



Shengqiong Wu¹ Hao Fei¹ Leigang Qu¹ Wei Ji¹ Tat-Seng Chua¹

Abstract

While recently Multimodal Large Language Models (MM-LLMs) have made exciting strides, they mostly fall prey to the limitation of only input-side multimodal understanding, without the ability to produce content in multiple modalities. As we humans always perceive the world and communicate with people through various modalities, developing any-to-any MM-LLMs capable of accepting and delivering content in any modality becomes essential to human-level AI. To fill the gap, we present an end-to-end general-purpose any-to-any MM-LLM system, **NExT-GPT**. We connect an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, image, video, and audio. By leveraging the existing well-trained high-performing encoders and decoders, NExT-GPT is tuned with only a small amount of parameter (1%) of certain projection layers, which not only benefits low-cost training but also facilitates convenient expansion to more potential modalities. Moreover, we introduce a modality-switching instruction tuning (MosIT) and manually curate a high-quality dataset for MosIT, based on which NExT-GPT is empowered with complex cross-modal semantic understanding and content generation. Overall, our research showcases the promising possibility of building a unified AI agent capable of modeling universal modalities, paving the way for more human-like AI research in the community. Project website: <https://next-gpt.github.io/>

1. Introduction

Recently, the topic of Artificial Intelligence Generated Content (AIGC) has witnessed unprecedented advancements

¹NExT++ Research Center, National University of Singapore, Singapore. Correspondence to: Hao Fei <haofei37@nus.edu.sg>.

with certain technologies, such as ChatGPT for text generation (OpenAI, 2022a) and diffusion models for visual generation (Fan et al., 2022). Among these, the rise of Large Language Models (LLMs) has been particularly remarkable, e.g., Flan-T5 (Chung et al., 2022), Vicuna (Chiang et al., 2023), LLaMA (Touvron et al., 2023) and Alpaca (Taori et al., 2023), showcasing their formidable human-level language reasoning and decision-making capabilities, shining a light on the path of Artificial General Intelligence (AGI). Our world is inherently multimodal, and humans perceive the world with different sensory organs for varied modal information, such as language, images, videos, and sounds, which often complement and synergize with each other. With such intuition, the purely text-based LLMs have recently been endowed with other modal understanding and perception capabilities of image, video, audio, etc.

A notable approach involves employing adapters that align pre-trained encoders in other modalities to textual LLMs. This endeavor has led to the rapid development of multimodal LLMs (MM-LLMs), such as BLIP-2 (Li et al., 2023c), Flamingo (Alayrac et al., 2022), MiniGPT-4 (Zhu et al., 2023), Video-LLaMA (Zhang et al., 2023c), LLaVA (Liu et al., 2023b), PandaGPT (Su et al., 2023), and SpeechGPT (Zhang et al., 2023b). Nevertheless, most of these efforts pay attention to the multimodal content understanding at the input side. Lately, fewer works have considered multimodal generation, such as Emu (Sun et al., 2023), DREAMLLM (Dong et al., 2023), GILL (Koh et al., 2023), SEED (Ge et al., 2023). Notably, these models are confined to generating interleaved texts and images. We emphasize that natural human cognition and communication indispensably require seamless transitions between any modalities of information. This makes the exploration of any-to-any MM-LLMs critical, i.e., the ability to accept inputs in any modality and deliver responses in any appropriate modality.

Certain efforts have been made to mimic the human-like any-to-any modality conversion. Lately, CoDi (Tang et al., 2023) has made strides in implementing the capability of simultaneously processing and generating arbitrary combinations of modalities; however, it lacks the reasoning and decision-making prowess of LLMs as its core, and is also limited to simple paired content generation. On the other hand, some efforts, e.g., Visual-ChatGPT (Wu et al., 2023)

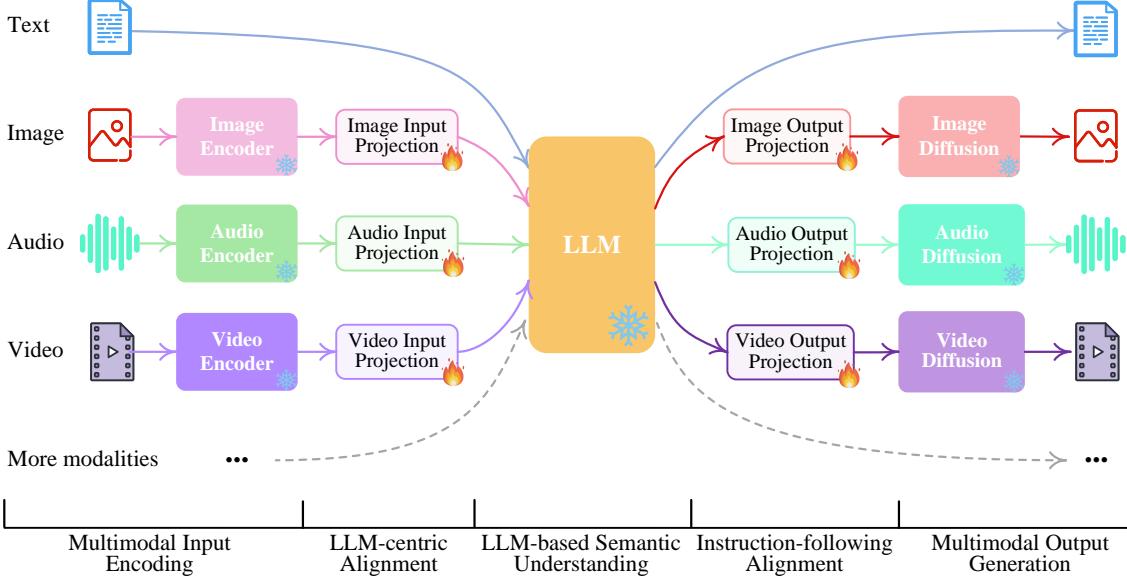


Figure 1. By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output. and represents the frozen and trainable modules, respectively.

and HuggingGPT (Shen et al., 2023), have sought to combine LLMs with external tools to achieve approximately the ‘any-to-any’ multimodal understanding and generation. Unfortunately, these systems suffer from critical challenges due to their complete pipeline architecture. First, the information transfer between different modules is entirely based on discrete texts produced by the LLM, where the cascading process inevitably introduces noise and propagates errors. More critically, the entire system leverages existing pre-trained tools for inference only. Due to the lack of overall end-to-end training, the capabilities of content understanding and multimodal generation can be very limited, especially in interpreting intricate and implicit user instructions. In a nutshell, there is a compelling need to construct an end-to-end MM-LLM of arbitrary modalities.

In pursuit of this goal, we present **NExT-GPT**, an any-to-any MM-LLM designed to seamlessly handle input and output in any combination of four modalities: text, image, video, and audio. As depicted in Figure 1, NExT-GPT comprises three tiers. **First**, we leverage established encoders to encode inputs in various modalities, where these representations are projected into language-like representations comprehensible to LLM through a projection layer. **Second**, we harness an existing open-sourced LLM as the core to process input information for semantic understanding and reasoning. The LLM not only directly generates text tokens but also produces unique ‘modality signal’ tokens that serve as instructions to dictate the decoding layers on whether and what modal content to output correspondingly. **Third**, after projection, the produced multimodal signals with specific instructions are routed to different encoders and finally

generate content in corresponding modalities.

As NExT-GPT encompasses encoding and generation of various modalities, training the system from scratch would entail substantial costs. Instead, we take advantage of the existing pre-trained high-performance encoders and decoders, such as CLIP (Radford et al., 2021), ImageBind (Girdhar et al., 2023) and the state-of-the-art latent diffusion models (Rombach et al., 2022; Ruiz et al., 2022; Cerspense, 2023; An et al., 2023; Liu et al., 2023a; Huang et al., 2023a). By loading the off-the-shelf parameters, we not only avoid cold-start training but also facilitate the potential growth of more modalities. For feature alignment across the three tiers, we only consider fine-tuning locally the input projection and output projection layers, with an encoding-side LLM-centric alignment and decoding-side instruction-following alignment, where the minimal computational overhead ensures higher efficiency. Furthermore, to empower our any-to-any MM-LLM with human-level capabilities in complex cross-modal generation and reasoning, we introduce a *modality-switching instruction tuning*, to equip the system with sophisticated cross-modal semantic understanding and content generation. To combat the absence of such cross-modal instruction tuning data in the community, we manually collect and annotate a MOSIT dataset consisting of 5,000 high-quality samples. By employing the LoRA technique (Hu et al., 2022), we fine-tune the overall NExT-GPT system on instruction tuning data, updating both input and output projection layers and certain LLM parameters.

Overall, this work showcases the promising possibility of developing a more human-like MM-LLM agent capable of modeling universal modalities. The contributions of this

research include:

- We, for the first time, present an end-to-end general-purpose any-to-any MM-LLM, named NExT-GPT, capable of semantic understanding and reasoning and generation of free input and output combinations of text, image, video, and audio.
- We introduce lightweight alignment learning techniques, the LLM-centric alignment at the encoding side, and the instruction-following alignment at the decoding side, efficiently requiring only minimal parameter adjustments (only 1% params) while maintaining highly effective semantic alignment.
- We annotate a high-quality modality-switching instruction tuning dataset covering intricate instructions across various modal combinations of text, image, video, and audio, aiding MM-LLM with human-like cross-modal content understanding and reasoning.

2. Related Work

Cross-modal Understanding and Generation Our world is replete with multimodal information, wherein we continuously engage in the intricate task of comprehending and producing cross-modal content. The AI community correspondingly emerges varied forms of cross-modal learning tasks (Zeng et al., 2023; Dessì et al., 2023; Yang et al., 2021; Ding et al., 2021; Liu et al., 2023a; Dorkenwald et al., 2021). Moreover, to generate high-quality content, a multitude of strong-performing methods have been proposed, such as Transformer (Vaswani et al., 2017; Zhang et al., 2022; Ding et al., 2021; Ge et al., 2022), GANs (Liu et al., 2020; Brock et al., 2019; Xu et al., 2018; Zhu et al., 2019), VAEs (Vahdat & Kautz, 2020; Razavi et al., 2019), Flow models (Shibata et al., 2022; Bashiri et al., 2021) and the current state-of-the-art diffusion models (Hoogeboom et al., 2021; Qu et al., 2023b; Mou et al., 2023; Feng et al., 2022; Rombach et al., 2022). In particular, the diffusion-based methods have recently delivered a remarkable performance in a plethora of cross-modal generation tasks, such as DALL-E (Ramesh et al., 2021), Stable Diffusion (Rombach et al., 2022). While all previous efforts of cross-modal learning are limited to the comprehension of multimodal inputs only, CoDi (Tang et al., 2023) lately presents groundbreaking development. Leveraging the power of diffusion models, CoDi possesses the ability to generate any combination of output modalities, including language, image, video, or audio, from any combination of input modalities in parallel. Regrettably, CoDi still falls short of achieving human-like deep reasoning of input content, because it can only deliver parallel cross-modal feeding&generation without any reasoning and decision-marking capabilities.

Multimodal Large Language Models LLMs have already made a profound impact and revolution on the entire AI community and beyond (OpenAI, 2022a;b), where a

series of open-source LLMs have greatly spurred advancement and made contributions to the community (Chieng et al., 2023; Touvron et al., 2023; Zhu et al., 2023; Zhang et al., 2023a). Building on top of these LLMs, significant efforts have been made to extend them to deal with multimodal inputs and tasks, leading to the development of MM-LLMs. On the one hand, most researchers build fundamental MM-LLMs by aligning the well-trained encoders of various modalities to the textual feature space of LLMs to perceive other modal inputs (Huang et al., 2023c; Zhu et al., 2023; Su et al., 2022; Koh et al., 2023). For example, Flamingo (Alayrac et al., 2022) uses a cross-attention layer to connect a frozen image encoder to the LLMs. BLIP-2 (Li et al., 2023c) employs a Q-Former to translate the input image queries to the LLMs. There are also various similar practices for building MM-LLMs that are able to understand video (e.g., Video-Chat (Li et al., 2023d) and Video-LLaMA (Zhang et al., 2023c)), audio (e.g., SpeechGPT (Zhang et al., 2023b)), etc. Profoundly, PandaGPT (Su et al., 2023) achieves a comprehensive understanding of six different modalities simultaneously by integrating the multimodal encoder, i.e., ImageBind (Girdhar et al., 2023).

