UnifiedMLLM: Enabling Unified Representation for Multi-modal Multi-tasks With Large Language Model

Zhaowei Li[1,2], **Wei Wang**[1,3], **Yiqing Cai**[1], **Qi Xu**[1], **Pengyu Wang**[2],
Dong Zhang[2], **Hang Song**[1], **Botian Jiang**[2], **Zhida Huang**[1], **Tao Wang**[1]
[1]ByteDance Inc, [2]Fudan University,
[3]University of Science and Technology of China
lizhaowei126@gmail.com

Abstract

Significant advancements has recently been achieved in the field of multi-modal large language models (MLLMs), demonstrating their remarkable capabilities in understanding and reasoning across diverse tasks. However, these models are often trained for specific tasks and rely on task-specific input-output formats, limiting their applicability to a broader range of tasks. This raises a fundamental question: Can we develop a unified approach to represent and handle different multi-modal tasks to maximize the generalizability of MLLMs? In this paper, we propose UnifiedMLLM, a comprehensive model designed to represent various tasks using a unified representation. Our model exhibits strong capabilities in comprehending the implicit intent of user instructions and preforming reasoning. In addition to generating textual responses, our model also outputs task tokens and grounding tokens, serving as indicators of task types and task granularity. These outputs are subsequently routed through the task router and directed to specific expert models for task completion. To train our model, we construct a task-specific dataset and an 100k multi-task dataset encompassing complex scenarios. Employing a three-stage training strategy, we equip our model with robust reasoning and task processing capabilities while preserving its generalization capacity and knowledge reservoir. Extensive experiments showcase the impressive performance of our unified representation approach across various tasks, surpassing existing methodologies. Furthermore, our approach exhibits exceptional scalability and generality.

