represented in a unified discrete token-based form, multimodal content can be intuitively generated through a next-token prediction approach with the help of modality-specific de-tokenizer (Zhan et al., 2024a; Lu et al., 2023a; Team, 2024; Ge et al., 2023b).

For the Type 2 hybrid output space, additional mapping modules are required to align the output space of LMM backbones with the input space of corresponding modality generators. Considering image generation, commonly used modules are built on linear projection (Dong et al., 2024a) or the transformer architecture (Koh et al., 2024; Zheng et al., 2024b). Similar to Q-Former, transformer-based modules learn a fixed number of queries to retrieve information from the LMM outputs through cross-attention, serving as the condition input of image diffusion models (Rombach et al., 2022). NextGPT (Wu et al., 2023d) expands the transformer-based mapping modules to fit the diffusion generators for image, video and audio modalities. Additionally, Emu series (Sun et al., 2024b,a) replace the linear projection with cross-attention in diffusion models to perform dimensional conversion.

In summary, we detail the architectural designs of current large vision-language models (LVLMs) in Table 1. Similarly, such alignment architectures can be extended to more modalities, as presented in Table 2. We kindly refer readers to Appendix B for how to train the constructed models.

4 Extension to Embodied Agents

Beyond modality extension, the representation space can be expanded to include various forms of signals in different scenarios, such as embodied environment. In this section, we introduce how to expand LMMs into embodied agents with the intelligence to interact with environments. We will firstly introduce categories of embodied tasks, then delve into how to adapt LMMs to embodied tasks by extending the representation spaces.

4.1 Embodied Tasks

Tasks are referred to as "embodied" because the agent needs to interact with a real or virtual environment. Based on the complexity of the interaction actions, we categorize embodied tasks as follows: (1) ***Embodied Question Answering (EQA)*** (Das et al., 2018; Gordon et al., 2018): In these tasks, the agent is required to answer user questions based on environment exploration. Broadly speaking, we consider such action spaces as discrete vocabularies. (2) ***Vision-and-Language Navigation (VLN)*** (Anderson et al., 2018; Krantz et al., 2020): These tasks involve navigation based on user instructions. However, these tasks do not require interactions with objects. Therefore, the action space is either discrete directional movements, such as forward, backward, left, and right, or it can involve continuous control parameters, such as speed and direction. (3) ***Vision-and-Language Manipulation (VLM)*** (Shridhar et al., 2020; Padmakumar et al., 2022; Yenamandra et al., 2023): These tasks require the agent to not only engage in question-answer dialogues with the user, but also navigate the environment and interact with objects based on user instructions. This action space builds upon the action space of VLN tasks by adding object manipulation actions. (4) ***Open-World Robot Control (ORC)*** (Gupta et al., 2019; Mees et al., 2022; Padalkar et al., 2023): In these tasks, agents are equipped with high-degree-of-freedom robotic arms, capable of performing precise object manipulations, such as grasping and moving objects. The action space for ORC tasks is continuous and determined by the complexity of the robotic arm movements, i.e. represented by a set of continuous values, such as the joint angles or velocities.

4.2 Input Extension: Environment

Since embodied agents interact with the environment as the subject, the egocentric observation be-

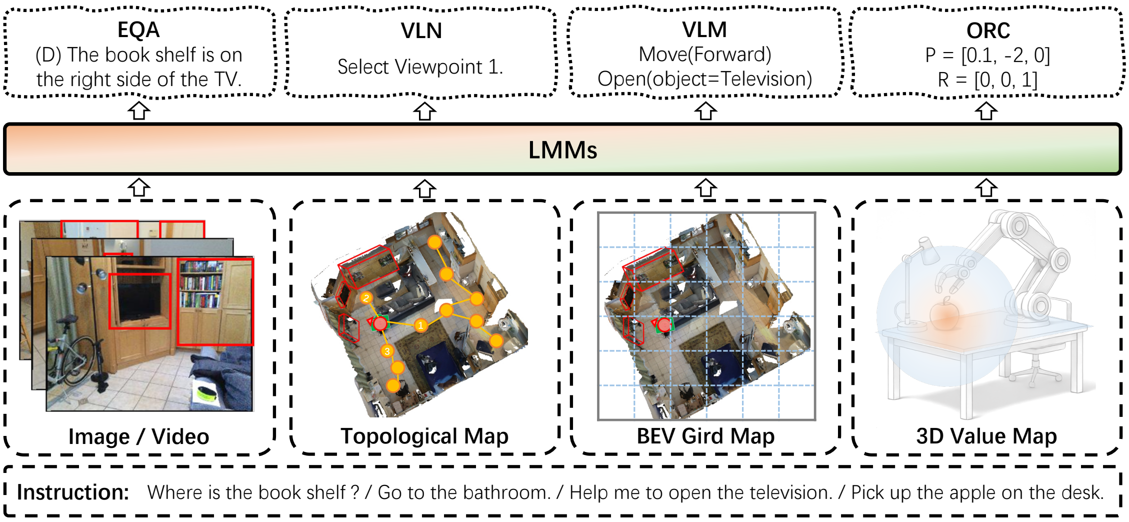

Figure 4: Examples of the input-output space for embodied tasks. Typically, the input includes the user instruction, the current observation (image or video), the environment and the history (optional). We omit the history here, as for different tasks, the content of history vary from pure texts to observation sequences, action sequences and so on.