Nevertheless, these MM-LLMs are all limited to only perceiving multimodal data, without the ability to generate content in arbitrary modalities. To enable LLMs with both multimodal input and output, some efforts explore employing LLMs as decision-makers, and utilizing existing off-the-shelf multimodal encoders and decoders as tools to execute multimodal input and output, such as Visual-ChatGPT (Wu et al., 2023), HuggingGPT (Shen et al., 2023), and AudioGPT (Huang et al., 2023b). As aforementioned, passing messages between modules with pure texts (i.e., LLM textual instruction) under the discrete pipeline scheme will inevitably introduce noises. Also, the lack of comprehensive tuning across the whole system significantly limits the efficacy of semantics understanding. Our work takes the mutual benefits of both the above two types, i.e., learning an any-to-any MM-LLM in an end-to-end manner.

3. Overall Architecture

Figure 1 presents the schematic overview of the NExT-GPT framework, consisting of three main stages: encoding, LLM understanding and reasoning, and decoding.

Multimodal Encoding Stage First, we leverage existing well-established models to encode inputs of various modalities. There are a set of alternatives of encoders for different modalities, e.g., CLIP (Radford et al., 2021), HuBERT (Hsu et al., 2021). Here we take advantage of the ImageBind (Girdhar et al., 2023), which is a unified high-performance encoder across six modalities. With ImageBind, we are spared from managing many numbers of heterogeneous modal encoders. Then, via a projection layer, different input

Table 1. Summary of NExT-GPT system configuration. Only 1% of parameters need updating during fine-tuning.

Encoder		Input Projection		LLM		Output Projection		Diffusion		
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	Vicuna (LoRA)	7B [✳] 33M [✳]	Transformer	31M [✳]	—	—
Image	—	—	Grouping	28M [✳]	—	—	Transformer	31M [✳]	SD	1.3B [✳]
Audio	ImageBind	1.2B [✳]	—	—	—	—	Transformer	32M [✳]	AudioLDM	975M [✳]
Video	—	—	—	—	—	—	Transformer	—	Zeroscope	1.8B [✳]

representations are mapped into language-like representations that are comprehensible to the LLM.

LLM Understanding and Reasoning Stage An LLM is used as the core agent of NExT-GPT. Technically, we employ the Vicuna (7B-v0) (Chiang et al., 2023), which is the open-source text-based LLM that is widely used in the existing MM-LLMs (Su et al., 2023; Zhang et al., 2023c). LLM takes as input the representations from different modalities and carries out semantic understanding and reasoning over the inputs. It outputs: 1) the textual responses directly, and 2) signal tokens of each modality that serve as instructions to dictate the decoding layers on whether to generate multimodal contents and what content to produce if yes.

Multimodal Generation Stage Receiving the multimodal signals with specific instructions from LLM (if any), the Transformer-based output projection layers map the signal token representations into the ones that are understandable to the following multimodal decoders. Technically, we employ the current off-the-shelf latent conditioned diffusion models of different modal generations, i.e., Stable Diffusion (SD-v1.5) for image synthesis (Rombach et al., 2022), Zeroscope (v2-576w) for video synthesis (Cer- spense, 2023), and AudioLDM (1-full) for audio synthesis (Liu et al., 2023a). After a projection layer, the signal representations are fed into the conditioned diffusion models for content generation. In Table 1 we summarize the overall system configurations. It is noteworthy that in the entire system, only the input and output projection layers of lower-scale parameters (compared with the overall huge capacity framework) are required to be updated during the following learning, with all the rest of the encoders and decoders frozen. This amounts to, $155M(=28+33+31+31+32)/[155M + 12.275B(=1.2+7+1.3+1.8+0.975)]$, or only **1%** of parameters need to be updated. This is also one of the key advantages of our MM-LLM.

4. Lightweight Multimodal Alignment Learning

To bridge the gap between the feature space of different modalities, and ensure fluent semantics understanding of different inputs, it is essential to perform alignment learning for NExT-GPT. Since we design the loosely-coupled system with mainly three tiers, we only need to update the two projection layers at the encoding side and decoding side.

4.1. Encoding-side LLM-centric Multimodal Alignment

Most existing MM-LLMs adopt the Transformer-architected multimodal encoders and generate patch-level grid features (e.g., for image, audio or video). They transform the multimodal features to be understandable to the core LLM by projecting them into the text feature space straightforwardly via linear layers. However, we note that the patch-based feature units might not best coincide with the intricate textual token semantics, as intuitively the language tokens always encapsulate separate concepts. This may result in suboptimal information perception (Zhong et al., 2022) in MM-LLMs. Thus, inspired by (Xu et al., 2022), we design a type of learnable *concept tokens* to hierarchically aggregate the grid-level features into semantic concept tokens via a grouping mechanism, and then the conceptual representation is fed into LLM.

To accomplish the alignment, we adopt an ‘X-to-text’ generation task trained on the ‘X-caption’ pair (‘X’ stands for image, audio, or video) data from existing corpus and benchmarks, i.e., given the representation of an ‘X’, to prompt the frozen LLM to generate the corresponding text description. Specifically, we utilize three types of ‘X-caption’ pair data, including 1) ‘Video-caption’ pair dataset: Webvid-2M (Bain et al., 2021), a large-scale dataset of short videos with textual description sourced from stock footage sites, 2) ‘Image-caption’ pair dataset: CC3M (Sharma et al., 2018), contains over 3 million images accompanied by diverse styles of natural-language descriptions, and 3) ‘Audio-caption’ pair dataset: AudioCaps (Kim et al., 2019), an extensive dataset of approximately 46k audio clips paired with human-written textual descriptions collected via crowdsourcing. Figure 2(a) illustrates the learning process.

4.2. Decoding-side Instruction-following Alignment

On the decoding end, we have integrated pre-trained conditional diffusion models from external resources. Our main purpose is to align the diffusion models with LLM’s output instructions. However, performing a full-scale alignment process between each diffusion model and the LLM would entail a significant computational burden. Alternatively, we explore a more efficient approach, decoding-side instruction-following alignment, as depicted in Figure 2(b). Specifically, instead of outputting straightforward textual instructions, we design three types of special tokens (Koh et al., 2023), i.e., ‘[IMG_i]’ ($i = 0, \dots, 4$) as image signal

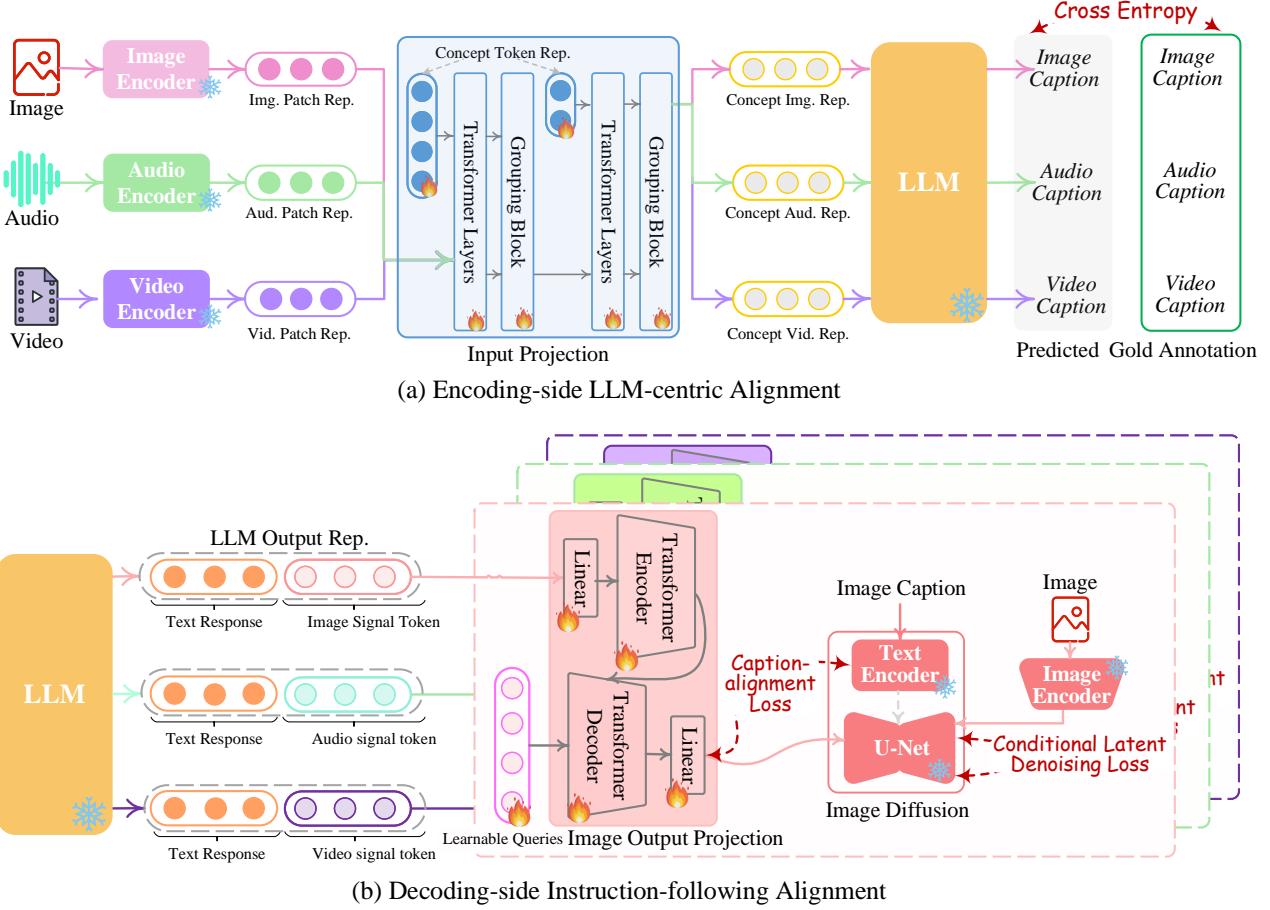


Figure 2. Illustration of the lightweight multimodal alignment learning of encoding and decoding, respectively.

tokens; ‘[AUD_i]’ ($i = 0, \dots, 8$) as audio signal tokens; and ‘[VID_i]’ ($i = 0, \dots, 24$) as video signal tokens; these tokens implicitly carry rich and flexible instructions for the downstream diffusion model. We want to enable the LLM to learn what content to generate, i.e., textual tokens, and modality signal tokens. If LLM identifies a certain modality content to be produced, a special type of token will be output indicating the activation of that modality; otherwise, no special token output means deactivation of that modality.

We notice that diffusion models generate contents conditioned solely on text-oriented representations, i.e., from the diffusion textual encoders. However, this text-centered conditioning diverges significantly from the modal signal tokens in our LLM. This leads to a gap that prevents the diffusion models from accurately interpreting the instructions from LLM. Thus, on the one hand, we consider taking the LLM’s modal signal token representations (after each Transformer-based project layer) as a conditional input in the denoising process to guide the diffusion model to generate appropriate images, videos, or audio. On the other hand, we also propose minimizing the distance between projected signal token representations and the conditional text representations of the diffusion models to accelerate alignment

learning. Note that all the diffusion backbones (i.e., U-Net) are frozen, which also ensures highly lightweight training.

In the alignment training phase, we take the captions from CC3M, WebVid, and AudioCaps as inputs and concatenate them with the signal tokens as outputs. The loss function comprises three key components: 1) the negative log-likelihood of producing signal tokens, and 2) the caption alignment loss: l_2 -distance between the hidden states of signal tokens produced by the LLM and the conditional text representations derived from the text encoder within diffusion models, and 3) conditional latent denoising loss (Rombach et al., 2022).

5. Modality-switching Instruction Tuning

5.1. Instruction Tuning

Despite aligning both the encoding and decoding ends with LLM, there remains a gap towards the goal of enabling the overall system to faithfully follow and understand users’ instructions and generate the desired multimodal outputs. To address this, further instruction tuning (IT) (Yin et al., 2023; Su et al., 2023; Liu et al., 2023b) is deemed necessary to enhance the capabilities and controllability of LLM.

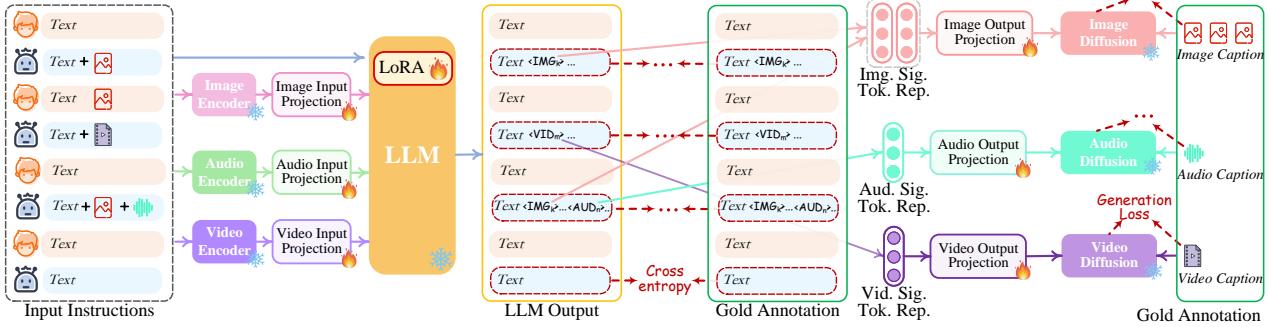


Figure 3. Illustration of modality-switching instruction tuning.