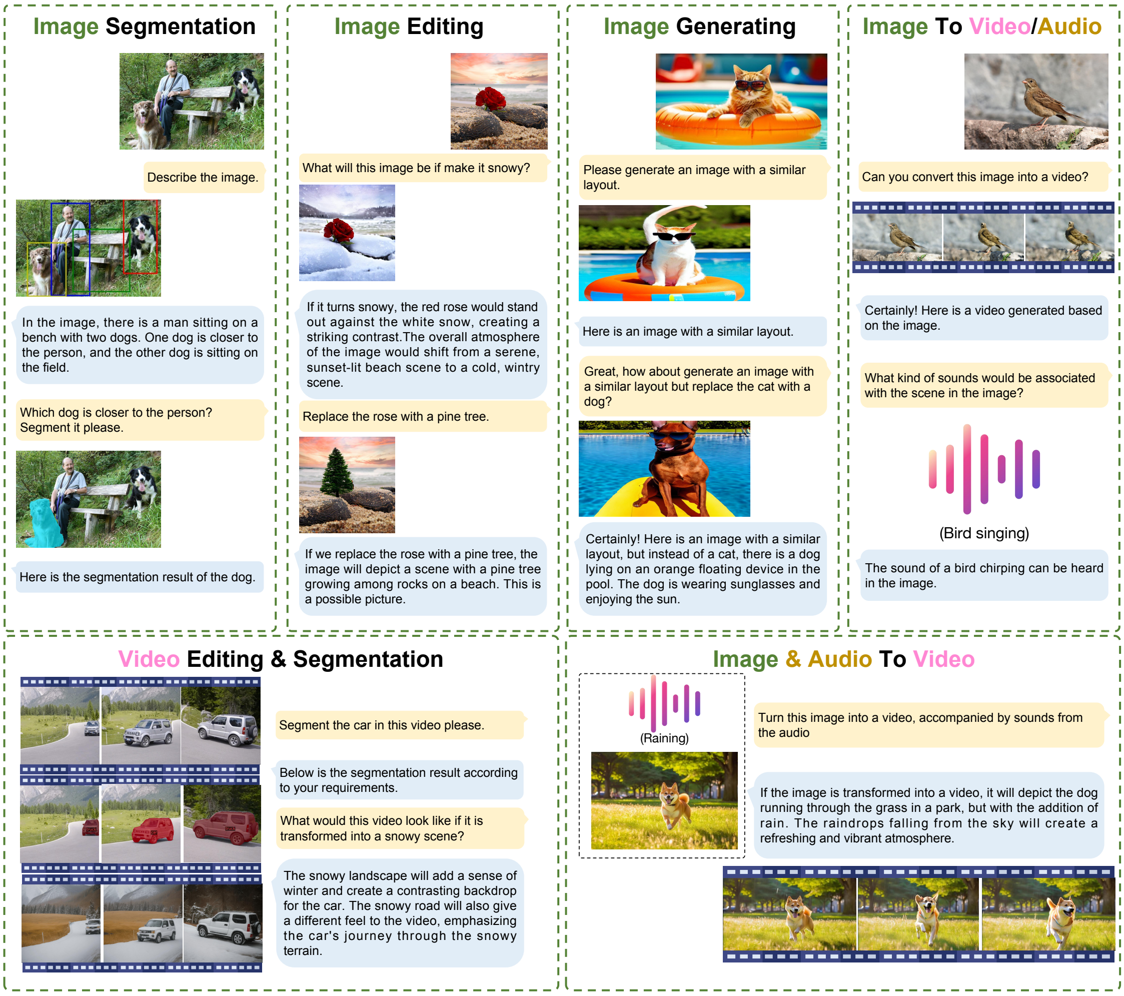
1 Introduction

Large language models have demonstrated remarkable performance in various natural language processing tasks, and the field of multi-modal large language models (MLLMs) has also made significant progress. Representative models like LLaVA (Liu et al., 2023b) and MiniGPT-4 (Zhu et al., 2023) have exhibited great capabilities in tasks such as image captioning and visual question answering. Some models have been designed to tackle a broader range of multi-modal tasks, including image segmentation (Lai et al., 2023) and image editing (Huang et al., 2023) using MLLMs. However, these models are primarily designed and trained for specific tasks, which constrains their applicability to a broader range of tasks and their overall generality in diverse scenarios due to their reliance on task-specific input-output formats. Some approaches have explored the utilization of MLLMs to accomplish more tasks. For example, LLaVA-Interactive (Chen et al., 2023) integrates multiple visual expert models with LLaVA to perform tasks such as image segmentation, editing, generation.

However, these methods view MLLMs as chatbots and heavily rely on scheduling expert models to handling visual tasks, thus failing to fully leverage the knowledge base and reasoning capabilities of MLLMs. Furthermore, while these models can handle multiple visual tasks simultaneously, they often rely on explicit instructions or predefined categories to execute visual tasks, lacking the ability to understand more implicit and complex human instructions. We expect models to comprehend implicit human intent, which encompasses understanding the tasks intend to perform and the specific regions where these tasks need to be executed. Therefore, it is necessary for models to possess strong reasoning and grounding abilities, which were lacking in previous work.

In this paper, we propose UnifiedMLLM, which models and handles different multi-modal tasks in a unified manner. Our approach introduce task tokens and grounding tokens to establish a unified representation across different tasks. The model understands the implicit intent behind user instructions and outputs not only the textual response but also our expanded special tokens indicating the task type and specific region to be processed. These

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 334–344
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics



Describe the image.

In the image, there is a man sitting on a bench with two dogs. One dog is closer to the person, and the other dog is sitting on the field.

Which dog is closer to the person? Segment it please.

Here is the segmentation result of the dog.

What will this image be if make it snowy?

If it turns snowy, the red rose would stand out against the white snow, creating a striking contrast.The overall atmosphere of the image would shift from a serene, sunset-lit beach scene to a cold, wintry scene.

Replace the rose with a pine tree.

If we replace the rose with a pine tree, the image will depict a scene with a pine tree growing among rocks on a beach. This is a possible picture.

Please generate an image with a similar layout.

Here is an image with a similar layout.

Great, how about generate an image with a similar layout but replace the cat with a dog?

Certainly! Here is an image with a similar layout, but instead of a cat, there is a dog lying on an orange floating device in the pool. The dog is wearing sunglasses and enjoying the sun.

Can you convert this image into a video?

Certainly! Here is a video generated based on the image.

What kind of sounds would be associated with the scene in the image?

The sound of a bird chirping can be heard in the image.

Segment the car in this video please.

Below is the segmentation result according to your requirements.

What would this video look like if it is transformed into a snowy scene?

The snowy landscape will add a sense of winter and create a contrasting backdrop for the car. The snowy road will also give a different feel to the video, emphasizing the car's journey through the snowy terrain.

Turn this image into a video, accompanied by sounds from the audio

If the image is transformed into a video, it will depict the dog running through the grass in a park, but with the addition of rain. The raindrops falling from the sky will create a refreshing and vibrant atmosphere.

Figure 1: We introduce UnifiedMLLM, a model that represents and handles different multi-modal tasks in a unified manner, enabling it to perform tasks involving multi-modal understanding, processing, and generation.

tokens are then routed through a task router, activating the corresponding expert model for task execution. Based on this design, as illustrated in Figure 1, our model exhibits excellent performance in accomplishing a wide range of multi-modal tasks.

To construct the datasets, we leverage publicly available datasets to create task-specific datasets. Additionally, we curate an 100k multi-task instruction instruction tuning dataset for complex scenarios using advanced grounding models (Li et al., 2024) and GPT-3.5.

