NExT-GPT: Any-to-Any Multimodal LLM

Shengqiong Wu [1] **Hao Fei** [1] **Leigang Qu** [1] **Wei Ji** [1] **Tat-Seng Chua** [1]

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 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)

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

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

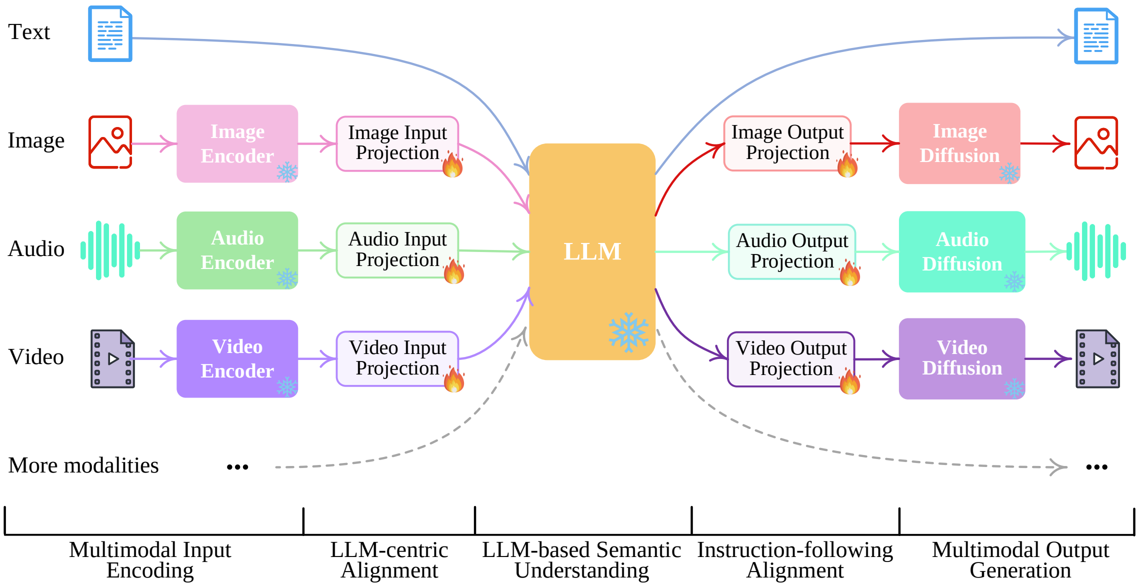


Figure 1. By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output. ❄and 🔥represents the frozen and trainable modules, respectively.

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 (Chiang 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	—	—	—	—			—	—	—	—
Image	ImageBind	1.2B❄	Grouping	28M🔥	Vicuna	7B❄	Transformer	31M🔥	SD	1.3B❄
Audio					(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM	975M❄
Video							Transformer	32M🔥	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 (Cerspense, 2023), and AudioLDM (`l-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.