IT involves additional training of overall MM-LLMs using ‘*INPUT, OUTPUT*’ pairs, where ‘*INPUT*’ represents the user’s instruction, and ‘*OUTPUT*’ signifies the desired model output that conforms to the given instruction. Technically, we leverage LoRA (Hu et al., 2022) to enable a small subset of parameters within NExT-GPT to be updated concurrently with two layers of projection during the IT phase. As illustrated in Figure 3, when an IT dialogue sample is fed into the system, the LLM reconstructs and generates the textual content of input (and represents the multimodal content with the multimodal signal tokens). The optimization is imposed based on gold annotations and LLM’s outputs. In addition to LLM tuning, we also fine-tune the decoding end of NExT-GPT. We align the modal signal tokens’ representation encoded by the output projection with the gold multimodal caption representation encoded by the diffusion condition encoder. Thereby, the comprehensive tuning process brings closer to the goal of faithful and effective interaction with users.

5.2. Instruction Dataset

For the IT of NExT-GPT, we first consider leveraging the well-established ‘Text’→‘Text+X’ datasets where ‘X’ could be the image, video, audio, or others, for example, LLaVA-150K (Liu et al., 2023b), and VideoChat (Li et al., 2023d). However, these IT datasets are limited to output textual responses from LLMs. In our any-to-any scenario, the target not only includes the generations of texts, but also the multimodal contents, i.e., ‘Text+X’. Thus, we construct the ‘Text’→‘Text+X’ dataset, i.e., text-to-multimodal (namely T2M) data. Based on the rich volume of ‘X-caption’ pairs from the existing corpus and benchmarks (Sharma et al., 2018; Lin et al., 2014; Bain et al., 2021; Kim et al., 2019), with some templates, we employ GPT-4 to produce varied textual instructions to wrap the captions, and result in the dataset.

MosIT Dataset Crafting high-quality instructions that comprehensively cover the desired target behaviors is non-trivial. We notice that the above IT datasets fail to meet the requirements for our any-to-any MM-LLM scenario. Firstly, during a human-machine interaction, users and LLM involve diverse and dynamically changing modalities in their

inputs and outputs. Additionally, we allow multi-turn conversations in the process, and thus the processing and understanding of complex user intentions is required. However, the above two types of datasets lack variable modalities, and also are relatively short in dialogues, failing to mimic real-world scenarios adequately.

To facilitate the development of any-to-any MM-LLM, we propose a novel Modality-switching Instruction Tuning (MosIT) approach. MosIT not only supports complex cross-modal understanding and reasoning but also enables sophisticated multimodal content generation. In conjunction with MosIT, we manually and meticulously construct a high-quality dataset. The MosIT dataset encompasses a wide range of multimodal inputs and outputs, offering the necessary complexity and variability to facilitate the training of MM-LLMs that can handle diverse user interactions and deliver the desired responses accurately. Specifically, we design some template dialogue examples between a ‘Human’ role and a ‘Machine’ role, based on which we prompt the GPT-4 to generate more conversations under various scenarios with more than 100 topics or keywords. The interactions are required to be diversified, e.g., including both straightforward and implicit requirements by the ‘Human’, and execution of perception, reasoning, suggestion, and planning, etc., by the ‘Machine’. And the interactive content should be logically connected and semantically inherent and complex, with in-depth reasoning details in each response by the ‘Machine’. Each conversation should include 3-7 turns (i.e., QA pairs), where the ‘Human’ - ‘Machine’ interactions should involve multiple modalities at either the input or output side, and switch the modalities alternately. Whenever multimodal contents (e.g., image, audio, and video) are present in the conversations, we look for the best-matched contents from the external resources, including the retrieval systems, e.g., Youtube, and even AIGC tools, e.g., Stable-XL (Podell et al., 2023), and Midjourney. After human inspections and filtering of inappropriate instances, we obtain a total of 5K high-quality dialogues. In Table 6 of Appendix §C.4, we compare the statistics of existing multimodal IT datasets with our MosIT data in detailed statistics.

Table 2. Zero-shot evaluation of image captioning with CIDEr (\uparrow) score on NoCaps (Agrawal et al., 2019), Flickr 30K (Young et al., 2014) and COCO (Karpathy & Fei-Fei, 2017), and image question answering on VQA^{v2} (Goyal et al., 2017), VizWiz (Gurari et al., 2018) and OKVQA (Marino et al., 2019), and two evaluation-only benchmarks, MMB (Liu et al., 2023c) and SEED (Li et al., 2023a). The best results are marked in bold, and the second ones are underlined.

Model	Version	Image Captioning			Image Question Answering			Comprehensive	
		NoCaps	Flickr 30K	COCO	VQA ^{v2}	VizWiz	OKVQA	MMB	SEED
InstructBLIP (Dai et al., 2023)	Vicuna-7B	<u>123.1</u>	82.4	102.2	-	33.4	33.9	36.0	-
LLaVA (Liu et al., 2023b)	LLaMA-2-7B-Chat	120.7	<u>82.7</u>	-	-	-	-	36.2	-
mPLUG-Owl (Ye et al., 2023b)	LLaMA-7B	117.0	80.3	<u>119.3</u>	-	39.0	-	46.6	<u>34.0</u>
Emu (Sun et al., 2023)	LLaMA-7B	-	-	117.7	40.0	35.4	34.7	-	-
DREAMLLM (Dong et al., 2023)	Vicuna-7B	-	-	115.4	56.6	45.8	44.3	49.9	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	-	-	74.7	<u>48.1</u>	-	60.9	-
NExT-GPT	Vicuna-7B	123.7	84.5	124.9	<u>66.7</u>	48.4	52.1	<u>58.0</u>	57.5

Table 3. Comparison of video reasoning tasks on MSRVTT (Xu et al., 2016), MSVD-QA and MSRVTT-QA (Xu et al., 2017) and NExTQA (Xiao et al., 2021), and the audio captioning task on AudioCaps (Kim et al., 2019). Scores with * means being fine-tuned on the training dataset.

Model	Version	Video Captioning		Video Question Answering			Audio Captioning	
		MSR-VTT	MSVD-QA	MSRVTT-QA	NExTQA	AudioCaps		
Codi (Tang et al., 2023)	-	<u>74.4</u> *	-	-	-	-	<u>78.9</u> *	-
UIO-2XXL (Lu et al., 2023)	6.8B	48.8*	41.5	52.1	-	-	48.9*	-
Video-LLaMA (Zhang et al., 2023c)	LLaMA-7B	-	51.6	-	<u>29.6</u>	-	-	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	70.7	<u>59.2</u>	-	-	-	-
Emu (Sun et al., 2023)	LLaMA-7B	-	32.4	14.0	6.8	-	-	-
NExT-GPT	Vicuna-7B	76.2 *	<u>64.5</u>	61.4	50.7	81.3 *	-	-

Table 4. Results on text-to-image/audio/video generation (MS COCO (Lin et al., 2014), AudioCaps (Kim et al., 2019), and MSRVTT (Xu et al., 2016)). \dagger : zero-shot results.

Model	Image	Audio	Video
	FID (\downarrow)	FAD (\downarrow)	CLIPSSIM (\uparrow)
SD-1.5 (Wang et al., 2022c)	11.21	-	-
Codi (Huang et al., 2023a)	11.26	1.80	28.90
AudioLDM-L (Liu et al., 2023a)	-	1.96	-
GILL-8B \dagger (Koh et al., 2023)	12.20	-	-
Emu-13B \dagger (Sun et al., 2023)	11.66	-	-
UIO-2XXL (Lu et al., 2023)	13.39	2.64	-
NExT-GPT	10.07	1.68	31.97
NExT-GPT \dagger	<u>11.18</u>	1.74	30.96

6. Experiments

In the experiments, we aim to quantify the performance of NExT-GPT on a range of downstream tasks requiring perceiving and generating any modalities. More settings and implementation details can be found in Appendix §C. Also due to the space limitation, we present a good number of more experimental results and analyses in Appendix §D.

6.1. Main Results

Multimodal Perception Firstly, we evaluate the semantic understanding capability of the NExT-GPT w.r.t. image, video, or audio, across multiple benchmarks. The results are shown in Table 2, and 3. Notably, NExT-GPT showcases exceptional performance in image comprehension, demonstrating significant improvements over baseline levels in

tasks such as image captioning and image question answering. Moreover, when evaluated on evaluation-only benchmark datasets like MMBench (MMB) and SEED-Bench (SEED), NExT-GPT consistently achieves comparable performance. Additionally, the model excels in video and audio comprehension. In comparison with Codi, NExT-GPT attains enhanced results attributed to its capability for direct text generation from LLM, leveraging the inherent expertise of the LLM.

Multimodal Generation We then examine the synthesis quality of the image, video, or audio conditioned on text. Table 4 presents the comparisons between ours and some state-of-the-art systems. Overall, NExT-GPT exhibits superior performance in generating images, audio, and video conditioned on text. Compared to LLM-centric models, i.e., GILL, Emu, and UIO-2XXL, ours stands out by supporting a more diverse range of modalities. Moreover, in the generation of individual modalities, NExT-GPT maintains optimal performance, even in zero-shot scenarios. Notably, in comparison with non-LLM-centered models, ours still demonstrates a clear improvement in generation quality.

6.2. In-depth Analysis

The Impact of Signal Token Numbers In Figure 6, we show the results of NExT-GPT utilizing varying numbers of proposed modality-specific signal tokens. The experimental results reveal that the number of tokens required varies for each modality. Notably, videos, due to their more intricate content generation, demand the highest number of signal

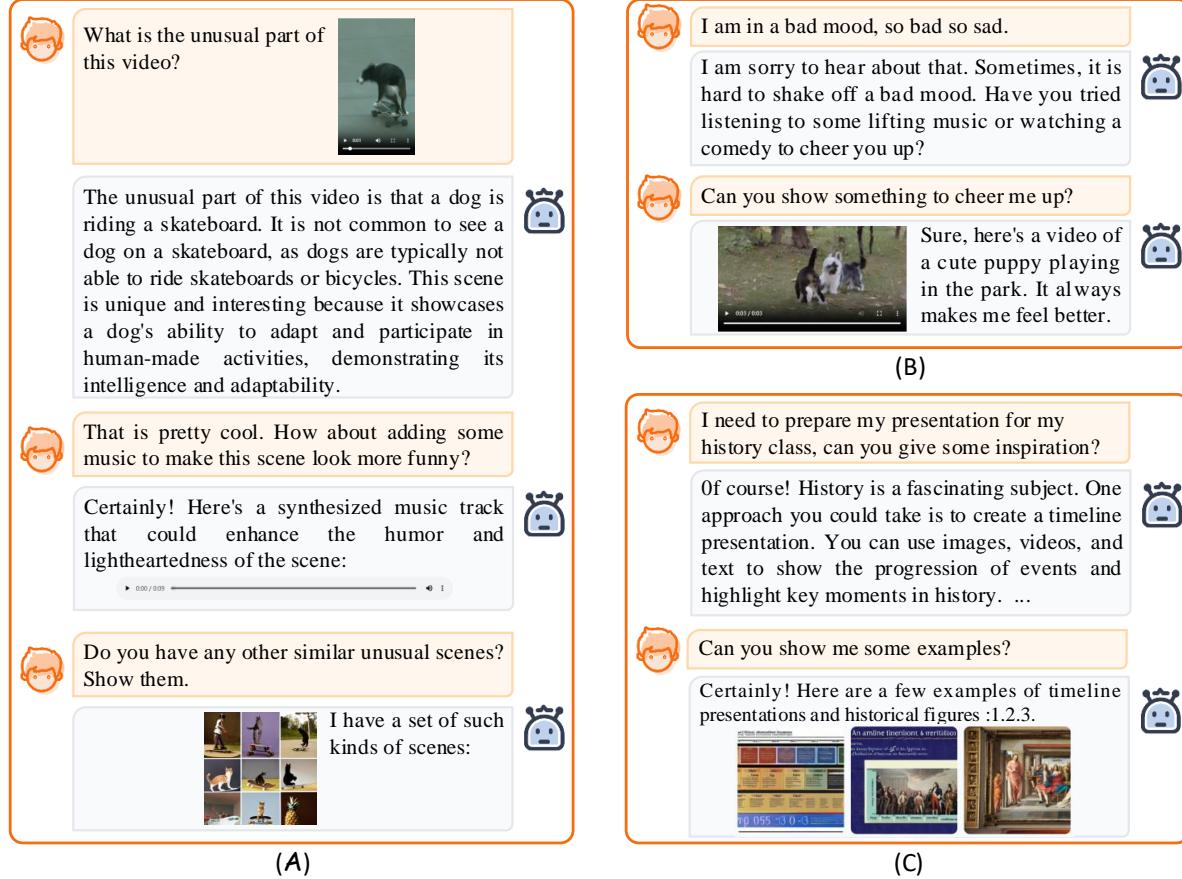


Figure 4. Qualitative examples showcasing the interpretative and generative capabilities of NExT-GPT across diverse modalities or their combinations.