During training, we adopt a three-stage training strategy. Initially, the model is trained to acquire perceptual understanding of multi-modal inputs. Subsequently, it is trained using the task-specific datasets, enabling the model to comprehend human intent, perform reasoning, and effectively accomplish a wide range of tasks. Finally, we further fine-tune the model with a multi-turn, multi-task dataset. Inspired by LoRAMOE (Dou et al., 2023), we incorporate its training methodology to ensure accurate understanding and execution of multiple tasks while mitigating knowledge forgetting and performance degradation. Experimental results across multiple multi-modal tasks demonstrate that our model effectively coordinates the MLLM and expert models, outperforming existing methods in task completion. Furthermore, our unified representation empowers our model to seamlessly integrate more tasks without the need for additional training, further demonstrating the generality and scalability of our approach. In summary, our contributions can be summarized as follows:

- We propose a unified representation for multi-task learning by introducing task tokens and grounding tokens to represent different tasks and regions. This enables us to seamlessly integrate multiple tasks.
- We construct task-specific datasets and multi-

task datasets for complex scenarios. We propose a three-stage training strategy to continuously improve the model's understanding and reasoning abilities while preserving its existing knowledge and capabilities.

- Extensive experiments conducted on various benchmarks validate the effectiveness and scalability of our unified approach.The results demonstrate the model's superior performance in handling multiple tasks and its ability to generalize across different domains.

2 Related Work

Multi-modal Large Language Models (MLLMs) In recent years, there has been significant advancements in large language models (LLMs) such as GPTs (OpenAI, 2023) and LLaMA (Touvron et al., 2023) due to their exceptional performance across various natural language processing tasks. The field of multi-modal language models (MLLMs) has also made notable progress, extending the capabilities of LLMs to handle multi-modal inputs and outputs beyond text alone. Models like LLaVA (Liu et al., 2023b) and MiniGPT-4 (Zhu et al., 2023) have demonstrated remarkable performance in visual question answering tasks. Similarly, video models like Video-LLaMA (Zhang et al., 2023c) and Video-Chatgpt (Maaz et al., 2023), as well as speech models like SpeechGPT (Zhang et al., 2023b), have also showcased their ability to comprehend the input in multiple modalities. In MoE-LLaVA (Lin et al., 2024), the exploration of incorporating the MOE (Mixture of Experts) structure into MLLMs has yielded outstanding performance while reducing the number of parameters.

Multi-tasks MLLMs Some studies (Wang et al., 2024; Zheng et al., 2025) have explored the application of MLLMs to other tasks, while other research has investigated MLLMs that can handle a greater number of modalities or tasks. Next-GPT (Wu et al., 2023) achieves multi-modal input and output by connecting modality-specific diffusion models at the output end. Trained on multi-modal and multi-granularity data, GroundingGPT (Li et al., 2024) is capable of understanding and grounding multi-modal inputs including images, videos, and audios. LLaVA-Interactive (Chen et al., 2023) integrates multiple models and enables tasks such as text-image dialogues, segmentation, generation, and editing, while also facilitating visual interactions. LLaVA-Plus (Liu et al., 2023c) incorporates a skill library comprising various pre-trained visual-language models. It dynamically combines the execution results of these models in real-time based on user's multi-modal inputs to accomplish these tasks. LLMBind (Zhu et al., 2024) integrates different tasks into an MLLM by designing specific tokens. It can handle multi-modal inputs and invoke corresponding models to accomplish various tasks. However, these methods lack uniformity in handling multiple tasks and also do not possess strong capabilities in understanding human intent and reasoning.

