arXiv:2309.05519v1 [cs.AI] 11 Sep 2023

NExT-GPT: Any-to-Any Multimodal LLM

Shengqiong Wu **Hao Fei**[*] **Leigang Qu** **Wei Ji** **Tat-Seng Chua**
NExT++, School of Computing, National University of Singapore

Project: `https://next-gpt.github.io/`

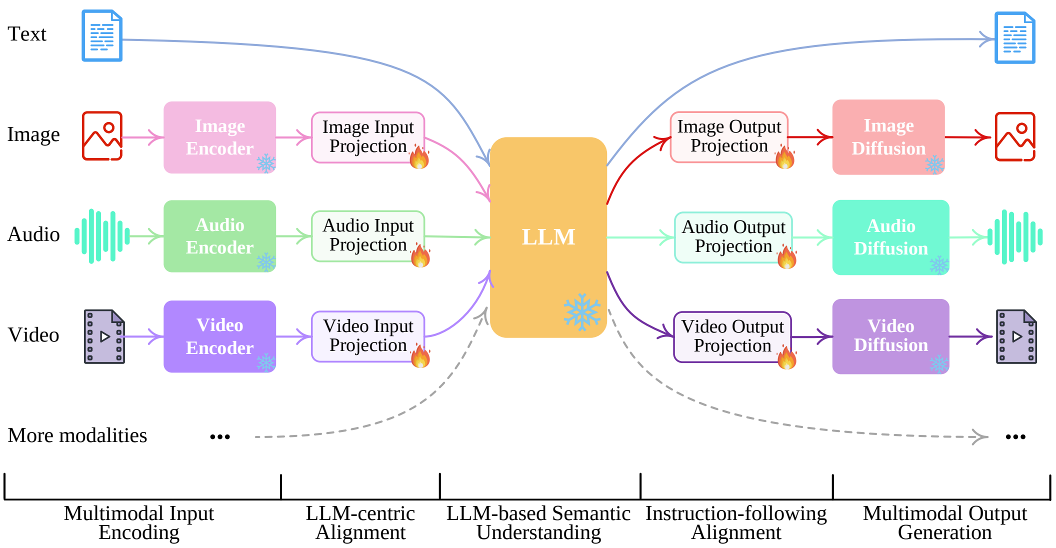


Figure 1: By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output.

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, images, videos, and audio. By leveraging the existing well-trained highly-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 and 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 an AI agent capable of modeling universal modalities, paving the way for more human-like AI research in the community.

---

[*] Hao Fei is the corresponding author: `haofei37@nus.edu.sg`

Preprint, work in progress.

1 Introduction

Recently, the topic of Artificial Intelligence Generated Content (AIGC) has witnessed unprecedented advancements with certain technologies, such as ChatGPT for text generation [59] and diffusion models for visual generation [21]. Among these, the rise of Large Language Models (LLMs) has been particularly remarkable, e.g., Flan-T5 [13], Vicuna [12], LLaMA [80] and Alpaca [79], 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 visual, 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 [43], Flamingo [1], MiniGPT-4 [110], Video-LLaMA [104], LLaVA [52], PandaGPT [77], SpeechGPT [103]. Nevertheless, most of these efforts pay the attention to the multimodal content understanding at the input side, lacking the ability to output content in multiple modalities more than texts. We emphasize that real human cognition and communication indispensably require seamless transitions between any modalities of information. This makes the exploration of any-to-any MM-LLMs critical to achieving real AGI, i.e., accepting inputs in any modality and delivering responses in the appropriate form of any modality.

Certain efforts have been made to mimic the human-like any-to-any modality conversion. Lately, CoDi [78] has made strides in implementing the capability of simultaneously processing and generating arbitrary combinations of modalities, while it lacks the reasoning and decision-making prowess of LLMs as its core, and also is limited to the simple paired content generation. On the other hand, some efforts, e.g., visual-ChatGPT [88] and HuggingGPT [72] 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 the complete pipeline architecture. First, the information transfer between different modules is entirely based on discrete texts produced by the LLM, where the cascade process inevitably introduces noise and propagates errors. More critically, the entire system only leverages existing pre-trained tools for inference only. Due to the lack of overall end-to-end training in error propagation, 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 for constructing 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, images, videos, 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 the 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 whether & what modal content to output correspondingly. **Third**, the produced multimodal signals with specific instructions, after projection, route 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 Q-Former [43], ImageBind [25] and the state-of-the-art latent diffusion models [68, 69, 8, 2, 51, 33]. By loading the off-the-shelf parameters, we not only avoid cold-start training but also facilitate the potential growth of more modalities. For the feature alignment across the three tiers, we consider fine-tuning locally only 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* (termed `Mosit`), equipping 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 samples of high quality. Employing the LoRA technique [32], we fine-tune the overall NExT-GPT system on `MosIT` data, updating the 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 project are as follows:

- We for the first time present an end-to-end general-purpose any-to-any MM-LLM, NExT-GPT, capable of semantic understanding and reasoning and generation of free input and output combinations of text, images, videos, 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 minimal parameter adjustments (only 1% params) for effective semantic alignment.
- We annotate a high-quality modality-switching instruction tuning dataset covering intricate instructions across various modal combinations of text, images, videos, and audio, aiding MM-LLM with human-like cross-modal content understanding and instruction reasoning.

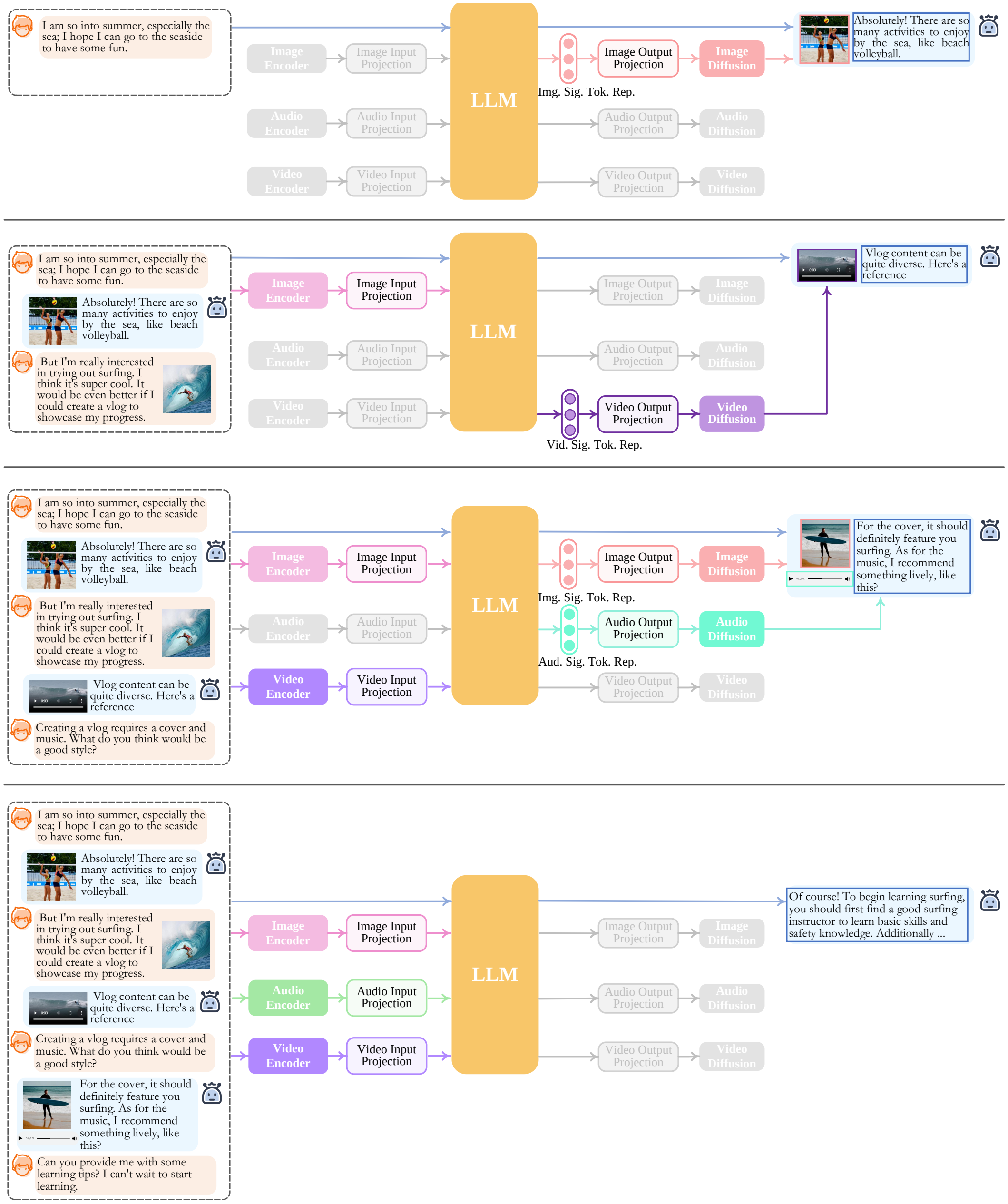
2 Related Work

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, such as Image/Video Captioning [100, 16, 56, 56, 27, 49], Image/Video Question Answering [94, 90, 48, 98, 3], Text-to-Image/Video/Speech Synthesis [74, 30, 84, 23, 17, 51, 33], Image-to-Video Synthesis [18, 37] 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 [82, 102, 17, 24], GANs [53, 7, 93, 111], VAEs [81, 67], Flow models [73, 6] and the current state-of-the-art diffusion models [31, 64, 57, 22, 68]. Especially, the diffusion-based methods have recently delivered remarkable performance in a plethora of cross-modal generation tasks, such as DALL-E [66], CogView [17], and Pariti [99]. While all previous efforts of cross-modal learning are limited to the comprehension of multimodal inputs only, CoDi [78] 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 [59] and GPT4 [60], with alignment techniques such as instruction tuning [61, 47, 105, 52] and reinforcement learning from human feedback (RLHF) [75], have demonstrated remarkable language understanding and reasoning abilities. And a series of open-source LLMs, e.g., Flan-T5 [13], Vicuna [12], LLaMA [80] and Alpaca [79], have greatly spurred advancement and made contributions to the community [110, 101]. 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 [35, 110, 76, 40]. For example, Flamingo [1] uses a cross-attention layer to connect a frozen image encoder to the LLMs. BLIP-2 [43] employs a Q-Former to translate the input image queries to the LLMs. LLaVA [52] 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 [44] and Video-LLaMA [104]), audios (e.g., SpeechGPT [103]), etc. Profoundly, PandaGPT [77] achieves a comprehensive understanding of six different modalities simultaneously by integrating the multimodal encoder, i.e., ImageBind [25].

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 visual-ChatGPT [88], HuggingGPT [72], and AudioGPT [34]. As aforementioned, passing


I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

Vlog content can be quite diverse. Here's a reference

Creating a vlog requires a cover and music. What do you think would be a good style?

For the cover, it should definitely feature you surfing. As for the music, I recommend something lively, like this?

Can you provide me with some learning tips? I can't wait to start learning.