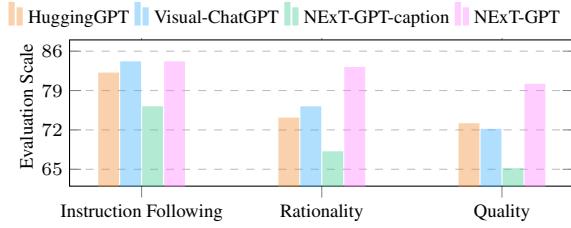


Figure 5. Human Evaluation (1-100 scale, results are on average) of NExT-GPT in comparison with pipeline baselines.

tokens. The other two modalities, images and audio, achieve satisfactory generation with merely 4 and 8 signal tokens, respectively. However, the choice of signal token numbers is contingent on factors such as training data size and the selection of the diffusion model. For example, with more extensive data and a robust diffusion model, increasing the number of signal tokens might lead to improved results.

The Impact of Grouping Mechanism To further illustrate the effectiveness of employing the grouping mechanism to align visual features with LLM, we conducted experiments with different projection architecture designs. These designs include ‘w Linea Layer’ which removes the grouping module and directly maps the output of Imagebind to the language embedding space through a single linear

layer, and ‘w Q-former + Linea Layer’ which integrates Q-former instead of the grouping mechanism. All variants undergo training following the same procedure as the original design. The results of two image QA datasets, two video QA datasets, and an audio captioning dataset are presented in Table 4. The experimental findings indicate a significant decrease in the model’s perceptual capabilities across three modalities when using a simple linear approach. In addition, the integration of Q-former yields a modest improvement in perceptual capabilities. This enhancement might be attributed to the Q-former’s ability to perform slight visual feature grouping, aligning effectively with complex textual token semantics, thus elevating perceptual capabilities. However, our grouping mechanism of NExT-GPT shows the optimal performance.

Evaluation on Pipeline vs End-to-End MM-LLMs To evaluate if the system really or how well it understands the input and generates output content (response text + image), we perform the human evaluation. For constructing the testing data, we first leverage GPT-4 to synthesize 100 complex instructions (e.g., involving intricate and semantically-rich scenes) that require implicit reasoning ability to generate image content. Then, the synthesized instructions are fed into

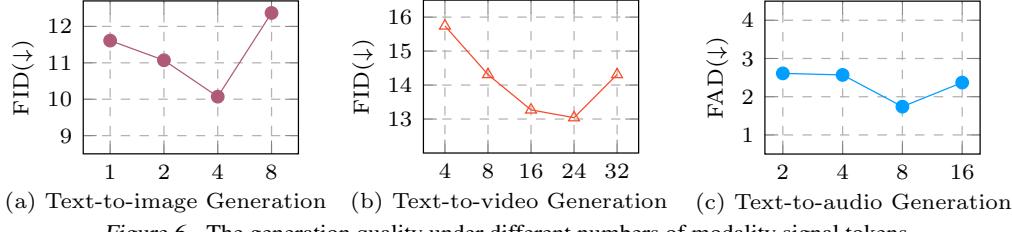


Figure 6. The generation quality under different numbers of modality signal tokens.

Table 5. The perception performance of NExT-GPT by varying input projection mechanisms.

Model	Image Question Answering		Video Question Answering		Audio Captioning
	VQA ^{v2}	VizWiz	MSVD-QA	MSRVTT-QA	AudioCaps
NExT-GPT	66.7	48.4	64.5	61.4	81.3
w Linear Layer	63.8	45.4	60.8	57.1	77.4
w Q-former + Linear Layer	65.1	46.9	63.4	58.1	79.7

the models to generate the response text + image content. Subsequently, five unbiased volunteers evaluate the generated results under three aspects, 1) **Instruction following**, identifying, among the four models, which of the generated text+image accurately responded to the input instructions, 2) **Rationality**, determining which of the generated images adhered to the input instructions, 3) **Quality**, evaluating which of the generated images exhibited the highest quality. Figure 5 illustrates superior performance in following complex instructions and generating high-quality images, compared to two existing systems and NExT-GPT-caption, which directly generates textual captions for downstream diffusion models.

Qualitative Analysis To directly demonstrate the effectiveness and potential of NExT-GPT in developing human-like conversational agents, we further offer qualitative examples that vividly illustrate the system’s capacity to comprehend and reason contents across various modalities in any combination, as shown in Figure 4. From example (A), we can note that NExT-GPT can understand the unusual part of the input video, and synthesize a light-heartedness audio and similar unusual scenes, i.e., a cat riding a skateboard. In addition, beyond responding to explicit queries prompting model synthesis in specific modalities, NExT-GPT demonstrates proficiency in inferring implicit user intentions. In example (B), when the user conveys a negative mood, NExT-GPT responds empathetically and autonomously, and decides to present a cheerful puppy video to uplift the user’s spirits. Similarly, when preparing a presentation for a history class, NExT-GPT exhibits flexibility in generating pertinent tips and visualizations. Kindly refer to Appendix §D.4 for more demonstrations with implicit and explicit instructions.

7. Conclusion

In this work, we presented an end-to-end general-purpose any-to-any multimodal Large Language Model (MM-LLM).

By connecting an LLM with multimodal adaptors and different diffusion decoders, NExT-GPT is capable of perceiving inputs and generating outputs in any combination of text, image, video, and audio. Harnessing the existing well-trained highly-performing encoders and decoders, training NExT-GPT only entails a few number of parameters (1%) of certain projection layers, which not only benefits low costs but also facilitates convenient expansion of more potential modalities in the future. To enable our NExT-GPT with complex cross-modal semantic understanding and content generation, we further introduced a modality-switching instruction tuning (MoSIT), and manually curated a high-quality dataset for MoSIT. Overall, our research showcases the potential of any-to-any MM-LLMs in bridging the gap between various modalities and paving the way for more human-like AI systems in the future.

Acknowledgements

This work is supported by CCF-Baidu Open Fund and NExT Research Center.

Impact Statement

This paper aims to develop a human-level AI agent, an end-to-end general-purpose any-to-any MM-LLM. The NExT-GPT, constrained by the quantity of fine-tuning data and the quality of base models, may produce low-quality or hallucinated content that could be harmful. Users are cautioned to interpret results carefully and adhere to licensing rules, with commercial use prohibited. We prioritize data privacy by following social media platform terms and obtaining user consent when necessary, ensuring all personal information is anonymized or obfuscated. Additionally, we are vigilant in minimizing bias in dataset collection, striving for a representative and fair dataset that does not favor or disfavor any particular group or perspective.

References

- Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., and Lee, S. nocaps: novel object captioning at scale. In *Proceedings of the ICCV*, pp. 8947–8956, 2019.
- Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J. L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In *Proceedings of the NeurIPS*, 2022.
- An, J., Zhang, S., Yang, H., Gupta, S., Huang, J., Luo, J., and Yin, X. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *CoRR*, abs/2304.08477, 2023.
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the CVPR*, pp. 6077–6086, 2018.
- Avrahami, O., Fried, O., and Lischinski, D. Blended latent diffusion. *ACM Trans. Graph.*, 42(4):149:1–149:11, 2023.
- Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the ICCV*, pp. 1708–1718, 2021.
- Bashiri, M., Walker, E. Y., Lurz, K., Jagadish, A., Muhammad, T., Ding, Z., Ding, Z., Tolias, A. S., and Sinz, F. H. A flow-based latent state generative model of neural population responses to natural images. In *Proceedings of the NeurIPS*, pp. 15801–15815, 2021.
- Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *Proceedings of the ICLR*, 2019.
- Cerspense. Zeroscope: Diffusion-based text-to-video synthesis. 2023. URL <https://huggingface.co/cerspense>.
- Ceylan, D., Huang, C. P., and Mitra, N. J. Pix2video: Video editing using image diffusion. *CoRR*, abs/2303.12688, 2023.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 902023.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V. Y., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022.
- Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. In *Proceedings of the ICLR*, 2023.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. C. H. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023.
- Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N. C., Franzon, F., and Baroni, M. Cross-domain image captioning with discriminative finetuning. In *Proceedings of the CVPR*, pp. 6935–6944, 2023.
- Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., and Tang, J. Cogview: Mastering text-to-image generation via transformers. In *Proceedings of the NeurIPS*, pp. 19822–19835, 2021.
- Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., and Yi, L. Dreamllm: Synergistic multimodal comprehension and creation. *CoRR*, abs/2309.11499, 2023.
- Dorkenwald, M., Milbich, T., Blattmann, A., Rombach, R., Derpanis, K. G., and Ommer, B. Stochastic image-to-video synthesis using cinns. In *Proceedings of the CVPR*, pp. 3742–3753, 2021.
- Fan, W., Chen, Y., Chen, D., Cheng, Y., Yuan, L., and Wang, Y. F. Frido: Feature pyramid diffusion for complex scene image synthesis. *CoRR*, abs/2208.13753, 2022.
- Feng, W., He, X., Fu, T., Jampani, V., Akula, A. R., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *CoRR*, abs/2212.05032, 2022.
- Feng, W., Zhu, W., Fu, T., Jampani, V., Akula, A. R., He, X., Basu, S., Wang, X. E., and Wang, W. Y. Layoutgpt: Compositional visual planning and generation with large language models. *CoRR*, abs/2305.15393, 2023.
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *CoRR*, abs/2208.01618, 2022.

- Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J., and Parikh, D. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *Proceedings of the ECCV*, pp. 102–118, 2022.
- Ge, Y., Ge, Y., Zeng, Z., Wang, X., and Shan, Y. Planting a SEED of vision in large language model. *CoRR*, abs/2307.08041, 2023.
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the ICASSP*, pp. 776–780, 2017.
- Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. *CoRR*, abs/2305.05665, 2023.
- Gontier, F., Serizel, R., and Cerisara, C. Automated audio captioning by fine-tuning BART with audioset tags. In *Proceedings of the DCASE*, pp. 170–174, 2021.
- Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the CVPR*, pp. 6325–6334, 2017.
- Gu, X., Chen, G., Wang, Y., Zhang, L., Luo, T., and Wen, L. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the CVPR*, pp. 18941–18951, 2023.
- Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the CVPR*, pp. 3608–3617, 2018.
- Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross-attention control. In *Proceedings of the ICLR*, 2023.
- Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *CoRR*, abs/2205.15868, 2022.
- Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., and Welling, M. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *CoRR*, 2021.
- Hsu, W., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR*, 2022.
- Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the ICML*, pp. 13916–13932, 2023a.
- Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., Ren, Y., Zhao, Z., and Watanabe, S. Audiogpt: Understanding and generating speech, music, sound, and talking head. *CoRR*, abs/2304.12995, 2023b.
- Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., and Wei, F. Language is not all you need: Aligning perception with language models. *CoRR*, abs/2302.14045, 2023c.
- Huang, W., Tu, S., and Xu, L. Pfb-diff: Progressive feature blending diffusion for text-driven image editing. *CoRR*, abs/2306.16894, 2023d.
- Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017.
- Karras, J., Holynski, A., Wang, T., and Klemelacher-Shlizerman, I. Dreampose: Fashion image-to-video synthesis via stable diffusion. *CoRR*, abs/2304.06025, 2023.
- Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the NAACL*, pp. 119–132, 2019.
- Kim, E., Kim, J., Oh, Y., Kim, K., Park, M., Sim, J., Lee, J., and Lee, K. Improving audio-language learning with mixgen and multi-level test-time augmentation. *CoRR*, abs/2210.17143, 2022.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *Proceedings of the ICLR*, 2015.
- Koh, J. Y., Fried, D., and Salakhutdinov, R. Generating images with multimodal language models. *CoRR*, abs/2305.17216, 2023.
- Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023a.

- Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425, 2023b.
- Li, J., Li, D., Savarese, S., and Hoi, S. C. H. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, pp. 19730–19742, 2023c.
- Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023d.
- Li, L., Yin, Y., Li, S., Chen, L., Wang, P., Ren, S., Li, M., Yang, Y., Xu, J., Sun, X., Kong, L., and Liu, Q. M³it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387, 2023e.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the ECCV*, pp. 121–137, 2020.
- Li, Y., Wang, X., Xiao, J., Ji, W., and Chua, T. Invariant grounding for video question answering. In *Proceedings of the CVPR*, pp. 2918–2927, 2022.
- Li, Y., Zhang, C., Yu, G., Wang, Z., Fu, B., Lin, G., Shen, C., Chen, L., and Wei, Y. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *CoRR*, abs/2308.10253, 2023f.
- Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *CoRR*, abs/2311.10122, 2023.
- Lin, K., Li, L., Lin, C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., and Wang, L. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the CVPR*, pp. 17928–17937, 2022.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Proceedings of the ECCV*, pp. 740–755, 2014.
- Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D. P., Wang, W., and Plumbley, M. D. Audioldm: Text-to-audio generation with latent diffusion models. In *Proceedings of the ICML*, pp. 21450–21474, 2023a.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023b.
- Liu, S., Wang, T., Bau, D., Zhu, J., and Torralba, A. Diverse image generation via self-conditioned gans. In *Proceedings of the CVPR*, pp. 14274–14283, 2020.
- Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281, 2023c.
- Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *CoRR*, abs/2312.17172, 2023.
- Maaz, M., Rasheed, H. A., Khan, S. H., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424, 2023.
- Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the CVPR*, pp. 3195–3204, 2019.
- Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *Proceedings of the ICLR*, 2022.
- Milewski, V. S. J., Moens, M., and Calixto, I. Are scene graphs good enough to improve image captioning? In *Proceedings of the ACL*, pp. 504–515, 2020.
- Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.
- Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the ICML*, pp. 16784–16804, 2022.
- OpenAI. Introducing chatgpt. 2022a.
- OpenAI. Gpt-4 technical report. 2022b.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the NeurIPS*, 2022.
- Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L. V., Gross, M. H., and Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the CVPR*, pp. 724–732, 2016.

- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023.
- Qu, L., Wu, S., Fei, H., Nie, L., and Chua, T. Layoutllm-t2i: Eliciting layout guidance from LLM for text-to-image generation. In *Proceedings of the ACM MM*, pp. 643–654, 2023a.
- Qu, L., Wu, S., Fei, H., Nie, L., and Chua, T. Layoutllm-t2i: Eliciting layout guidance from LLM for text-to-image generation. *CoRR*, abs/2308.05095, 2023b.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the ICML*, pp. 8748–8763, 2021.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *Proceedings of the ICML*, pp. 8821–8831, 2021.
- Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with VQ-VAE-2. In *Proceedings of the NeurIPS*, pp. 14837–14847, 2019.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the CVPR*, pp. 10674–10685, 2022.
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CoRR*, abs/2208.12242, 2022.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the ACL*, pp. 2556–2565, 2018.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.
- Shibata, H., Hanaoka, S., Cao, Y., Yoshikawa, M., Takenaga, T., Nomura, Y., Hayashi, N., and Abe, O. Local differential privacy image generation using flow-based deep generative models. *CoRR*, abs/2212.10688, 2022.
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-a-video: Text-to-video generation without text-video data. *CoRR*, abs/2209.14792, 2022.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In *Proceedings of the NeurIPS*, 2020.
- Su, Y., Lan, T., Liu, Y., Liu, F., Yogatama, D., Wang, Y., Kong, L., and Collier, N. Language models can see: Plugging visual controls in text generation. *CoRR*, abs/2205.02655, 2022.
- Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.
- Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Generative pretraining in multimodality. *CoRR*, abs/2307.05222, 2023.
- Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. Any-to-any generation via composable diffusion. *CoRR*, abs/2305.11846, 2023.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. 2023. URL https://github.com/tatsu-lab/stanford_alpaca.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lamplie, G. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.
- Vahdat, A. and Kautz, J. NVAE: A deep hierarchical variational autoencoder. In *Proceedings of the NeurIPS*, 2020.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the NeurIPS*, pp. 5998–6008, 2017.
- Veaux, C., Yamagishi, J., MacDonald, K., et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *CSTR*, 6:15, 2017.
- Voynov, A., Chu, Q., Cohen-Or, D., and Aberman, K. P+: extended textual conditioning in text-to-image generation. *CoRR*, abs/2303.09522, 2023.
- Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. GIT: A generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.*, 2022, 2022a.

- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the ICML*, volume 162, 2022b.
- Wang, T., Yi, J., Fu, R., Tao, J., and Wen, Z. Campnet: Context-aware mask prediction for end-to-end text-based speech editing. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:2241–2254, 2022c.
- Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.
- Wu, J. Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *CoRR*, abs/2212.11565, 2022.
- Xiao, J., Shang, X., Yao, A., and Chua, T. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the CVPR*, pp. 9777–9786, 2021.
- Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., and Chua, T. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI*, pp. 2804–2812, 2022.
- Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM MM*, pp. 1645–1653, 2017.
- Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., and Zhou, J. mplug-2: A modularized multimodal foundation model across text, image and video. In *Proceedings of the ICML*, pp. 38728–38748, 2023.
- Xu, J., Mei, T., Yao, T., and Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*, pp. 5288–5296, 2016.
- Xu, J., Mello, S. D., Liu, S., Byeon, W., Breuel, T. M., Kautz, J., and Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the CVPR*, pp. 18113–18123, 2022.
- Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the CVPR*, pp. 1316–1324, 2018.
- Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the ICCV*, pp. 1666–1677, 2021.
- Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:1720–1733, 2023.
- Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., Qi, Q., Zhang, J., and Huang, F. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499, 2023a.
- Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., and Huang, F. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023b.
- Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Sheng, L., Bai, L., Huang, X., Wang, Z., Shao, J., and Ouyang, W. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687, 2023.
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.
- Yu, B., Fu, C., Yu, H., Huang, F., and Li, Y. Unified language representation for question answering over text, tables, and images. *CoRR*, abs/2306.16762, 2023.
- Zeng, Z., Zhang, H., Lu, R., Wang, D., Chen, B., and Wang, Z. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the CVPR*, pp. 23465–23476, 2023.
- Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., and Chua, T. Transfer visual prompt generator across llms. *CoRR*, abs/2305.01278, 2023a.
- Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., and Guo, B. Styleswin: Transformer-based GAN for high-resolution image generation. In *Proceedings of the CVPR*, pp. 11294–11304, 2022.
- Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *CoRR*, abs/2305.11000, 2023b.
- Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858, 2023c.
- Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*, abs/2306.17107, 2023d.

Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., and Zha, Z. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the CVPR*, pp. 13275–13285, 2020.

Zhao, B., Wu, B., and Huang, T. SVIT: scaling up visual instruction tuning. *CoRR*, abs/2307.04087, 2023a.

Zhao, L., Yu, E., Ge, Z., Yang, J., Wei, H., Zhou, H., Sun, J., Peng, Y., Dong, R., Han, C., and Zhang, X. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *CoRR*, abs/2307.09474, 2023b.

Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., and Kang, B. Bubogpt: Enabling visual grounding in multi-modal llms. *CoRR*, abs/2307.08581, 2023c.

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., and Gao, J. Regionclip: Region-based language-image pretraining. In *Proceedings of the CVPR*, pp. 16772–16782, 2022.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.

Zhu, M., Pan, P., Chen, W., and Yang, Y. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the CVPR*, pp. 5802–5810, 2019.

A. Potential Limitation and Future work

As future work, there are at least following four avenues to explore.

- **i) Modalities & Tasks Expansion:** Due to resource limitations, currently, our system supports input and output in four modalities: language, images, videos, and audio. Next, we plan to extend this to accommodate even more modalities (e.g., web page, 3D vision, heat map, tables&figures) and tasks (e.g., object detection, segmentation, grounding, and tracking), broadening the system’s applicability to become more universal.
- **ii) LLM Variants:** Currently, we have implemented the 7B Vicuna version of the LLM. Our next plans involve incorporating various LLM types and sizes, allowing practitioners to choose the most suitable one for their specific requirements.
- **iii) Multimodal Generation Strategies:** While our system excels in generating content across modalities, the quality of generative outputs can sometimes be limited by the capabilities of the diffusion model. It is very promising to explore the integration of retrieval-based approaches to complement the generative process, potentially improving the overall system’s performance.
- **iv) MosIT Dataset Expansion:** Currently, our IT dataset has room for expansion. We intend to significantly increase the amount of annotated data, ensuring a more comprehensive and diverse set of instructions to further enhance the MM-LLMs’ ability to understand and follow user prompts effectively.

B. Full Related Work

Cross-modal Understanding and Generation Our world is replete with multimodal information, wherein we continuously engage in the intricate task of comprehending and producing cross-modal content. The AI community correspondingly emerges varied forms of cross-modal learning tasks, such as Image/Video Captioning (Zeng et al., 2023; Dessì et al., 2023; Milewski et al., 2020; Gu et al., 2023; Lin et al., 2022), Image/Video Question Answering (Yang et al., 2021; Xiao et al., 2022; Li et al., 2022; Yu et al., 2023; Anderson et al., 2018), Text-to-Image/Video/Speech Synthesis (Singer et al., 2022; Hong et al., 2022; Voynov et al., 2023; Gal et al., 2022; Ding et al., 2021; Liu et al., 2023a; Huang et al., 2023a), Image-to-Video Synthesis (Dorkenwald et al., 2021; Karras et al., 2023) and more, all of which have experienced rapid advancements in past decades. Researchers have proposed highly effective multimodal encoders, with the aim of constructing unified representations encompassing various modalities. Meanwhile, owing to the distinct feature spaces of different modalities, it is essential to undertake modality alignment learning. Moreover, to generate high-quality content, a multitude of strong-performing methods have been proposed, such as Transformer (Vaswani et al., 2017; Zhang et al., 2022; Ding et al., 2021; Ge et al., 2022), GANs (Liu et al., 2020; Brock et al., 2019; Xu et al., 2018; Zhu et al., 2019), VAEs (Vahdat & Kautz, 2020; Razavi et al., 2019), Flow models (Shibata et al., 2022; Bashiri et al., 2021) and the current state-of-the-art diffusion models (Hoogeboom et al., 2021; Qu et al., 2023b; Mou et al., 2023; Feng et al., 2022; Rombach et al., 2022). Especially, the diffusion-based methods have recently delivered a remarkable performance in a plethora of cross-modal generation tasks, such as DALL-E (Ramesh et al., 2021), Stable Diffusion (Rombach et al., 2022). While all previous efforts of cross-modal learning are limited to the comprehension of multimodal inputs only, CoDi (Tang et al., 2023) lately presents groundbreaking development. Leveraging the power of diffusion models, CoDi possesses the ability to generate any combination of output modalities, including language, images, videos, or audio, from any combination of input modalities in parallel. Regrettably, CoDi might still fall short of achieving human-like deep reasoning of input content, with only parallel cross-modal feeding&generation.

Multimodal Large Language Models LLMs have already made profound impacts and revolutions on the entire AI community and beyond. The most notable LLMs, i.e., OpenAI’s ChatGPT (OpenAI, 2022a) and GPT4 (OpenAI, 2022b), with alignment techniques such as instruction tuning (Ouyang et al., 2022; Li et al., 2023f; Zhang et al., 2023d; Liu et al., 2023b) and reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020), have demonstrated remarkable language understanding and reasoning abilities. And a series of open-source LLMs, e.g., Flan-T5 (Chung et al., 2022), Vicuna (Chiang et al., 2023), LLaMA (Touvron et al., 2023) and Alpaca (Taori et al., 2023), have greatly spurred advancement and made contributions to the community (Zhu et al., 2023; Zhang et al., 2023a). Afterward, significant efforts have been made to construct LLMs dealing with multimodal inputs and tasks, leading to the development of MM-LLMs.

On the one hand, most of the researchers build fundamental MM-LLMs by aligning the well-trained encoders of various modalities to the textual feature space of LLMs, so as to let LLMs perceive other modal inputs (Huang et al., 2023c; Zhu et al., 2023; Su et al., 2022; Koh et al., 2023). For example, Flamingo (Alayrac et al., 2022) uses a cross-attention layer to connect a frozen image encoder to the LLMs. BLIP-2 (Li et al., 2023c) employs a Q-Former to translate the input image

queries to the LLMs. LLaVA (Liu et al., 2023b) employs a simple projection scheme to connect image features into the word embedding space. There are also various similar practices for building MM-LLMs that are able to understand videos (e.g., Video-Chat (Li et al., 2023d) and Video-LLaMA (Zhang et al., 2023c)), audios (e.g., SpeechGPT (Zhang et al., 2023b)), etc. Profoundly, PandaGPT (Su et al., 2023) achieves a comprehensive understanding of six different modalities simultaneously by integrating the multimodal encoder, i.e., ImageBind (Girdhar et al., 2023).