3 Method

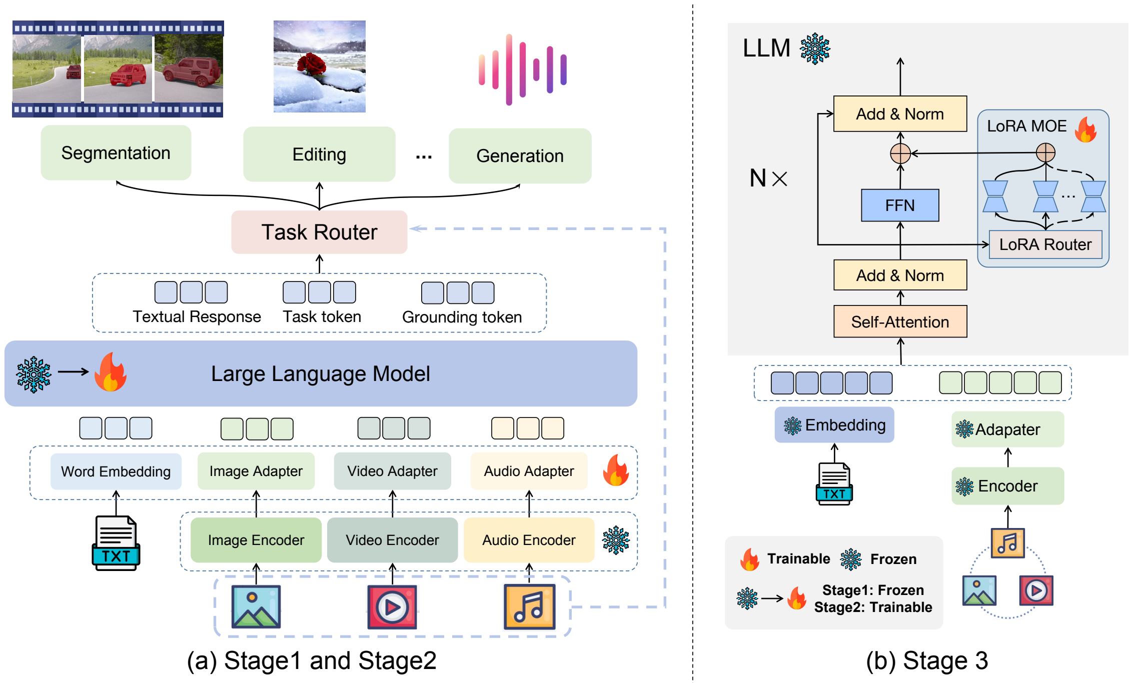
We propose UnifiedMLLM, a unified multi-modal model capable of handling various tasks in a unified manner. We will introduce the structure of UnifiedMLLM and its multi-task unified representation. Then we describe the pipeline for constructing the training dataset and our training strategy.

3.1 Architecture

Figure 2 presents the overall architecture of the model. Then we will proceed to introduce each component of the model structure.

Encoder and Adapter For each modality input, we employ different encoders to extract features, followed by modality-specific adapters. Specifically, we employ the CLIP visual encoder ViT-L/14 (Radford et al., 2021) to extract image features. For videos, we extract image features by uniformly sampling M frames. After adding temporal positional encoding to the video frames, we aggregate the video features using Q-Former, which has a structure similar to BLIP-2 (Li et al., 2023a). For audio modality, we sample N 2-second audio segments and extract features using the audio encoder from Imagebind (Girdhar et al., 2023). Similar to the video branch, a Q-Former is used to aggregate the audio features with the added temporal positional encoding. Following the feature extraction process, each modality input obtains a fixed-length embedding. Then, we use modality-specific adapters, which are two-layer MLPs, to map the features to the embedding space of LLMs.

Unified Representation Due to the different input-output formats across different tasks, achieving a unified approach to scaling and modeling



(a) Stage1 and Stage2

(b) Stage 3

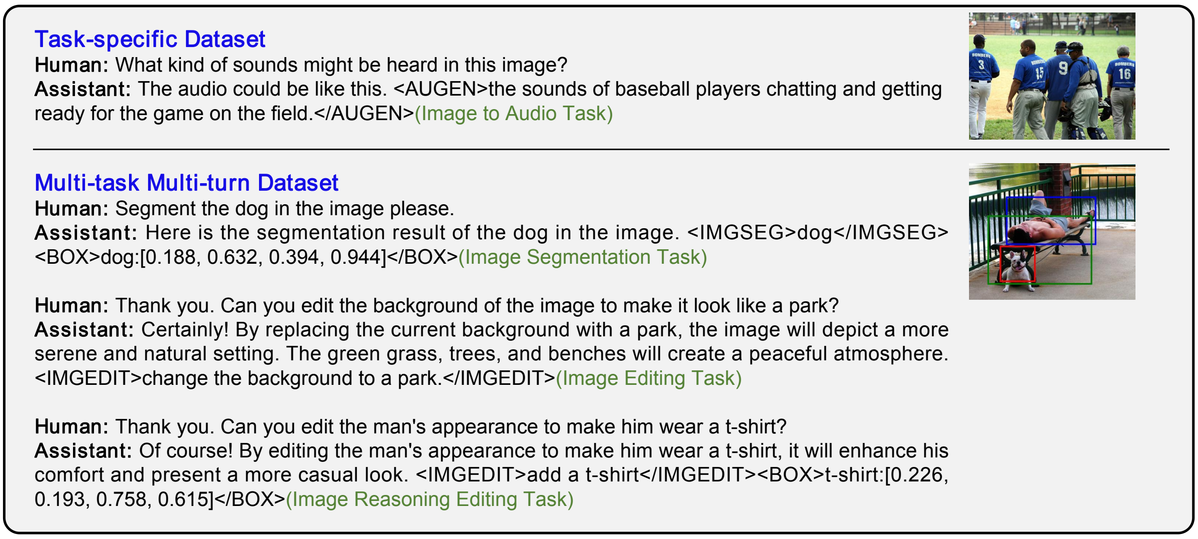
Figure 2: The model structure and three-stage training strategy of UnifiedMLLM.(a) Model structure and training strategy in the first two stages. (b) Training strategy in the third stage.

various tasks poses a significant challenge.