comes an essential choice (Anderson et al., 2018; Chen et al., 2019; Qi et al., 2020; Padmakumar et al., 2022; Du et al., 2024). Under egocentric observations, the environment is often represented as a local image (Fried et al., 2018; Ahn et al., 2022; Driess et al., 2023) corresponding to the current orientation or by rotating 360 degrees, which could be satisfactory for EQA tasks. However, VLN and VLM tasks require an integrated understanding of observed environments. To obtain a complete picture, the agent must engage in thorough and repeated exploration of the environment (Chaplot et al., 2020a,b). Therefore, the ability to integrate temporal local information and transform it into a long-term global perspective is crucial for embodied agents. Several works utilize topological map (Cartillier et al., 2021, 2024) to record the spatial semantics during navigation, either for obtaining a better visual representation for the environment (Hong et al., 2023b), or for constructing reasoning chains (Zhan et al., 2024b). Others employ bird's-eye-view grid maps to structure the visited environment (Chen et al., 2023a; Xiong et al., 2023; Wang et al., 2023b). For ORC tasks, a detailed 3D modeling of the environment is essential for executing precise actions with a robotic arm. For example, VoxPoser (Huang et al., 2023b) take the 3D value map derived from interactions between a LLM and a vision-language model to enable exact and efficient object manipulations.

4.3 Output Extension: Embodied Action

As stated in Section 4.1, different embodied tasks have distinct embodied action spaces, necessitating the extensions to model outputs to accommodate the specific demands of each task.

Discrete Action Space For embodied tasks of VLN and VLM with discrete action spaces, embodied actions are divided into a fixed set of categories. One line of work, i.e. LLaRP (Szot et al., 2023), utilizes an additional action prediction module to decode discrete actions. Another line of work leverages the powerful language decoding capabilities of LLMs. For example, NavGPT (Zhou et al., 2024b) and NaviLLM (Zheng et al., 2024a) predict actions as plain-text, which are then parsed into specific action commands. This design is simple and effective, yet limits the decoding of complex operations like robotic arm control in ORC tasks. To mitigate this issue, RT-2 (Brohan et al., 2023a) adds special action tokens into the vocabulary. The discrete tokens are then de-tokenized into continuous signals.

Continuous Action Space To better adapt to ORC tasks, the extension to continuous actions is necessary. Since the direct outputs of LVLMs are discrete tokens, decoding continuous actions typically requires an extra action decoding head. RoboFlamingo (Li et al., 2023d) experiments with different action decoding head architectures (e.g., MLP, RNN, and Transformer) to enable language-conditioned robotic control. Octo (Team et al., 2024b) employs a modular framework, integrating diffusion model-based action policies to predict continuous actions. Unlike RoboFlamingo, the advantage of Octo lies in its ability to flexibly connect different task encoders, observation encoders, and action decoders, making it highly adaptable.

Hierarchical Action Space This separates the level of action control into high-level task planning and low-level control policies (could be either discrete or continuous), each handled by separate modules or models. Specifically, PALM-E (Driess et al., 2023) uses high-level instructions generated by LVLMs to guide low-level control policies in executing specific embodied actions.

4.4 Multi-Modal Alignment

Input-level Alignment To bridge the gap between the newly introduced environment representation and other modalities, SMNet (Cartillier et al., 2021), GridMM (Wang et al., 2023b) and Trans4Map (Chen et al., 2023a) employ end-to-end imitation learning, continuously adjusting the model parameters to optimize the updating processes of allocentric map. However, the obtained map representations are highly dependent on the UNet and GRU modules nested within the model architecture, lacking the ability to transfer between different language backbones. To address this issue, Ego^2-Map (Hong et al., 2023b) takes a self-supervised contrastive learning strategy, comparing egocentric view features with their corresponding semantic maps. Such representations exhibit strong generalizable capability on various environments.

Output-level Alignment Adapting the outputs to different action spaces is essential for agents to understand and execute complex tasks. There are two major strategies: (1) *Direct Alignment*: This approach maps instructions directly to executable actions in an end-to-end manner, as exemplified by RoboFlamingo (Li et al., 2023d) and Octo (Team et al., 2024b). During training, both RoboFlamingo and Octo collect sequential actions covering various scenarios and tasks, enhancing the model's generalization capability during pretraining. They also allow the policy module to be fine-tuned with a small amount of trajectory data so as to quickly adapt to new tasks. Besides, LEO (Huang et al., 2023a) adopts a two-stage training process involving pre-training for 3D vision-language alignment and fine-tuning on 3D vision-language-action instructions, enhancing the agent's adaptability to different action spaces. (2) *Indirect Alignment*: This method breaks down user instructions into language plans that can be understood by downstream models, with representative works as PALM-E (Driess et al., 2023). PALM-E pre-trains on large datasets of robotic manipulation planning, visual question answering and captioning, converting complex environmental perceptions into multi-step task planning. It integrates the task plans with downstream action controller SayCan (Brohan et al., 2023b) for specific action execution.