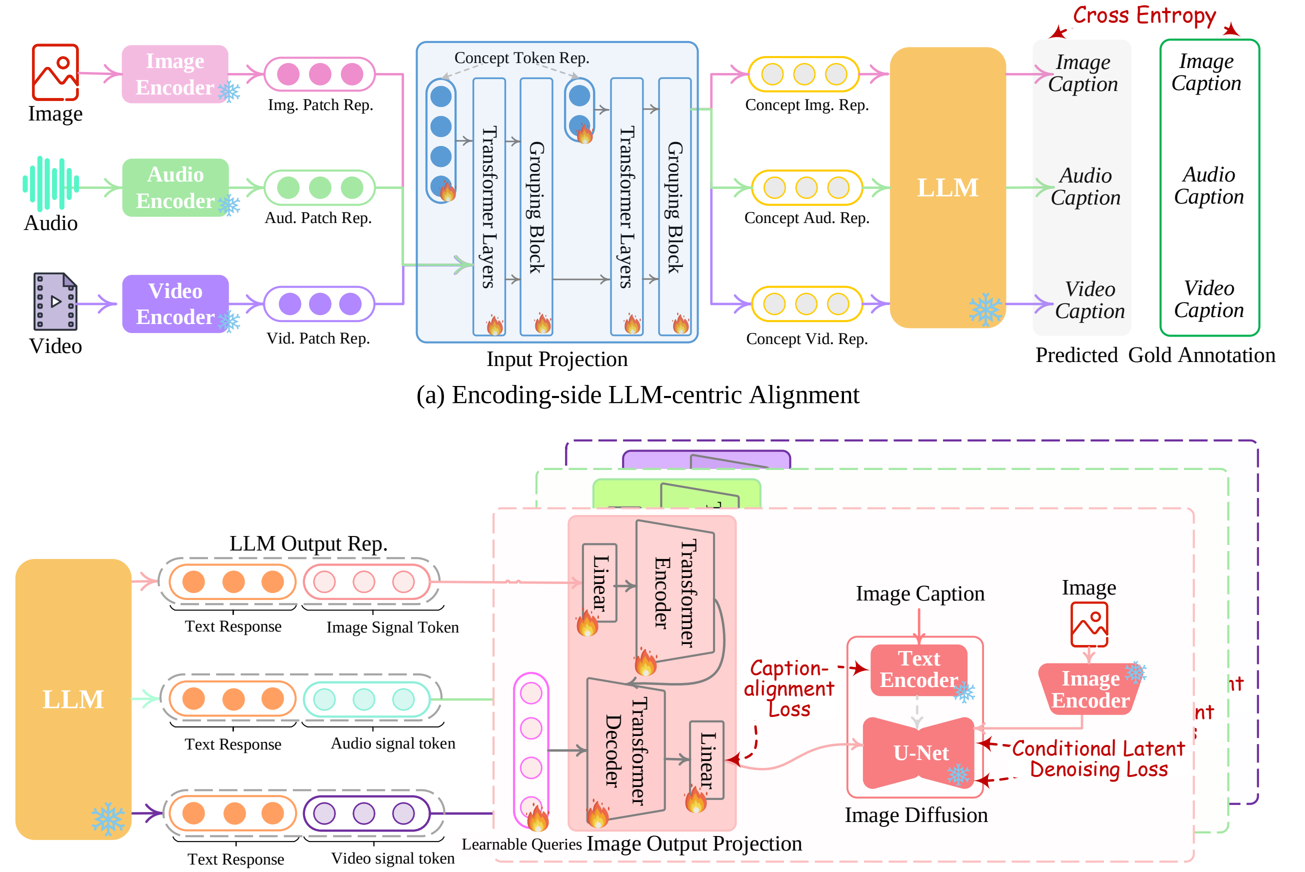
4.1. Encoding-side LLM-centric Multimodal Alignment

Most existing MM-LLMs adopt the Transformer-architectured 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$_i$]*' ($i = 0, \cdots, 4$) as image signal



(a) Encoding-side LLM-centric Alignment

(b) Decoding-side Instruction-following Alignment

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

tokens; ‘*[AUD$_i$]*’ ($i = 0, \cdots, 8$) as audio signal tokens; and ‘*[VID$_i$]*’ ($i = 0, \cdots, 24$) 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.