Moreover, for fine-grained tasks like image reasoning editing, it is essential for the model to accurately identify the specific regions to be edited based on the given instruction, as this greatly influences the successful completion of the user's requested task. As shown in Figure 2, in addition to generating textual responses, our language model generates task tokens and grounding tokens. To achieve this, we expand the vocabularies of the LLM and introduce multiple task-specific tokens and grounding tokens, which appear in pairs (e.g., <Edit></Edit>). The content between task tokens indicates the task to be executed, while the content between grounding tokens contains region-relative coordinates expressed in text format. To handle the various tasks and modalities, we employ a task router component, which utilizes the special tokens to determine the type and region of the task to be executed. The task router will activate the corresponding expert model to perform the task based on the special tokens. This representation approach facilitates seamless integration of different tasks across multiple modalities. Furthermore, decoupling the LLM from the subsequent expert models not only reduces training costs but also ensures excellent scalability.

Experts Integration We activate different external modules based on the output of the task router to execute different tasks, enabling seamless integration of various tasks. For text-to-image generation and layout-based generation tasks, we utilize Stable Diffusion (Rombach et al., 2022) and GLIGEN (Li et al., 2023b) models. Instructpix2pix (Brooks et al., 2023) and GLIGEN are employed for image global editing and reasoning editing tasks in image editing. SEEM (Zou et al., 2024) is utilized for image and video segmentation tasks. For video editing tasks, we utilize the FRESCO (Yang et al., 2024) model. ModelScopeT2V (Wang et al., 2023) and I2vgen-xl (Zhang et al., 2023d) are used for text-based video generation and image-based video generation, respectively. Additionally, Auffusion (Xue et al., 2024) is employed for audio generation.

LoRA Mixture of Experts It has been observed (Dou et al., 2023) that when LLMs introduce a large amount of instruction data during the SFT stage to enhance performance on multiple tasks, it may compromise the stored world knowledge within the model. In order to mitigate this issue and ensure that the model retains its reservoir of knowledge and reasoning abilities during the training process, we adopt a strategy where



Task-specific Dataset

Human: What kind of sounds might be heard in this image?

Assistant: The audio could be like this. <AUGEN>the sounds of baseball players chatting and getting ready for the game on the field.</AUGEN>(Image to Audio Task)

Multi-task Multi-turn Dataset

Human: Segment the dog in the image please.

Assistant: Here is the segmentation result of the dog in the image. <IMGSEG>dog</IMGSEG> <BOX>dog:[0.188, 0.632, 0.394, 0.944]</BOX>(Image Segmentation Task)

Human: Thank you. Can you edit the background of the image to make it look like a park?

Assistant: Certainly! By replacing the current background with a park, the image will depict a more serene and natural setting. The green grass, trees, and benches will create a peaceful atmosphere. <IMGEDIT>change the background to a park.</IMGEDIT>(Image Editing Task)

Human: Thank you. Can you edit the man's appearance to make him wear a t-shirt?

Assistant: Of course! By editing the man's appearance to make him wear a t-shirt, it will enhance his comfort and present a more casual look. <IMGEDIT>add a t-shirt</IMGEDIT><BOX>t-shirt:[0.226, 0.193, 0.758, 0.615]</BOX>(Image Reasoning Editing Task)

Figure 3: Examples of UnifiedMLLM Dataset in Unified Representation Format. We provide examples of task-specific datasets and multi-task multi-turn datasets.

the backbone of the model is frozen to preserve its capabilities. Additionally, multiple expert models are introduced to handle various downstream tasks. We employ LoRA as the structure for the expert models, which enhances the efficiency of both training and inference processes. For the transformers architecture, the forward propagation of the feed-forward neural (FFN) network block can be denoted as follows:

$$f(x) = x + f_{\text{FNN}}(x). \tag{1}$$

The linear layer in the FFN can be expressed as:

$$o = Wx = W_0x + \Delta W x \tag{2}$$

where $W_0 \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ represents the parameter of the backbone while $\Delta W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ denotes the updated parameter during training. We replace the linear layer with the MoE, then the forward process of the layer can be denoted as follows:

$$\begin{gathered} o = W_0x + \Delta W x = W_0x + \sum_{i=1}^{N} G(x)_i E_i(x), \\ G(x) = Softmax(x \cdot W_g), \end{gathered} \tag{3}$$

where $E_i(\cdot)$ denotes the i-th expert, $G(\cdot)$ represents the router network and the W_g is a trainable parameter of the route network. With this design, the experts are able to efficiently handle different tasks through collaboration.