5 Discussion and Outlook

Design of representation spaces Constructing multi-modal representation spaces involves hybrid and unified approaches (Section 2 and Section 3), each with trade-offs. Hybrid models, which encode continuous modality signals into discrete text spaces, excel in comprehension but require complex alignment modules and struggle with generation tasks. Unified discrete models simplify comprehension and generation but face challenges with weaker encoders and training stability. Addressing granularity mismatches between textual and other modality tokens is key for future improvement.

We kindly refer readers to Appendix C for more discussions in this direction.

A promising way towards world models As demonstrated in Section 4, our perspective of representation space extension works beyond modalities, encompassing any form of information or signals. By encoding these into input/output spaces and aligning them via model architecture and training strategies, models can be applied for downstream tasks. The proposed framework highlights the potential for models to understand the physical world. The statement, "*predicting the next token is to understand the world*", holds if the defined token space has been expanded to cover a sufficient amount of information and signals from the world.

6 Conclusion

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

7 Limitations

Although our analysis from the perspective of representation space extension is general, this paper does not delve deeply into the evaluation of existing LMMs. Notably, evaluation tasks and datasets can be systematically categorized based on the input and output representation spaces. For example, VQA tasks that only require Type 1 outputs (see Figure 2) could be considered “understanding" tasks, while image editing tasks that require Type 2 or Type 3 outputs could be categorised as “generation" tasks. This aspect remains an open question and we reserve it for future investigation.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

AI@Meta. 2024. Llama 3 model card.

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. *NIPS*, 35:23716–23736.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 2024. 4m-21: An any-to-any vision model for tens of tasks and modalities. *Preprint*, arXiv:2406.09406.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. 2024. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Fuyu-8b.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023a. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023b. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Vincent Cartillier, Neha Jain, and Irfan Essa. 2024. 3d semantic mapnet: Building maps for multi-object re-identification in 3d. *arXiv preprint arXiv:2403.13190*.

Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. 2021. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 964–972.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. 2020a. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. 2020b. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12875–12884.

Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. 2023a. Trans4map: Revisiting holistic bird's-eye-view mapping from egocentric images to allocentric semantics with vision transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4013–4022.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023b. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*.

Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, et al. 2024a. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023c. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023d. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*.

Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024b. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023e. Pali-x: On scaling up a multilingual vision and language model. *Preprint*, arXiv:2305.18565.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024a. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024b. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

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

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

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024a. Dreamllm: Synergistic multimodal comprehension and creation. *Preprint*, arXiv:2309.11499.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024b. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024c. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Preprint*, arXiv:2404.06512.

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. *Preprint*, arXiv:2010.11929.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, et al. 2024. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*.

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

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems*, 31.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. 2024. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*.

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

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023b. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. 2019. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv:2309.03905*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

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

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

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

Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, and Hao Tan. 2023b. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3055–3067.

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

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024a. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024b. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Wenbo Hu, Yifan Xu, Y Li, W Li, Z Chen, and Z Tu. 2023. Bliva: A simple multimodal llm for better handling of text-rich visual questions. *arXiv:2308.09936*.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023a. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023b. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*.

IDEFICS. 2023. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Yadong Mu, et al. 2023. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*.

Khen Bo Kan, Hyunsu Mun, Guohong Cao, and Youngseok Lee. 2024. Mobile-llama: Instruction fine-tuning open-source llm for network analysis in 5g networks. *IEEE Network*.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. 2024. Brave: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 104–120. Springer.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, et al. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *Preprint*, arXiv:2305.03726.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. 2024c. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv preprint arXiv:2405.05949*.

Jian Li and Weiheng Lu. 2024. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv:2301.12597*.

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

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. 2023d. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.

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. *arXiv preprint arXiv:2403.18814*.

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

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023e. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024a. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024b. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Gosh, Luke Zettlemoyer, and Armen Aghajanyan. 2024c. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022a. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. A convnet for the 2020s. arxiv e-prints.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *arXiv:2403.05525*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023a. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *Preprint*, arXiv:2312.17172.

Junyu Lu, Ruyi Gan, Dixiang Zhang, Xiaojun Wu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2023b. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *arXiv preprint arXiv:2312.05278*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024b. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023a. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv:2305.15023*.

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

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334.