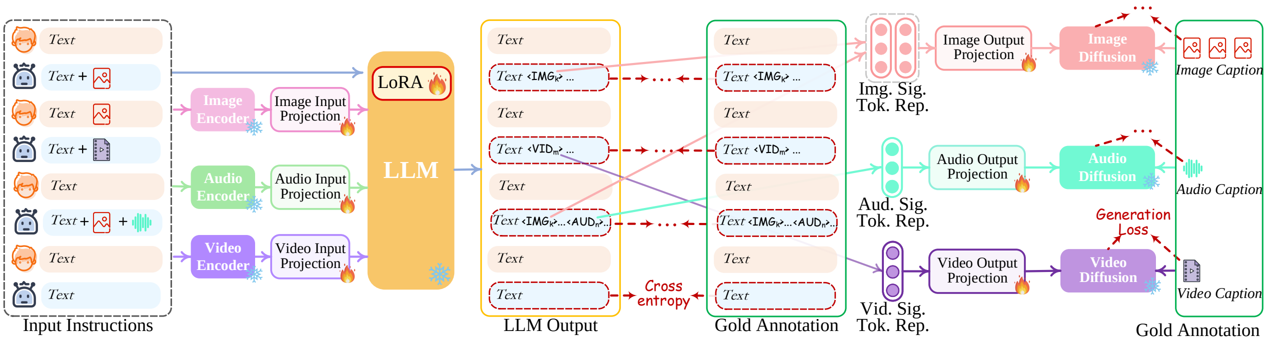


Figure 3. Illustration of modality-switching instruction tuning.

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 (↑) 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	123.1	82.4	102.2	-	33.4	33.9	36.0	-
LLaVA (Liu et al., 2023b)	LLaMA-2-7B-Chat	120.7	82.7	-	-	-	-	36.2	-
mPLUG-Owl (Ye et al., 2023b)	LLaMA-7B	117.0	80.3	119.3	-	39.0	-	46.6	34.0
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**	48.1	-	**60.9**	-
NExT-GPT	Vicuna-7B	**123.7**	**84.5**	**124.9**	66.7	**48.4**	**52.1**	58.0	**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)	-	74.4*	-	-	-	78.9*
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	-	29.6	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	**70.7**	59.2	-	-
Emu (Sun et al., 2023)	LLaMA-7B	-	32.4	14.0	6.8	-
NExT-GPT	Vicuna-7B	**76.2***	64.5	**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)). †: zero-shot results.

Model	Image	Audio	Video
	FID (↓)	FAD (↓)	CLIPSIM (↑)
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† (Koh et al., 2023)	12.20	-	-
Emu-13B† (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†	11.18	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.