To enhance the training efficiency, we replace the experts in the MoE layer with a low-rank format. The parameter matrix $\Delta W_E \in \mathbb{R}^{d_{in} \times d_{out}}$ of a single expert can be expressed as follows:

$$\triangle\, W_E = BA, \tag{4}$$

where $A \in \mathbb{R}^{d_{in} \times r}, B \in r \times \mathbb{R}^{d_{in}}$ and the rank $r << min(d_{in}, d_{out})$. The forward process of the LoRAMoE layer can be written as follows:

$$o = W_0x + \frac{\alpha}{r} \sum_{i=1}^{N} w_i B_i A_i x, \tag{5}$$

where w_i denotes the weight of i-th expert and α is a constant. This low-rank design significantly reduces training costs, improves training speed, and avoids degradation of model knowledge and capabilities during the training process.

3.2 Dataset

Task-specific Dataset To enable the model to handle different tasks in a unified manner, we construct task-specific datasets following the representation method described in section 3.1. For each task, we select task-relevant datasets and transform them into a conversation format, where the model's output includes task tokens and grounding tokens. Additionally, to further enhance the model's reasoning ability, we utilize several reasoning datasets constructed in our work. These include the reasoning segmentation dataset from LISA (Lai et al., 2023), the reasoning editing dataset from SmartEdit (Huang et al., 2023), and the layout-based image generation dataset from LayoutGPT (Feng et al., 2024). These datasets further enhance the model's understanding of human intent. Figure 3 showcases some task-specific datasets.

Multi-turn Multi-task Dataset The existing multi-task datasets are quite limited, especially those with coordinates for regions like the reasoning segmentation dataset. Due to our model's unified representation for different tasks, we can maximize the model's performance across a wider range of tasks, even with limited data availability. To further expand our dataset with grounding tokens, we utilize advanced grounding model GroundingGPT (Li et al., 2024) for data generation. Given an input image, we first use GroundingGPT to generate captions with bounding boxes. Then we utilize GPT-3.5 for multi-turn dialogue data construction. By providing GPT-3.5 a system prompt that outlines roles, requirements, and several human-annotated examples, we ask GPT-3.5 to generate multi-turn dialogues using the provided captions. Subsequently, we filter the generated data to remove samples that do not adhere to the expected output format. We totally generate 100k instances of multi-turn, multi-task dialogues, covering various multi-modal tasks in complex scenarios.

3.3 Training

We adopted a three-stage training strategy. Firstly, we train the model to acquire the ability to perceive and understand different modal inputs. Secondly, we train the model using multiple task-specific datasets to develop its capability to understand human intent and complete different tasks. Lastly, we further optimize the model's responses and enhance its reasoning ability to enable it to complete a variety of tasks in complex scenarios.

Modality-perception Pretraining In this stage, we expect the model to understand multi-modal inputs and establish the knowledge base, which serves as the foundation for subsequent reasoning and completion of various multi-modal tasks. During training, we utilize publicly available multi-modal training data, consisting of three pre-training datasets for each modality. Throughout the training process, we keep the LLM and encoder frozen and only train the adapters for each modality.

Task Adaptation Tuning After the the first stage of training, where the model gains the ability to understand inputs, it still lacks the capability to handle various multi-modal tasks. In this stage, we train the model to understand human intent and accomplish a variety of tasks. The training data used in this stage includes task-specific datasets that we constructed based on publicly available data, following the unified representation format described in section 3.1. These datasets contain replies with task tokens and grounding tokens, enabling the model to comprehend human intent. Additionally, we also use some open source general instruction fine-tuning datasets for training to improve the model's ability to understand general instructions. During this stage of training, we keep the encoders for each modality frozen while jointly training the LLM and adapters.

Multi-task LoRAMoE Tuning To enable the model to further understand human intent, perform reasoning, and accomplish a variety of tasks in complex scenarios while avoiding knowledge forgetting and performance degradation caused by further training, we utilize the constructed multi-turn multi-task dataset for training. As depicted in Figure 2, during the training process, we keep all parameters frozen except for LoRAMOE, which is updated. This training strategy enhances the model's capability to handle different tasks in complex scenarios while preserving its general ability and maintaining training efficiency.

4 Experiment

4.1 Experimental Setup

We employ Vicuna-v1.5 (Chiang et al., 2023) as the language model. Each training stage lasts for one epoch. During the training process, all images were padded to a square shape and resized to a resolution of 336 × 336. For each video, 64 frames were sampled, and for each audio, three 2-second segments were sampled and processed. All experiments were conducted on 8 A100-80G GPUs.

4.2 Quantitative Evaluation

Referring Segmentation For the reference segmentation task, the model needs to segment the objects in the image corresponding to the given expressions. We conduct experiments using the RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016) datasets and evaluate the models based on the cIoU metric.

As shown in Table 2, we have achieved excellent results on multiple datasets due to the strong grounding ability of our model.