MistralAITeam. 2023. Mixtral of experts a high quality sparse mixture-of-experts. [EB/OL]. https://mistral.ai/news/mixtral-of-experts/ Accessed December 11, 2023.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 2023. 4m: Massively multimodal masked modeling. *Preprint*, arXiv:2312.06647.

OpenAI. 2023. Chatgpt (august 3 version).

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

Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. 2023. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models. *Preprint*, arXiv:2310.02992.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.

Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. 2024. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, and Jiajun Ding. 2024. Imp: Highly capable large multimodal models for mobile devices. *arXiv preprint arXiv:2405.12107*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

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

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024a. Generative multimodal models are in-context learners. *Preprint*, arXiv:2312.13286.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Emu: Generative pretraining in multimodality. *Preprint*, arXiv:2307.05222.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. 2023. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*.

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

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023b. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023c. Codi-2: In-context, interleaved, and interactive any-to-any generation. *Preprint*, arXiv:2311.18775.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. 2024b. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

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

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023a. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. 2023b. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15625–15636.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. 2024a. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14205–14215.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023b. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023c. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

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

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. 2024b. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

Yonghui Wu. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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. *arXiv preprint arXiv:2405.08603*.

Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. 2023. Neural map prior for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17535–17544.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. 2023. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362*.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024a. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024b. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *ArXiv*, abs/2403.11703.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023c. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv*, abs/2311.04257.

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. 2023. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3).

Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024a. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *CoRR*, abs/2404.09204.

Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. 2024b. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. 2023. What matters in training a gpt4-style language model with multimodal inputs? *arXiv:2307.02469*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

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, Yugang Jiang, and Xipeng Qiu. 2024a. Anygpt: Unified multimodal llm with discrete sequence modeling. *Preprint*, arXiv:2402.12226.

Zhaohuan Zhan, Lisha Yu, Sijie Yu, and Guang Tan. 2024b. Mc-gpt: Empowering vision-and-language navigation with memory map and reasoning chains. *arXiv preprint arXiv:2405.10620*.

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. *arXiv preprint arXiv:2305.11000*.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechgpt-gen: Scaling chain-of-information speech generation. *arXiv preprint arXiv:2401.13527*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024b. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *Preprint*, arXiv:2407.03320.

Pan Zhang, Xiaoyi Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, Xinyu Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Y. Qiao, Da Lin, and Jiaqi Wang. 2023c. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *ArXiv*, abs/2309.15112.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023d. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023e. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024a. Cobra: Extending mamba to multi-modal large language model for efficient inference.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024b. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024a. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2024b. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *Preprint*, arXiv:2310.02239.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024a. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024b. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

A More about Visual Encoder

In Section 2.1.2, we illustrate the visual representations obtained via different visual encoders. As shown in Figure 5, the training strategies of these encoders vary. Most visual encoders are pre-trained in supervised or self-supervised manner. For supervised learning, early exploration utilize image categories as supervision signals (Dosovitskiy et al., 2021), while CLIP-like models (Radford et al., 2021; Sun et al., 2023; Zhai et al., 2023) use language supervision to learn generalized representations. Additionally, SAM (Kirillov et al., 2023) leverages segmentation tasks as training objectives. In contrast, self-supervised learning only requires images for training. One line of works employ contrastive self-supervised methods to distinguish representations between different images (He et al., 2020; Caron et al., 2021; Zhou et al., 2021; Oquab et al., 2023). Another line of approaches construct auto-encoders, where models are demanded to reconstruct images from the encoded visual representations, which is often used to support downstream image generation (Van Den Oord et al., 2017; Esser et al., 2021; Ge et al., 2023a; Sun et al., 2024a).

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

To support high-resolutional image processing, the direct method is to increase the resolution accepted by the visual encoder, including interpolating position embeddings in vision Transformers (Zhu et al., 2023a; Bai et al., 2023b) and using CNN-based models to enhance the encoding efficiency of high-resolution images while compressing the size of encoded feature maps (Yuan et al., 2024; Ge et al., 2024). Other works propose to crop high-resolution images into multiple sub-images and input them into the low-resolution encoder along with the down-sampled full image (Ye et al., 2023a; Li et al., 2023e; Gao et al., 2024; Xu et al., 2024b; Liu et al., 2024a). Different sub-image partitioning templates also help address issues caused by varying aspect ratios of images.