Nevertheless, these MM-LLMs all are subject to the limitation of only perceiving multimodal data, without generating content in arbitrary modalities. To achieve LLMs with both multimodal input and output, some thus explore employing LLMs as decision-makers, and utilizing existing off-the-shelf multimodal encoders and decoders as tools to execute multimodal input and output, such as Visual-ChatGPT (Wu et al., 2023), HuggingGPT (Shen et al., 2023), and AudioGPT (Huang et al., 2023b). As aforementioned, passing messages between modules with pure texts (i.e., LLM textual instruction) under the discrete pipeline scheme will inevitably introduce noises. Also lacking comprehensive tuning across the whole system significantly limits the efficacy of semantics understanding. Our work takes the mutual benefits of both the above two types, i.e., learning an any-to-any MM-LLM in an end-to-end manner.

C. Implementation Details

C.1. Detailed Input Projection Layer

Through multimodal encoder, we can obtain patch-level multimodal tokens, denoting as $\mathbf{X}^* = \{\mathbf{x}_i^*\}_{i=1}^{N^*}$, where $* \in \{i, a, v\}$ represents image, audio, and video, respectively. For brevity, we eschew modal-specific notation. Differing from the existing works that directly embed multimodal tokens into LLMs by a linear projection layer, we propose a multi-stage grouping mechanism, where patch-level tokens are grouped into concept-level tokens to facilitate the subsequent cross-modal interaction. Formally, we apply L grouping stages, and in each stage, we randomly initialize M_l learnable concept tokens $\mathbf{C}^l = \{c_j\}_{j=1}^{M_l}$. Then, we concatenate input features \mathbf{X}^l and \mathbf{C}^l together and then input them into a transformer layers: $\hat{\mathbf{C}}^l, \hat{\mathbf{X}}^l = \text{Transformer}([\mathbf{C}^l; \mathbf{X}^l])$, where $\mathbf{X}^1 = \mathbf{X}$, and $[;]$ denotes the concatenation operator. Within l grouping block, we group the updated M_l concept tokens $\hat{\mathbf{X}}^l$ into M_{l+1} new concept tokens $\hat{\mathbf{X}}^{l+1}$ based on the feature similarity.

Specifically, we firstly compute a similarity matrix \mathbf{A}^l between $\hat{\mathbf{C}}^l$ and $\hat{\mathbf{X}}^l$ via a Gumbel-softmax: $\mathbf{A}^l = \text{Softmax}((\text{Norm}(\hat{\mathbf{C}}^l)\text{Norm}(\hat{\mathbf{X}}^l) + G)/\tau)$, where G are i.i.d random samples drawn from the Gumbel(0, 1) distribution, and τ is the learnable significance coefficient to assist to find a more suitable assign boundary. We compute the group to assign a concept token by taking the one-hot operation on the argmax over all the groups. Since the one-hot assignment operation via argmax is not differentiable, we instead use the straight-through trick to compute the assignment matrix as $\hat{\mathbf{A}}^l = \text{Onehot}(\text{Argmax}(\mathbf{A}^l)) + \mathbf{A}^l - \text{Sg}(\mathbf{A}^l)$, where $\text{Sg}(\cdot)$ is the stop gradient operator. Finally, we integrate the features to updated concept tokens: $\mathbf{X}^{l+1} = \hat{\mathbf{C}}^l + \text{MLP}(\hat{\mathbf{A}}^l, \hat{\mathbf{X}}^l)$. After L stages grouping, we can obtain M_L concept tokens \mathbf{X}^L , which are then fed into the LLM for perception and reasoning.

C.2. Model Training

For NExT-GPT model training, we consider a three-stage learning process:

- **Stage-1: Encoding-side Alignment Learning.** As discussed in Section §4.1, we bridge the alignment to perform the caption generation task. The cross-entropy is employed as the loss function. During training, we only keep the input projection layer trainable while the other part of NExT-GPT is frozen. We employ Adam (Kingma & Ba, 2015) optimizer to update the parameters. This stage can be understood as training a compatible multimodal tokenizer for the frozen LLM.
- **Stage-2: Decoding-side Alignment Learning.** The output projection layer adopts a transformer-based architecture characterized by a hidden size of 512, 4 attention heads, 4 encoder layers, and 4 decoder layers. Additionally, the dropout ratio is set as 0.1. The optimization process for the three output projection layers involves a combination of three training objectives: cross-entropy focusing on the generated signal tokens, l_2 -distance measuring the alignment between the representation of signal tokens and captions, and conditional latent denoising loss, as shown in Section §4.2. We employ the Adam optimizer for this stage, with only the parameters of the output projection layers being learnable, while others remain frozen.
- **Stage-3: End-to-end Instruction-Tuning.** In this stage, we train the whole NExT-GPT using instruction-following datasets, as enumerated in Section §5.2. We incorporate LoRA to fine-tune the weights of the LLM. Moreover, both the input and output projection layers are trainable. The training objectives include two parts: 1) cross-entropy between the

generated and gold response, 2) generation loss. The Adam optimizer is applied to update the learnable parameters.

C.3. Detailed Dataset

Here, we enumerate the datasets employed for training and fine-tuning NExT-GPT:

- ‘Text-X’ Pair Dataset.
 - CC3M (Sharma et al., 2018): contains over 3 million images accompanied by diverse styles of natural-language descriptions.
 - COCO-caption (Lin et al., 2014): is a large-scale image-text pair dataset which is taken as image captioning, or text-to-image generation task benchmark.
 - WebVid-2M (Bain et al., 2021): is a large-scale dataset of short videos with textual description sourced from stock footage sites.
 - AudioCaps (Kim et al., 2019): with 46K audio-text pairs derived from the AudioSet (Gemmeke et al., 2017) dataset.
- ‘Text’ → ‘Text’ Instruction Datasets.
 - Cleaned-Alpaca¹: is a textual instruction dataset used to train the Alpaca LLM (Large Language Model).
- ‘Text+X’ → ‘Text’ Instruction Datasets.
 - LLaVA-150K (Liu et al., 2023b): is a set of GPT-generated multimodal instruction-following data. It is constructed for visual instruction tuning and for building large multimodal towards GPT-4 vision/language capability.
 - VideoChat (Li et al., 2023d): comprising thousands of videos paired with detailed dataset textual descriptions and conversations generated using dense captions fed to ChatGPT in temporal order.
- Modality-switching Dataset
 - MosIT: encompasses a wide range topic of dialogues between ‘Human’ and ‘Machine’. Each dialogue includes 3-7 turns (i.e., QA pairs), where the ‘Human’-‘Machine’ interactions should involve multiple modalities at either the input or output side, and switch the modalities alternately.

C.4. Multimodal IT Datasets Comparison

Here, we compare the existing multimodal instruction tuning (IT) datasets, as detailed in Table 6. As can be seen, the response modality of the existing IT datasets is merely limited to text. In this work, we leverage GPT-4 to generate a T2M IT dataset, comprising 15k instances, which serves as a foundation for instructing the model to generate responses in other modalities, such as image, video, and audio. Furthermore, we construct a modality-switching IT dataset with 5k instances, named MosIT. This dataset is designed to emulate the human-machine complex interaction featuring diverse and dynamic shifts in modalities within both inputs and outputs.

C.5. Training Recipes

In Table 7, we list the detailed hyper-parameters setting at three stages.

C.6. Inference Process

In Figure 7 we further illustrate the inference procedure of NExT-GPT. Given certain user inputs of any combination of modalities, the corresponding modal encoders, and projectors transform them into feature representations and pass them to LLM². Then, LLM decides what content to generate, i.e., textual tokens, and modality signal tokens. If LLM identifies a certain modality content (except language) to be produced, a special type of token (Koh et al., 2023) will be output indicating the activation of that modality; otherwise, no special token output means deactivation of that modality.

¹<https://github.com/gururise/AlpacaDataCleaned>

²Except the text inputs, which will be directly fed into LLM.

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
► Existing data							
MiniGPT-4 (Zhu et al., 2023)	CC, CC3M	T+I→T	Auto		134M/-	1	5K
StableLLAVA (Li et al., 2023f)	SD	T+I→T	Auto+Manu.		126K/-	1	126K
LLaVA (Zhang et al., 2023d)	COCO	T+I→T	Auto		81K/-	2.29	150K
SVIT (Zhao et al., 2023a)	MS-COCO, VG	T+I→T	Auto		108K/-	5	3.2M
LLaVAR (Zhang et al., 2023d)	COCO, CC3M, LAION	T+I→T	LLaVA+Auto		20K/-	2.27	174K
VideoChat (Li et al., 2023d)	WebVid	T+V→T	Auto		-8K/-	1.82	-
Video-ChatGPT (Maaz et al., 2023)	ActivityNet	T+V→T	Inherit		-100K/-	1	100K
Video-LLaMA (Zhang et al., 2023e)	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto		81K/8K/-	2.22	171K
InstructBLIP (Dai et al., 2023)	Multiple	T+I/V→T	Auto		-	-	~1.6M
MIMIC-JT (Li et al., 2023b)	Multiple	T+I/V→T	Auto		8.1M/502K/-	1	2.8M
PandaGPT (Su et al., 2023)	MiniGPT-4, LLaVA	T+I→T	Inherit		81K/-	2.29	160K
MGVLID (Zhao et al., 2023b)	Multiple	T+I+B→T	Auto+Manu.		108K/-	-	108K
M ³ IT (Li et al., 2023e)	Multiple	T+I/V/B→T	Auto+Manu.		-/-	1	2.4M
LAMM (Yin et al., 2023)	Multiple	T+I+PC→T	Auto+Manu.		91K/-	3.27	196k
BuboGPT (Zhao et al., 2023c)	Clotho, VGGSS	T+A/(I+A)→T	Auto		5k/79K	-	9K
mPLUG-DocOwl (Ye et al., 2023a)	Multiple	T+I/Tab/Web→T	Inherit		-	-	-
► In this work							
T2M	Webvid, CC3M, AudioCap	T→T+I/A/V	Auto		5K/5K/5K	1	15K
MoSIT	Youtube, Google, Flickr, Midjourney, etc.	T+H+A+V → T+I+A+V	Auto+Manu.		4K/4K/4K	4.8	5K

Table 6. Summary and comparison of existing datasets for multimodal instruction tuning. T: text, I: image, V: video, A: audio, B: bounding box, PC: point cloud, Tab: table, Web: web page.

Table 7. Training recipes for NExT-GPT. The three training stages are introduced in Section C.2. Stage-1: Encoding-side Alignment Learning, Stage-2: Decoding-side Alignment Learning, Stage-3: End-to-end Instruction Tuning.

Configuration	Stage-1	Stage-2	Stage-3
Optimizer	Adam	Adam	Adam
Learning Rate	0.0004	0.0004	0.0005
Weight Decay	0.001	0.001	0.001
Training Epochs	1	1	1
Warmup Ratio	0.1	0.1	0.1
Learning Rate Scheduler	Linear	Linear	Linear
Batch Size Per GPU	18	8	4
Maximum Token Length	512	512	512
Unfreeze LLM	✗	✗	✓
Training Data			
Dataset	CC3M	CC3M	LLaVA-150K, VideoChat
	WebVid	WebVid	cleaned-Alpaca
	AudioCaps	AudioCaps	Text→Text+X, MosIT

Table 8. Text-to-image generation results on COCO-caption (Lin et al., 2014).

Method	FID (↓)
CogView (Ding et al., 2021)	27.10
GLIDE (Nichol et al., 2022)	12.24
CoDi (Tang et al., 2023)	11.26
SD (Rombach et al., 2022)	11.21
NExT-GPT	11.18

Table 10. Text-to-audio generation results on AudioCaps (Kim et al., 2019).

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
CoDi (Tang et al., 2023)	22.90	8.77
NExT-GPT	23.25	8.67

Table 9. Text-to-video generation results (zero-shot) on MSR-VTT (Xu et al., 2016).

Method	FID (↓)	CLIPSIM (↑)
CogVideo (Hong et al., 2022)	23.59	26.31
MakeVideo (Singer et al., 2022)	13.17	30.49
Latent-VDM (Rombach et al., 2022)	14.25	27.56
Latent-Shift (An et al., 2023)	15.23	27.73
CoDi (Tang et al., 2023)	—	28.90
NExT-GPT	12.69	31.97

Table 11. Audio-to-text generation (audio captioning) results on AudioCaps (Kim et al., 2019).

Method	SPIDEr	CIDEr
AudioCaps (Kim et al., 2019)	0.369	0.593
BART (Gontier et al., 2021)	0.465	0.753
AL-MixGen (Kim et al., 2022)	0.466	0.755
CoDi (Tang et al., 2023)	0.480	0.789
NExT-GPT	0.534	0.807

D. Additional Experiments

D.1. Additional Multimodal Comprehension and Generation Results

Firstly, we examine the synthesis quality of the image, video, or audio conditioned on text compared with the non-LLM-based methods. Table 8, 10, 9 present the comparisons between ours and some state-of-the-art systems. On text-to-image and text-to-audio generation tasks, NExT-GPT shows a nice performance on par with that of the best-performing baselines. Notably, under the zero-shot setting, NExT-GPT shows a significant superiority in video generation conditioning on text, demonstrating the remarkable generalization capability of NExT-GPT.