Reasoning Editing The image reasoning edit task requires the model to reason the areas that

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
VLT (Ding et al., 2021)	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS (Wang et al., 2022)	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
LAVT (Yang et al., 2022)	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
LISA (Lai et al., 2023)	74.9	**79.1**	72.3	65.1	70.8	58.1	67.9	**70.6**
NExT-Chat (Zhang et al., 2023a)	74.7	78.9	69.5	65.1	71.9	56.7	67.0	67.0
UnifiedMLLM	**76.3**	78.8	**72.7**	**66.4**	**72.4**	**59.1**	**68.0**	69.6

Table 1: Quantitative results of image referring image segmentation on three referring segmentation datasets: RefCOCO, RefCOCO+, and RefCOCOg with metric cIoU.

Methods	Understanding Scenarios			Reasoning Scenarios		
	PSNR↑	SSIM↑	CLIP Score↑	PSNR↑	SSIM ↑	CLIP Score↑
InstructPix2Pix (Brooks et al., 2023)	21.576	0.721	22.762	24.234	0.707	19.413
MagicBrush (Zhang et al., 2024)	18.120	0.68	22.620	22.101	0.694	19.755
InstructDiffusion (Geng et al., 2024)	**23.258**	0.743	23.080	21.453	0.666	19.523
SmartEdit (Huang et al., 2023)	22.049	0.731	23.611	25.258	0.742	20.950
UnifiedMLLM	20.670	**0.776**	**23.633**	**26.667**	**0.808**	**21.104**

Table 2: Quantitative results (PSNR/SSIM/CLIP Score) of reasoning editing on Reason-Edit (Huang et al., 2023).

need editing based on user instructions and perform editing. We conducted experiments on the Reason-Edit (Huang et al., 2023) dataset. For the background, we evaluated the models using the PSNR and SSIM metrics. For the foreground regions that require editing, we calculated the CLIP Score between the edited foreground regions in the image and the ground truth labels. The results are shown in Table 2. It can be observed that our method achieves better results in both understanding and reasoning the scenes compared to other methods. Our model successfully edits the target regions while avoiding interference with the background areas.

Layout-based Image Generation The layout-based image generation task is used to evaluate the controllable generation capability of the model,where the objective was to generate coherent images by arranging the layout based on user instructions. Evaluations are conducted using the NSR-1K (Feng et al., 2024) dataset to examine the model's proficiency in comprehending quantity and spatial relationships for layout tasks. Following LayoutGPT (Feng et al., 2024), for the numerical reasoning subset, we report precision, recall, and accuracy based on generated bounding box counts and spatial positions. For spatial seasoning, we use the bounding box center for evaluation. For evaluating the generated images, we use GLIP (Li et al., 2022) to obtain bounding boxes and compute average accuracy based on the bounding box counts or spatial relations. Additionally, we also report the CLIP cosine similarity between text prompts and generated images. As shown in Table 3, our model is capable of generating layouts that are more reasonable and accurate compared to other models. Additionally, the generated images exhibit better consistency with the prompts, validating the reasoning and planning ability of our model.

Multi-modality Generation In this section, we evaluate the performance of our model in multi-modal text-based generation tasks. Specifically, in the text-to-image generation task, we evaluate using the COCO-caption (Lin et al., 2014) dataset, and the evaluation metric is the Fréchet Inception Distance (FID) score. In the text-to-video generation task, we evaluate using the MSR-VTT (Xu et al., 2016) dataset, and the evaluation metrics are FID for content quality and CLIPSIM for textual alignment. Furthermore, in the text-to-audio generation task, we conduct experiments on the Audio-Caps (Kim et al., 2019) dataset and evaluate using the Frechet Distance (FD) and Inception Score (IS) metrics. As observed from Tables 4, 5, 6, compared

Methods	Numerical Reasoning					Spatial Reasoning		
	Layout Eval.			Image Eval.		Layout Eval	Image Eval.	
	P	R	Acc	Acc-G	Sim-C	Acc	Acc-G	Sim-C
Stable Diffusion (v2.1)	-	-	-	42.44	0.256	-	17.81	0.256
Attend-and-Excite (Chefer et al., 2023)	-	-	-	45.74	0.254	-	26.86	0.264
LayoutTransformer (Gupta et al., 2021)	75.70	61.69	22.26	40.55	0.247	6.36	28.13	0.241
LayoutGPT (GPT3.5) (Feng et al., 2024)	**94.81**	**96.49**	**86.33**	51.20	0.258	82.54	52.86	0.264
LayoutGPT (GPT-4)	78.36	86.29	78.43	55.64	0.261	91.73	60.64	0.268
UnifiedMLLM	93.03	95.02	85.43	**57.94**	**0.266**	**92.93**	**61.78**	**0.270**

Table 3: Quantitative results of layout-guided image generation on NSR-1K (Feng et al., 2024), evaluated using counting and spatial correctness. "P" refers to precision. "R" refers to recall. "Acc-G" refers to accuracy calculated based on the GLIP (Li et al., 2022) model, while "Sim-C" refers to similarity calculated based on the CLIP (Radford et al., 2021) model.