Regarding feature enhancement, common practices consider to ensemble visual representations encoded by different encoders, such as combining encoders trained with different strategies (Lu et al., 2024a; Zhao et al., 2024a), or integrating high-resolution and low-resolution encoders to-

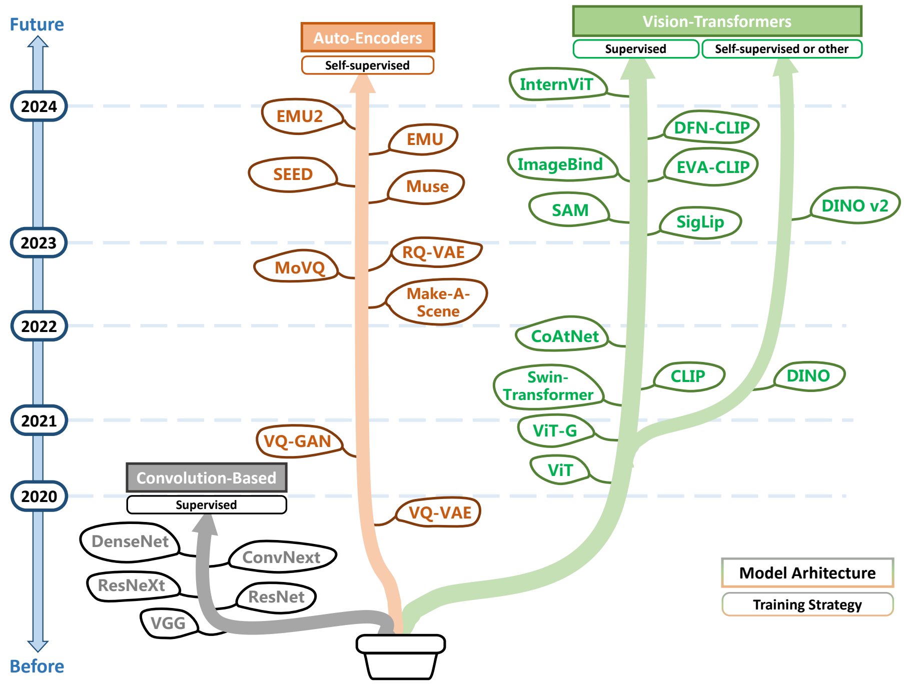

Figure 5: The evolution of commonly adopted visual encoder architectures and training strategies.

gether (Hong et al., 2023a; Li et al., 2024d). Specialized modules have been introduced to better fuse features from different encoders (Li et al., 2024d; Tong et al., 2024; Fan et al., 2024)

B Multi-Modal Alignment Training

Here, we provide additional details about modality alignment training that could not be fully discussed in the main text due to page limits. In Section 3, we have demonstrated the alignment architectures of contemporary LMMs, as summarized in Table 1 and Table 2. In this section, we will further illustrate the alignment training of LMMs.

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

Pre-training The primary goal of pre-training is to align and associate the input representations of various modalities within the multi-modal input space, enabling the backbone to uniformly model and understand inputs across modalities. Figure 6 illustrates the commonly applied settings in the pre-training phase which is described below. At this stage, commonly used data include X-text pairs ("X" means modality X) and multi-modal inter-

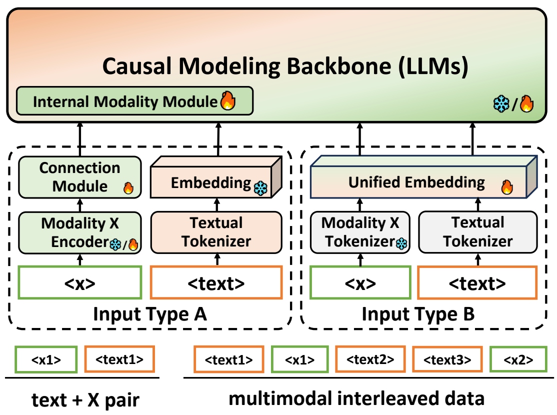

Figure 6: Illustration of common settings during the **pre-training stage**, including data and trainable parameters. "<x>" represents inputs of modalities other than text.

leaved documents. Besides the multi-modal data, text-only data can be adopted to maintain the language modeling capabilities of backbones (Zhang et al., 2023c; Lin et al., 2023; Lu et al., 2024a).

Instruction Fine-tuning The instruction fine-tuning stage enables the model to understand and follow instructions to generate appropriate responses, thereby enhancing the interactivity. Figure 7 provides a straightforward illustration for this stage. At this stage, to obtain better generalization under unseen scenarios and tasks, the training data must contains various instructions. Therefore, most LMMs adopt different strategies to construct a mixed dataset based on different re-

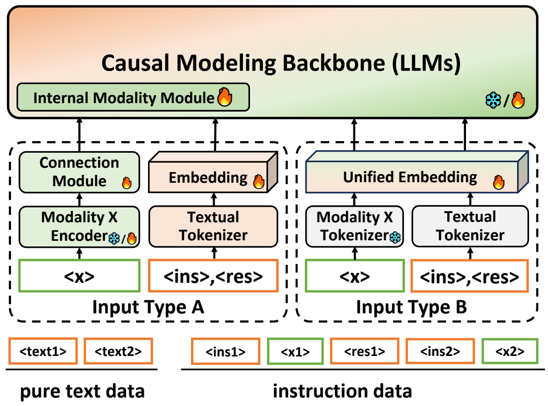

Figure 7: Illustration of common settings during the **instruction fine-tuning stage**, where <x>, <ins>, and <res> denote inputs of modalities other than text, instruction, and response, respectively.

quirements, such as mixing task-oriented data with self-instructed data (Liu et al., 2023; Laurençon et al., 2024b), combining general data with data from specific scenarios (Chen et al., 2023c; Cai et al., 2024), unifying data from various modalities (Wu et al., 2023d; Zhan et al., 2024a; Li et al., 2024b), integrating understanding and generation data (Dong et al., 2024a; Sun et al., 2024a), and blending multi-modal data and text-only data (Lin et al., 2024b; McKinzie et al., 2024).