Secondly, we evaluate the NExT-GPT on the tasks of textual caption generation to test the semantic understanding capability w.r.t. image, video, or audio. The results on different tasks are shown in Table 12, 11, and 13. Significantly, NExT-GPT mostly achieves much better performance on the X-to-text generation than that of the CoDi baseline, owing to the direct generation of texts from LLM, which is inherently expertized by the LLM.

Thirdly, we test our model on text-conditioned modal editing tasks. Table 14, 16 and 15 show the performances on different tasks. Compared with the above two types of tasks, although NExT-GPT did not demonstrate superior performance on the text-conditioned modal editing tasks, it still shows competitive performance.

NExT-GPT: Any-to-Any Multimodal LLM

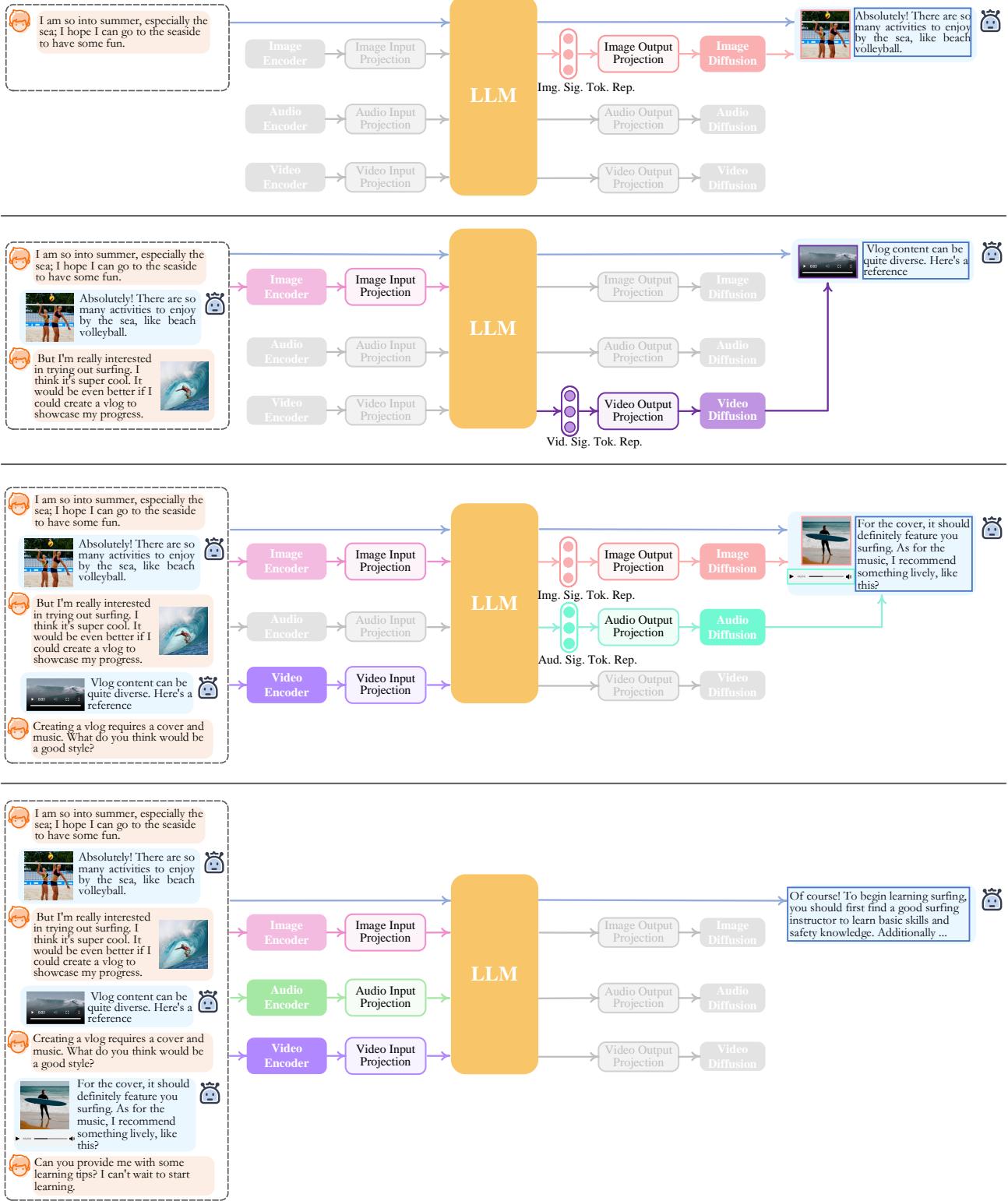


Figure 7. NExT-GPT inference process. Grey colors denote the deactivation of the modules.

Table 12. Image-to-text generation (image captioning) results on COCO-caption (Lin et al., 2014).

Method	B@4	METEOR	CIDEr
Oscar (Li et al., 2020)	36.58	30.4	124.12
BLIP-2 (Li et al., 2023c)	43.7	—	145.8
OFA (Wang et al., 2022b)	44.9	32.5	154.9
CoDi (Tang et al., 2023)	40.2	31.0	149.9
NExT-GPT	45.1	34.1	158.3

Table 14. Text+image-to-image generation (text-conditioned image editing) results on COCO-caption (Lin et al., 2014).

Method	Object		Background	
	CLIP (\uparrow)	FID (\downarrow)	CLIP (\uparrow)	FID (\downarrow)
PTP (Hertz et al., 2023)	30.33	9.58	31.55	13.92
BLDM (Avrahami et al., 2023)	29.95	6.14	30.38	20.44
DiffEdit (Couairon et al., 2023)	29.30	3.78	26.92	1.74
PFB-Diff (Huang et al., 2023d)	30.81	5.93	32.25	13.77
NExT-GPT	29.32	6.62	27.31	14.27

Table 13. Video-to-text generation (video captioning) results on MSR-VTT (Xu et al., 2016).

Method	B@4	METEOR
ORG-TRL (Zhang et al., 2020)	43.6	28.8
GIT (Wang et al., 2022a)	54.8	33.1
mPLUG-2 (Xu et al., 2023)	57.8	34.9
CoDi (Tang et al., 2023)	52.1	32.5
NExT-GPT	58.8	39.6

Table 15. Text+video-to-video generation (text-conditioned video editing) results on DAVIS (Perazzi et al., 2016).

Method	CLIP-T	CLIP-I
CogVideo (Hong et al., 2022)	0.2391	0.9064
TuneVideo (Wu et al., 2022)	0.2758	0.9240
SDEdit (Meng et al., 2022)	0.2775	0.8731
Pix2Video (Ceylan et al., 2023)	0.2891	0.9767
NExT-GPT	0.2684	0.9647

Table 16. Text+audio-to-audio generation (text-conditioned speech editing) results on VCTK (Veaux et al., 2017).

Method	MCD (\downarrow)
CampNet (Wang et al., 2022c)	0.380
MakeAudio (Huang et al., 2023a)	0.375
AudioLDM-L (Liu et al., 2023a)	0.349
NExT-GPT	0.300

D.2. Human Evaluation on Complex Any-to-any QA

We also carry out evaluation on some more scenarios where there are complicated cross-modal interactions between inputs and outputs. We mainly compare the model performance for the settings with different modality conversions. As no standard benchmark can be leveraged, here we adopt human evaluation. We ask several evaluators to score the performance of NExT-GPT on a scale from 1 to 10. Figure 8 shows the comparisons. We find NExT-GPT is more competent in producing images, compared with the generations on videos and audio. Also generating mixed combinations of multimodal content is slightly inferior to the generation of single-modal content, due to the complexity of the latter.

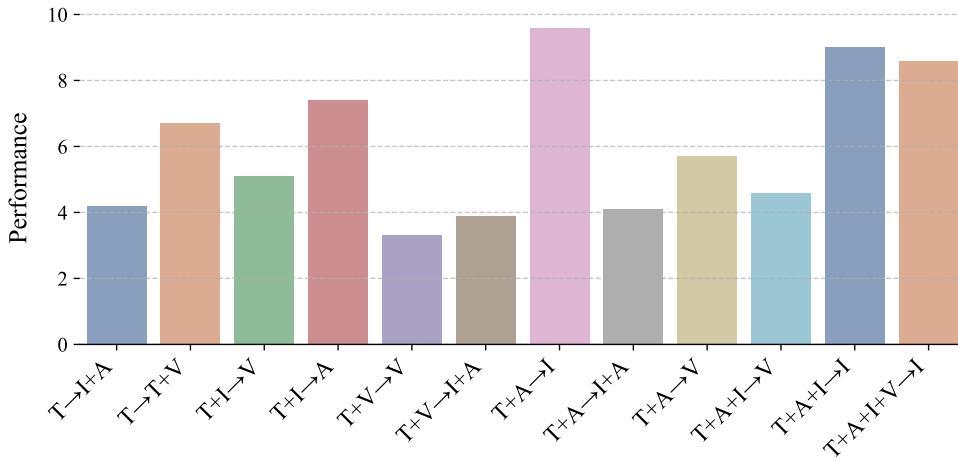


Figure 8. Comparative performance of NExT-GPT on various complex cross-modal conversions.

D.3. Case Study on Pipeline-style vs. End-to-end Unification

We earlier have elaborated on the difference as well as the necessity of building a unified any-to-any multimodal LLM in an end-to-end manner, compared with the existing pipeline-style systems that generate intermediate captions and then pass to

the downstream tools (e.g., diffusion models for generation). The cascade process inevitably introduces noise and propagates errors. Meanwhile, the entire system only leverages existing pre-trained tools for inference only, whereas, without end-to-end updating throughout the whole system, the capability to more accurately interpret complex user instructions and generate content will be compromised. Here we add a few illustrations, where we make comparisons with these pipeline-style systems: 1) Visual-ChatGPT and HuggingGPT, which are existing systems that have free open access; 2) NExT-GPT variant with captions as the messenger (which we mark as NExT-GPT-caption). To implement NExT-GPT-caption, the captions directly generated by LLM will be fed into the following generation models, instead of using the soft representations of the signal tokens. As Visual-ChatGPT only supports image generation, we here consider the evaluation on the Text-to-Text&Image setting.

Instruction (a): Generate me a figure of a cat running in the garden.

Response: I have generated an image of a cat running in the garden, the file name is image/00e38ab0.png.



{*Intermediate Caption:* a cat running in the garden}

Visual-ChatGPT

Response: Here is a figure of a cat running in the garden: <Image>A cat running in the garden</Image>



{*Intermediate Caption:* A cat running in the garden}

NExT-GPT-caption

Response: Certainly! Here's an image of cat running a garden.



NExT-GPT

Figure 9. Illustration of case study, image generation from a simple instruction on Visual-ChatGPT, NExT-GPT-caption, and NExT-GPT.

Figure 9 presents the case of image generation from a simple input user instruction; while Figure 10 and 11 present two cases of image generation from comparatively complex input user instructions. On the simple one, all generated image content from both pipeline-style and end-to-end (ours) systems seem correct and coincide with the input prompt. However, when handling complex instructions, as seen in Figure 10 and 11, the generated image content can be wrong and biased to the user’s intention. The problems are rooted in the core of different modalities, i.e., there are inherent gaps between language and visual modalities that cannot be eliminated. Here are two representative attributes: **the numeration of vision** (cf. Figure 10) and **the visual-spatial relational semantics** (cf. Figure 11), which could be hard to (or even cannot) be expressed by the intermediate captions conveniently. Utilizing textual captions as intermediate representations runs the risk of overlooking these modality-specific features when expressing non-linguistic (e.g., visual) modalities solely through language.

By the way, we kindly note a fact that, with the intermediate captions produced from the pipeline-style systems in Figure 10 and 11, the Stable Diffusion model just has difficulty in accurately understanding the vision numeration and visual-spatial relation and generating correct answers, i.e., they are the problems inherent to the Stable Diffusion model itself, and Stable Diffusion alone is tricky to overcome. Most recent work tries to solve this issue by integrating the vision-specific features into the Stable Diffusion (Feng et al., 2023; Qu et al., 2023a) via additional feature engineering. But, in our NExT-GPT with an end-to-end solution, the implicit modality signal token embeddings that carry rich modality-specific features of non-linguistic will be naturally encoded and passed to the downstream modules (e.g., Stable Diffusion), without any further external effort.

D.4. Example Demonstrations

Figure 12, 13, 14, 15, 16, 17, 18 and 19 show several real examples produced by NExT-GPT.