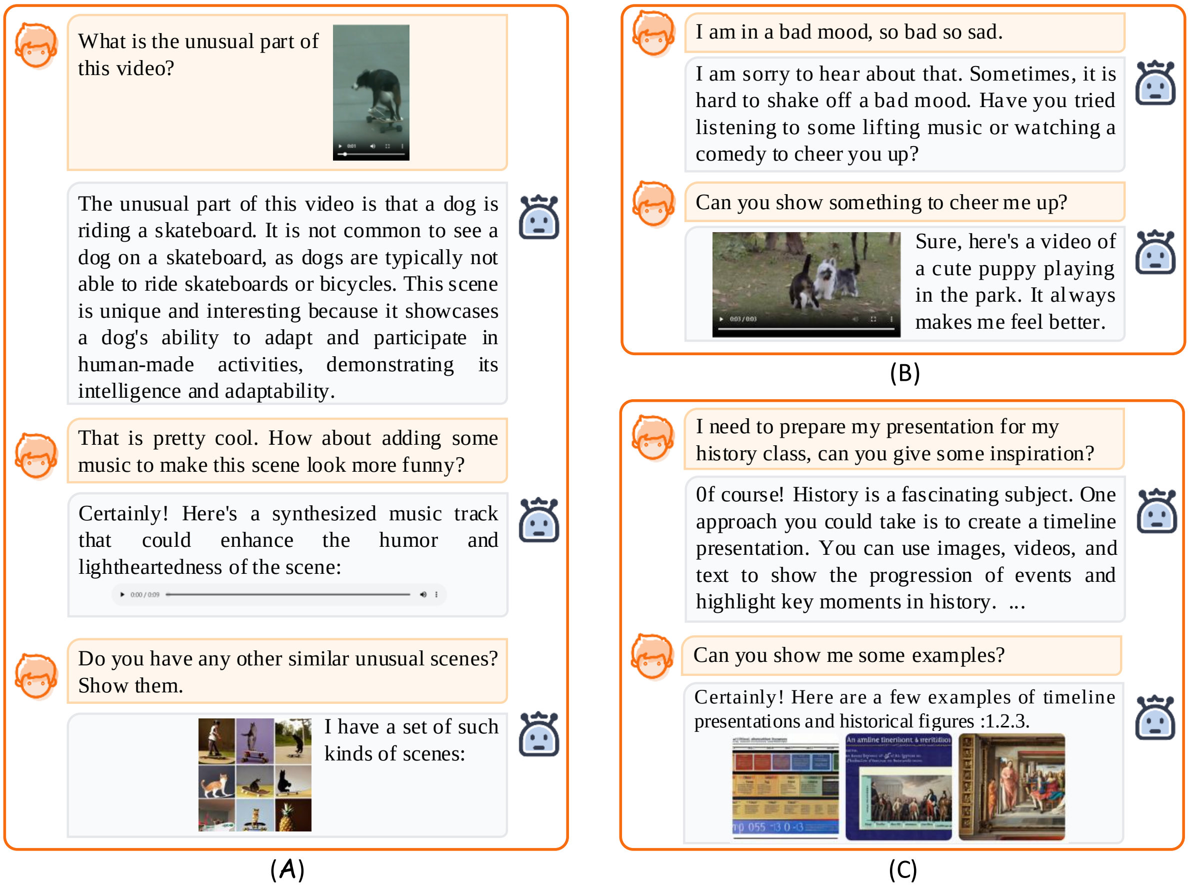
6.1. Main Results

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



What is the unusual part of this video?

The unusual part of this video is that a dog is riding a skateboard. It is not common to see a dog on a skateboard, as dogs are typically not able to ride skateboards or bicycles. This scene is unique and interesting because it showcases a dog's ability to adapt and participate in human-made activities, demonstrating its intelligence and adaptability.

That is pretty cool. How about adding some music to make this scene look more funny?

Certainly! Here's a synthesized music track that could enhance the humor and lightheartedness of the scene:

Do you have any other similar unusual scenes? Show them.

I have a set of such kinds of scenes:

(A)

I am in a bad mood, so bad so sad.

I am sorry to hear about that. Sometimes, it is hard to shake off a bad mood. Have you tried listening to some lifting music or watching a comedy to cheer you up?

Can you show something to cheer me up?

Sure, here's a video of a cute puppy playing in the park. It always makes me feel better.

(B)

I need to prepare my presentation for my history class, can you give some inspiration?

0f course! History is a fascinating subject. One approach you could take is to create a timeline presentation. You can use images, videos, and text to show the progression of events and highlight key moments in history. ...

Can you show me some examples?

Certainly! Here are a few examples of timeline presentations and historical figures :1.2.3.

(C)

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

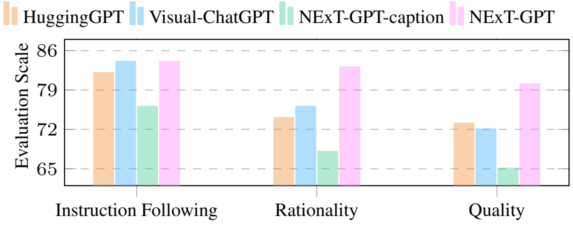
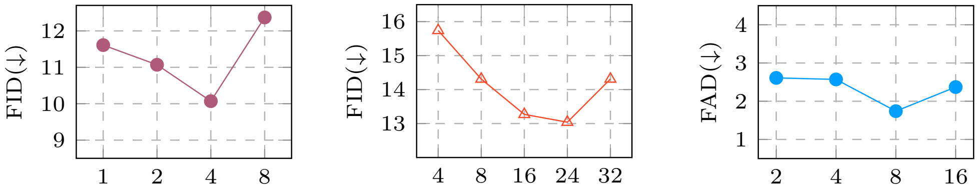


Figure 5. Human Evaluation (1-100 scale, results are on average) of NExT-GPT in comparison with pipeline baselines.

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



(a) Text-to-image Generation (b) Text-to-video Generation (c) Text-to-audio Generation

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.