Additional Alignment Training In addition to the regular pre-training and instruction fine-tuning stages, some specialized models require additional training stages to achieve alignment for specific objectives. To enable LMMs to generate multi-modal response, output-level alignment is required. Benefiting from unified multi-modal discrete representation and the pre-trained tokenizer and detokenizer for each modality, models with Type 3 output space can achieve output-level alignment directly through conventional pre-training and instruction fine-tuning (Jin et al., 2023; Ge et al., 2023b; Team, 2024; Zhan et al., 2024a). For models with Type 2 hybrid output space, an additional alignment stage may be required. By rearranging the order of text and other-modality information in “text + X” pairs and interleaved sequences, the text-to-other-modalities generation ability can be learned in the autoregressive setting. A line of approaches keeps modality decoders frozen and train the output mapping modules through gradients passed from the decoder for alignment (Tang et al., 2023c; Dong et al., 2024a; Zheng et al., 2024b). Since most modality decoders are originally conditioned on text for generation, the representations from the decoders’ corresponding text encoder can be utilized as supervision signal (Wu et al., 2023d; Koh et al., 2024). Another line of methods, represented by Emu series (Sun et al., 2024b,a), propose to construct an autoencoder between modality encoders and decoders. These methods first train LMMs to align the visual input and output spaces, then align the modality decoders to this space.

C Further Discussions

How to construct multi-modal representation spaces with discretely or continuously encoded modality signal? Currently, mainstream LMMs follow the hybrid structure, where modality signals are continuously encoded and integrated into the text space. This method is simple yet effective, leveraging encoders like CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2023), which are aligned with text through large-scale pre-training, to achieve impressive performance in comprehension tasks. However, this approach introduce additional design costs for corresponding alignment modules for the input and output ends.

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

On the other hand, the unified discrete space structure is more straightforward, supporting both comprehension and generation tasks through a unified approach (e.g., next-token prediction). However, they are currently limited by the absence of strong discrete encoders across various modalities, akin to CLIP, resulting in slightly weaker performance on comprehension tasks compared to hybrid models. Ovis (Lu et al., 2024b), however, has shown that by carefully designing and expanding the visual vocabulary, discrete models can also perform well on comprehension tasks. Additionally, due to the competitive relationship between modalities, improving training stability is also a challenge that needs to be addressed for unified discrete representation models.