Instruction (b): There were originally 7 apples on the table, but I ate one. Then, how many apples are left now? Please generate a picture to describe the result.

Response: Sure, I'd be happy to help you understand the process and results.

Based on your statement, "There were originally 7 apples on the table, but I ate one", I inferred that there are now 6 apples left on the table.

To visualize this, I used a text-to-image model to generate an image based on your description. The model I used is called 'runwayml/stable-diffusion-v1-5', which is currently the only available model for this task. The image generated can be found at the following path: '/images/ec4d.jpg'.

Please note that the image is a representation of the text and may not be 100% accurate. If there's anything else you need, feel free to ask!

HuggingGPT



{*Intermedia Caption:* 6 apples left on the table}

Response: Sure! As you can see, the number of apples has decreased from 7 to 6. Here's an image that shows the results after one apple was eaten: <Image>An image of a table with 6 apples</Image>

NExT-GPT-caption



{*Intermedia Caption:* An image of a table with 6 apples}

Response: Certainly! In the end, you have the following outcomes: 6 apples on the table.

NExT-GPT



Figure 10. Illustration of image generation from a complex instruction on HuggingGPT, NExT-GPT-caption, and NExT-GPT. In this case, the numeration understanding of vision has been wrong by pipeline-style methods.

Instruction (c): In the park, an old lady is sitting to the right of an old man, and a small dog with yellow and white fur is lying at the old man's arm. What a heartwarming scene! Can you help me visualize it in an image?

Response: Sure, I'd be happy to help you visualize the scene you described. Based on your description, I used a model to generate an image that represents the scene. Unfortunately, as an AI, I can't directly show you the image, but I can tell you where to find it. The generated image is stored at this location: '/images/a62c.jpg'. To accomplish this, I used the 'runwayml/stable-diffusion-v1-5' model. This model was chosen because it was the only one available for this task. The model took your description and used it to generate an image that represents the scene you described.

Please note that while the model does its best to accurately represent the scene, the results may not always be perfect due to the complexity of interpreting and visualizing text descriptions.



HuggingGPT

{*Intermedia Caption:* In the park, an old lady is sitting to the right of an old man, and a small dog with yellow and white fur is lying at the old man's arms}

Response: Certainly! Here's an image that captures an old woman is sitting on a bench, surrounded by a park with a small dog lying at the old man's arms. <Image>An image of an old woman sitting on a bench, surrounded by a park with a small dog lying at the old man's arms</Image>



NExT-GPT-caption

{*Intermedia Caption:* An image of an old woman sitting on a bench, surrounded by a park with a small dog lying at the old man's arms}

Response: Certainly! Here's an image that captures the warmth and companionship between the old lady and the old man, as well as the peaceful presence of the small dog at their arms.



NExT-GPT

Figure 11. Illustration of image generation from another complex instruction on HuggingGPT, NExT-GPT-caption, and NExT-GPT. In this case, the understanding of visual-spatial relational semantics has been wrong by pipeline-style methods.

How many cars are in this picture?



There are two cars shown in this picture.

Tell me what kind of car are they?

The two cars in the picture are SUVs.

Can you show me the sound of starting the engines of them.

Sure, here's the sound of the engine of the SUV:

▶ 0:00 / 0:09

Figure 12. Example of Text+Image → Text+Audio.

I'm planning a trip to New Orleans, and I'd love some recommendations for fun things to do and places to visit.



New Orleans is a vibrant city with a rich history and culture. Some popular attractions include the French Quarter, the National WWI Museum, and the St. Louis Cathedral. You can also explore the historic Garden District and visit the iconic Bourbon Street for some night life. Additionally, don't miss the opportunity to try some delicious Cajun cuisine!



What kind of cajun cuisine? can you show me some pictures?



Certainly! Cajun cuisine is known for its hearty and flavorful dishes. Here are some pictures to give you a taste of what you can expect:

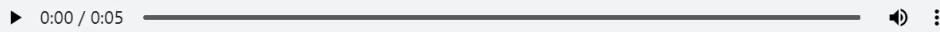


I've heard that New Orleans is famous for jazz music. Show me a piece of jazz music please.



You are right. Here is a jazz music:





Fantastic! I often see videos in which a bunch of people dance to jazz music, can you show me some?



Of course!

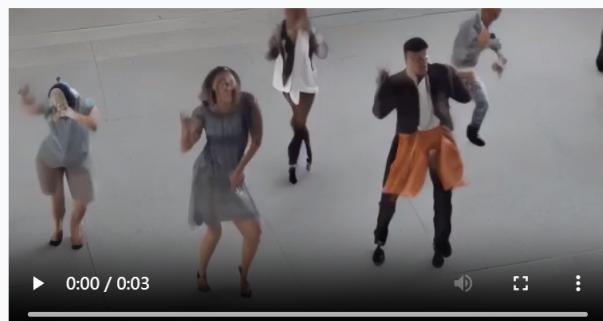


Figure 13. Example of Text → Text+Image+Video+Audio.

Show me another image in similar contents.



Is there any suitable place to see this scene?



Yes, there is a suitable place to see this scene. You can see the beautiful pink cherry blossoms in full bloom on a tree next to a river. The image is taken from a distance, allowing you to appreciate the vibrant colors and the serene atmosphere.

Use a sound to describe the beautiful scene.

Sure!



It would be great to see a video that shows the stunning scene.

Certainly! Here is the video scene.



Figure 14. Example of Text+Image → Text+Image+Video+Audio.

Is the guy in the video iron man?



0:00 / 0:08

▶ ⏸ ⏹ ⏷

⌚ Yes, the guy in the video is iron man.

⌚ Could you show me a comic-style picture of iron man?

⌚ Certainly! Here is a comic-style picture of Iron Man:



⌚ I need a more abstract one.

⌚ Sure, here's an abstract image of Iron Man:



Figure 15. Example of Text+Video → Text+Image.

Describe what can be heard from the sound:

▶ 0:00 / 0:01





The sound of a growling lion is a deep, rumbling noise that starts low in the chest and gradually builds in intensity. It is often associated with a sense of power and strength, as well as a sense of danger or warning. The growl of a lion can be heard from a distance and is often used by the animal to communicate with other members of its pride.

Displaying an image to help visualize it would be fantastic.



Sure, here is an image of a lion growling:



That looks great! How about showing a video that explains how the lions catch its prey?

▶ 0:00 / 0:03



Figure 16. Example of Text+Audio → Text+Image+Video.

What is the man doing in the video?



The man in the video is celebrating his birthday by lighting a candle on a cake and making a wish.

What is the man wearing?

The man is wearing a white shirt.

How about synthesizing a song to celebrate his birthday please?

It would be a great idea to synthesize a song to celebrate the man's birthday. Here's a sample of a song that could be used:

- Verse 1: Happy birthday, dear friend, You're shining bright, With a smile so wide. You're the light in our lives
- Chorus: Celebrate, celebrate, With laughter and joy, With love and happiness, Bringing happiness to those around
- Verse 2: You're the one who brings us together, With your warmth and your grace, You're the one who makes us feel, A part of this special place.
- Chorus: Celebrate, celebrate, With laughter and joy, With love and happiness, Bringing happiness to those around
- Bridge: So let's raise a glass, To a life filled with love and grace. With laughter and joy, May your birthday be filled with peace
- Chorus: Celebrate, celebrate, With laughter and joy, With love and happiness, Bringing happiness to those around

I guess you made me wrong. I mean, show the song in an audio.

Sorry for the mistake! Here's an audio of the song:



Figure 17. Example of Text+Video → Text+Audio.

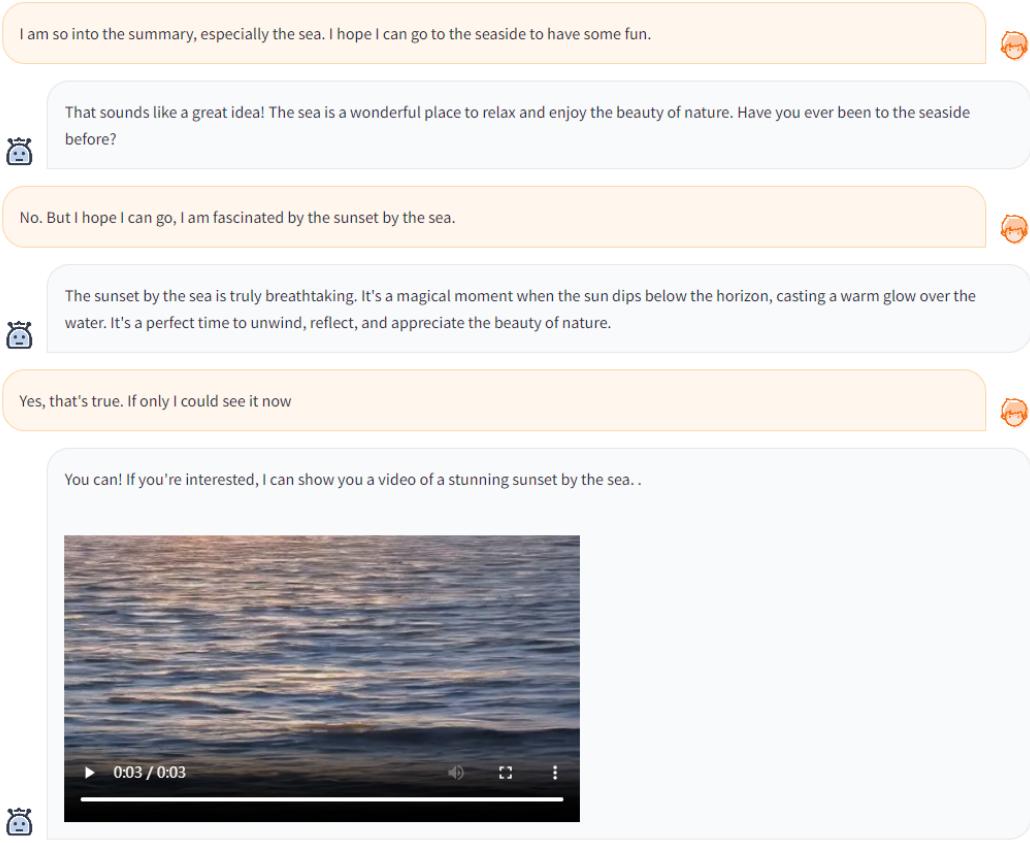


Figure 18. Example of Text → Text+Video.

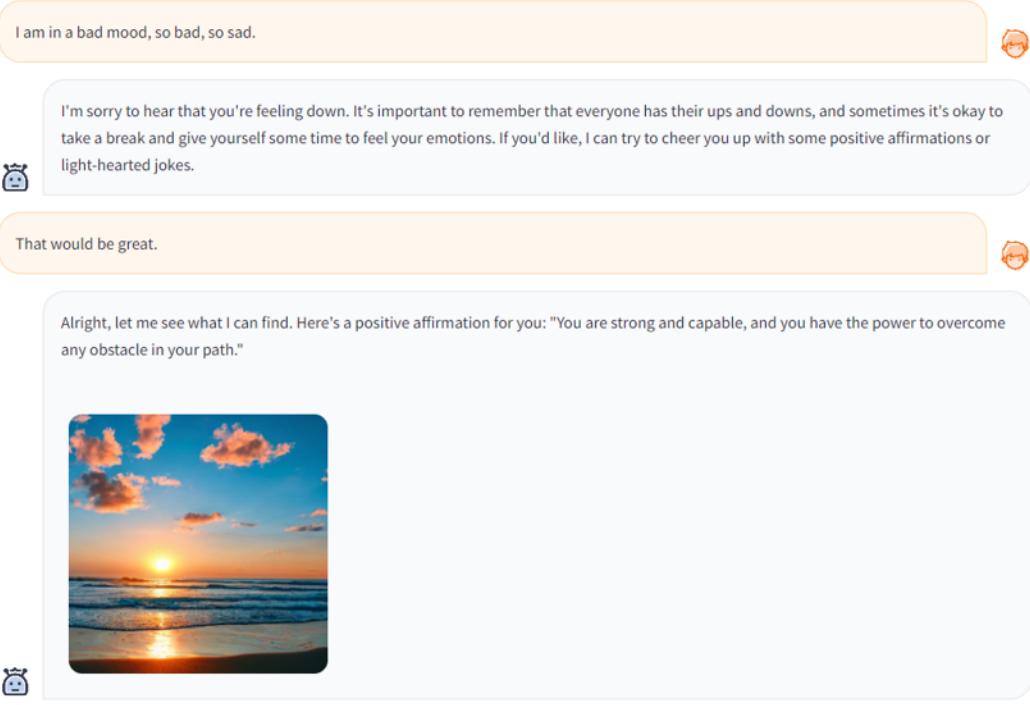


Figure 19. Example of Text → Text+Image.