Method	FID ↓
GLIDE (Nichol et al., 2021)	12.24
GILL (Koh et al., 2024)	12.20
Emu (Sun et al., 2023)	11.66
Codi (Tang et al., 2024b)	11.26
NExT-GPT (Wu et al., 2023)	11.28
UnifiedMLLM	**10.84**

Table 4: Quantitative results of text-to-image generation on COCO-captions dataset, evaluated with FID.

Method	FID↓	CLIPSIM↑
CogVideo (Hong et al., 2022)	23.59	0.2631
Make-A-Video (Singer et al., 2022)	13.17	0.3049
Latent-Shift (An et al., 2023)	15.23	0.2773
NExT-GPT (Wu et al., 2023)	13.04	0.3085
UnifiedMLLM	**11.15**	**0.3120**

Table 5: Quantitative results of text-to-video generation on MSR-VTT dataset, evaluated with FID and CLIPSIM.

Method	FD↓	IS↑
DiffSound (Yang et al., 2023)	47.68	4.01
AudioLDM-S (Liu et al., 2023a)	29.48	6.90
AudioLDM-L (Liu et al., 2023a)	23.31	8.13
NExT-GPT (Wu et al., 2023)	23.58	8.35
UnifiedMLLM	**22.42**	**9.95**

Table 6: Quantitative results of text-to-audio generation on AudioCaps dataset, evaluated with FID and IS.

to previous expert models or multi-task models, our model demonstrates strong performance across various multi-modal generation tasks.

4.3 Qualitative Results

Figure 1 presents a selection of visual results that effectively demonstrate the remarkable capabilities of our model. These results demonstrate our model's exceptional performance in tasks involving multi-modal understanding, segmentation, generation, editing and so on. As depicted in the Image Editing example, our model is able to comprehend and reason implicit human intent, enabling it to select the appropriate regions for editing. Additionally, our model exhibits robust generalization capabilities, successfully completes tasks that it has not encountered during training. For instance, tasks such as generating videos from images and audio, as depicted in the bottom right of the figure, validate the scalability of our model.

5 Conclusion

In this paper, we propose UnifiedMLLM, a multi-modal large language model that handles various multi-modal tasks using a unified representation. By introducing task tokens, grounding tokens, and a task router, we seamlessly integrate multiple tasks with excellent scalability and versatility. We construct a task-specific dataset and a multi-task multi-turn instruction-tuning dataset, and employ a three-stage training approach to enable the model to effectively perform diverse multi-modal tasks while avoiding degradation of general capabilities. Due to the powerful reasoning and grounding abilities of our model, a significant number of quantitative experiments and visual results demonstrate the effectiveness of our approach.

6 Limitation

Model Architecture Due to limited training resources and the complexity of tasks, our model primarily relies on external models to accomplish various multi-modal tasks. This approach ensures the effectiveness and scalability of completing visual tasks. However, the scope and effectiveness of the model are still constrained by the expert models. A future research direction is to construct an end-to-end trainable multi-modal system. One possible approach is to discretize various modal information, following the methodology of AnyGPT (Zhan et al., 2024).

Multi-modal Interleaving Currently, our model mainly focuses on processing single-modal inputs. Effectively handling multi-modal information simultaneously or interleaved is a challenge that needs to be addressed. CoDi-2 (Tang et al., 2024a) provides some insights, but due to the lack of this type of data, the number of tasks that can be handled is relatively limited. A future research direction is to explore how to achieve interleaved understanding and generation of inputs and outputs.

References

Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. 2023. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2021. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720.

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

Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2021. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1004–1014.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. 2023. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *arXiv preprint arXiv:2312.06739*.

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

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

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

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

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

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *CoRR*.

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

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

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

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

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

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024a. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2024b. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36.

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

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wei Wang, Zhaowei Li, Qi Xu, Linfeng Li, YiQing Cai, Botian Jiang, Hang Song, Xingcan Hu, Pengyu Wang, and Li Xiao. 2024. Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models. *arXiv preprint arXiv:2411.09691*.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695.

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

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

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. 2024. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2024. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8703–8712.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. 2023a. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

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

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2024. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. 2023d. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025. Unveiling fake news with adversarial arguments generated by multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7862–7869.

Bin Zhu, Peng Jin, Munan Ning, Bin Lin, Jinfa Huang, Qi Song, Mingjun Pan, and Li Yuan. 2024. Llmbind: A unified modality-task integration framework. *arXiv preprint arXiv:2402.14891*.

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

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36.