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

Model	Input Space		Output Space		Architecture				Max Res.	Date
	Modality	Type	Modality	Type	Backbone	Modality Encoder	Connection	Internal Module		
Flamingo (2022)	Text, Vision	A	Text	1	Chinchilla	NFNet	Perceiver	Cross-Attention	480	2022/04
BLIP-2 (2023b)	Text, Vision	A	Text	1	Flan-T5 / OPT	CLIP ViT-L/14 / Eva-CLIP ViT-G/14	Q-Former	-	224	2023/01
LLaMA-adapter (2023d)	Text, Vision	A	Text	1	LLaMA	CLIP-ViT-L/14	MLP	Adaption Prompt	224	2023/03
MiniGPT-4 (2023b)	Text, Vision	A	Text	1	Vicuna	Eva-CLIP ViT-G/14	Q-Former	-	224	2023/04
LLaVA (2024b)	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Linear	-	224	2023/04
mPLUG-Owl (2023b)	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Abstractor	-	224	2023/04
LLaMA-adapter v2 (2023)	Text, Vision	A	Text	1	LLaMA	CLIP-ViT-L/14	MLP	Adaption Prompt	224	2023/04
InstructBLIP (2023)	Text, Vision	A	Text	1	Flan-T5 / Vicuna	Eva-CLIP ViT-G/14	Q-Former	-	224	2023/05
Otter (2023a)	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Perceiver	Cross-Attention	224	2023/05
LAVIN (2023a)	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	MLP	MM-Adapter	224	2023/05
MultimodalGPT (2023)	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Perceiver	Cross-Attention	224	2023/05
Shikra (2023c)	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Linear	-	224	2023/06
VideoChatGPT (2023)	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Linear	-	224	2023/06
Valley (2023b)	Text, Vision	A	Text	1	Stable-Vicuna	CLIP ViT-L/14	Temporal Module + Linear	-	224	2023/06
Lynx (2023)	Text, Vision	A	Text	1	Vicuna	EVA-1B	Resampler	Adapter	420	2023/07
Qwen-VL (2023b)	Text, Vision	A	Text	1	Qwen	OpenCLIP ViT-bigG	Cross-Attention	-	448	2023/08
BLIVA (2023)	Text, Vision	A	Text	1	Flan-T5 / Vicuna	Eva-CLIP ViT-G/14	Q-Former + MLP	-	224	2023/08
IDEFICS (2023)	Text, Vision	A	Text	1	LLaMA	OpenCLIP ViT-H/14	Perceiver	Cross-Attention	224	2023/08
OpenFlamingo (2023)	Text, Vision	A	Text	1	LLaMA, MPT	CLIP ViT-L/14	Perceiver	Cross-Attention	224	2023/08
InterLM-XC (2023c)	Text, Vision	A	Text	1	InternLM	Eva-CLIP ViT-G/14	Perceiver	-	224	2023/09
LLaVA-1.5 (2023)	Text, Vision	A	Text	1	Vicuna 1.5	CLIP ViT-L/14	MLP	-	336	2023/10
MiniGPT-v2 (2023b)	Text, Vision	A	Text	1	LLaMA-2	EVA	Linear	-	448	2023/10
Fuyu-8B (2023)	Text, Vision	A	Text	1	Persimmon	-	Linear	-	unlimited	2023/10
UReader (2023a)	Text, Vision	A	Text	1	LLaMA	CLIP ViT-L/14	Abstractor	-	224*20	2023/10
CogVLM (2023a)	Text, Vision	A	Text	1	Vicuna 1.5	EVA2-CLIP-E	MLP	Visual Expert	490	2023/11
Monkey (2023e)	Text, Vision	A	Text	1	Qwen	OpenCLIP ViT-bigG	Cross-Attention	-	896	2023/11
ShareGPT4V (2023d)	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14	MLP	-	336	2023/11
mPLUG-Owl2 (2023c)	Text, Vision	A	Text	1	LLaMA-2	CLIP ViT-L/14	Abstractor	Modality-Adaptive Module	448	2023/11
Sphinx (2023)	Text, Vision	A	Text	1	LLaMA-2	CLIP ViT-L/14 + CLIP ConvNeXt-XXL + DINOv2 ViT-G/14	Linear + Q-Former	-	672	2023/11
InternVL (2023f)	Text, Vision	A	Text	1	Vicuna	InternViT	QLLaMA / MLP	-	336	2023/12
MobileVLM (2023a)	Text, Vision	A	Text	1	MobileLLaMA	CLIP ViT-L/14	LDP (conv-based)	-	336	2023/12
VILA (2024b)	Text, Vision	A	Text	1	LLaMA-2	CLIP ViT-L	Linear	-	336	2023/12
Osprey (2024)	Text, Vision	A	Text	1	Vicuna	CLIP ConvNeXt-L	MLP	-	512	2023/12
Honeybee (2024)	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14	C-Abstractor / D-Abstractor	-	336	2023/12
Omni-SMoLA (2024a)	Text, Vision	A	Text	1	UL2	Siglip ViT-G/14	Linear	LoRA MoE	1064	2023/12
LLaVA-Next (2024a)	Text, Vision	A	Text	1	Vicuna / Mistral / Hermes-2-Yi	CLIP ViT-L/14	MLP	-	672	2024/01
InterLM-XC2 (2024b)	Text, Vision	A	Text	1	InternLM-2	CLIP ViT-L/14	MLP	Partial LoRA	490	2024/01
Mousi (2024)	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14 + MAE + LayoutLMv3 + ConvNeXt + SAM + DINOv2 ViT-G	Poly-Expert Fusion	-	1024	2024/01
LLaVA-MoLE (2024b)	Text, Vision	A	Text	1	Vicuna1.5	CLIP ViT-L/14	MLP	LoRA MoE	336	2024/01
MoE-LLaVA (2024a)	Text, Vision	A	Text	1	StableL / Qwen / Phi-2	CLIP ViT-L/14	MLP	FFN MoE	336	2024/01
MobileVLM v2 (2024a)	Text, Vision	A	Text	1	MobileLLaMA	CLIP ViT-L/14	LDP v2		336	2024/02
Bunny (2024)	Text, Vision	A	Text	1	Phi-1.5 / LLaMA-3 StableLM-2 / Phi-2	SigLIP, EVA-CLIP	MLP	-	1152	2024/02
TinyLLaVA (2024a)	Text, Vision	A	Text	1	TinyLLaMA / Phi-2 / StableLM-2	SigLIP-L, CLIP ViT-L	MLP	-	336/384	2024/02
Sphinx-X (2024)	Text, Vision	A	Text	1	TinyLLaMA / InternLM2 / LLaMA2 / Mixtral	CLIP ConvNeXt-XXL + DINOv2 ViT-G/14	Linear	-	672	2024/02
Mini-Gemini (2024d)	Text, Vision	A	Text	1	Gemma / Vicuna / Mixtral / Hermes-2-Yi	CLIP ViT-L + ConvNext-L	Cross-Attention + MLP	-	1536	2024/03
Deepseek-VL (2024a)	Text, Vision	A	Text	1	Deepseek LLM	SigLIP-L, SAM-B	MLP	-	1024	2024/03
LLaVA-UHD (2024b)	Text, Vision	A	Text	1	Vicuna	CLIP ViT-L/14	Perceiver	-	336*6	2024/03
Yi-VL (2024)	Text, Vision	A	Text	1	Yi	CLIP ViT-H/14	MLP	-	448	2024/03
MM1 (2024)	Text, Vision	A	Text	1	in-house LLM	CLIP ViT-H*	C-Abstractor	-	1792	2024/03
VL Mamba (2024)	Text, Vision	A	Text	1	Mamba LLM	CLIP-ViT-L / SigLIP-SO400M	VSS + MLP	-	384	2024/03
Cobra (2024b)	Text, Vision	A	Text	1	Mamba-Zephyr	DINOv2 + SigLIP	MLP	-	384	2024/03
InternVL 1.5 (2024c)	Text, Vision	A	Text	1	InternLM2	InternViT-6B	MLP	-	448*40	2024/04
Phi-3-Vision (2024)	Text, Vision	A	Text	1	Phi-3	CLIP ViT-L/14	MLP	-	336*16	2024/04
PLLaVA (2024a)	Text, Vision	A	Text	1	Vicuna / Mistral / Hermes-2-Yi	CLIP ViT-L/14	MLP + Adaptive Pooling	-	336	2024/04
TextHawk (2024a)	Text, Vision	A	Text	1	InternLM-1	SigLIP-SO400M/14	Resampler + MLP	-	unlimited	2024/04
Imp (2024)	Text, Vision	A	Text	1	Phi-2	SigLIP	MLP	-	384	2024/05
IDEFICS2 (2024b)	Text, Vision	A	Text	1	Mistral-v0.1	SigLIP-SO400M/14	Perceiver + MLP	-	384*4	2024/05
ConvLLaVA (2024)	Text, Vision	A	Text	1	Vicuna-	CLIP-ConvNeXt-L*	MLP	-	1536	2024/05
Ovis (2024b)	Text, Vision	B	Text	1	LLaMA3 / Qwen1.5	CLIP ViT-L + Visual Embedding	-	-	336	2024/05
Deco (2024)	Text, Vision	A	Text	1	Vicuna-1.5	CLIP ViT-L/14	MLP + Adaptive Pooling	-	336	2024/05
CuMo (2024c)	Text, Vision	A	Text	1	Mistral / Mixtral	CLIP ViT-L/14	MLP	FFN + MLP MoE	336	2024/05
Cambrian-1 (2024)	Text, Vision	A	Text	1	Vicuna-1.5 / LLaMA-3 / Hermes-2-Yi	CLIP ViT-L/14 + DINOv2 ViT-L/14 + SigLIP ViT-SO400M + OpenCLIP ConvNeXt-XXL	Spatial Vision Aggregator	-	1024	2024/06
GLM-4v (2024)	Text, Vision	A	Text	1	GLM4	EVA-CLIP-E	Conv + SwiGLU	-	1120	2024/06
InterLM-XC2.5 (2024b)	Text, Vision	A	Text	1	InternLM-2	CLIP ViT-L/14	MLP	Partial LoRA	560*24	2024/07
IDEFICS3 (2024a)	Text, Vision	A	Text	1	LLaMA 3.1	SigLIP-SO400M/14	Perceiver + MLP	-	1820	2024/08
mPLUG-Owl3 (2024)	Text, Vision	A	Text	1	Qwen2	SigLIP-SO400M/14	Linear	Hyper Attention	384*6	2024/08
CogVLM2 (2024)	Text, Vision	A	Text	1	LLaMA3	EVA-CLIP-E	Conv + SwiGLU	Visual Expert	1344	2024/08
CogVLM2-video (2024)	Text, Vision	A	Text	1	LLaMA3	EVA-CLIP-E	Conv + SwiGLU	-	224	2024/08
LLaVA-OV (2024a)	Text, Vision	A	Text	1	Qwen-2	SigLIP-SO400M/14	MLP	-	384*36	2024/09
Qwen2-VL (2024)	Text, Vision	A	Text	1	Qwen-2	ViT-675M	MLP	-	unlimited	2024/09

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

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

Table 2: Supplement to Table 1. In the "Modality" column, T, V, A and 3D are abbreviations for text, vision, audio, and 3D point cloud, respectively.

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

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

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

As introduced in Section 3, there is a variety of designs for each module, with different structures having trade-offs across various dimensions. No structure consistently performs better across different scenarios and requirements. Finding ways to quickly validate the effectiveness of an optimization direction is essential. Luckily, there have already been relevant explorations to provide some general conclusions (Laurençon et al., 2024b; McKinzie et al., 2024), offering heuristic approaches to narrow down the model design space.