Of course! To begin learning surfing, you should first find a good surfing instructor to learn basic skills and safety knowledge. Additionally ...

Figure 2: NExT-GPT inference process. Grey colors denote the deactivation of the modules.

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.

3 Overall Architecture

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

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., Q-Former [43], ViT [19], CLIP [65]. Here we take advantage of the ImageBind [25], which is a unified high-performance encoder across six modalities. With ImageBind, we are spared from managing

	Encoder		**Input Projection**		**LLM**		**Output Projection**		**Diffusion**	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	Vicuna [12] (LoRA	7B❄ 33M🔥)	—	—	—	—
Image	ImageBind [25]	1.2B❄	Linear	4M🔥			Transformer	31M🔥	SD [68]	1.3B❄
Audio							Transformer	31M🔥	AudioLDM [51]	975M❄
Video							Transformer	32M🔥	Zeroscope [8]	1.8B❄

Table 1: Summary of system configuration. Only 1% parameters need updating.

many numbers of heterogeneous modal encoders. Then, via the linear projection layer, different input 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[2] [12], which is the open-source text-based LLM that is widely used in the existing MM-LLMs [77, 104]. 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 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 following multimodal decoders. Technically, we employ the current off-the-shelf latent conditioned diffusion models of different modal generations, i.e., Stable Diffusion (SD)[3] for image synthesis [68], Zeroscope[4] for video synthesis [8], and AudioLDM[5] for audio synthesis [51]. The signal representations are fed into the condition encoders of 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 encoders and decoders frozen. That is, 131M(=4+33+31+31+32) / [131M + 12.275B(=1.2+7+1.3+1.8+0.975)], only **1%** parameters are to be updated. This is also one of the key advantages of our MM-LLM.

In Figure 2 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 LLM[6]. 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 [40] will be output indicating the activation of that modality; otherwise, no special token output means deactivation of that modality. Technically, we design the '<IMG_i>' ($i = 0, \cdots, 4$) as image signal tokens; '<AUD_i>' ($i = 0, \cdots, 8$) as audio signal tokens; and '<VID_i>' ($i = 0, \cdots, 24$) as video signal tokens. After LLM, the text responses are output to the user; while the representations of the signal tokens of certain activated modalities are passed to the corresponding diffusion decoders for content generation.

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

Following the common practice of existing MM-LLMs, we consider aligning different inputting multimodal features with the text feature space, the representations that are understandable to the core

---

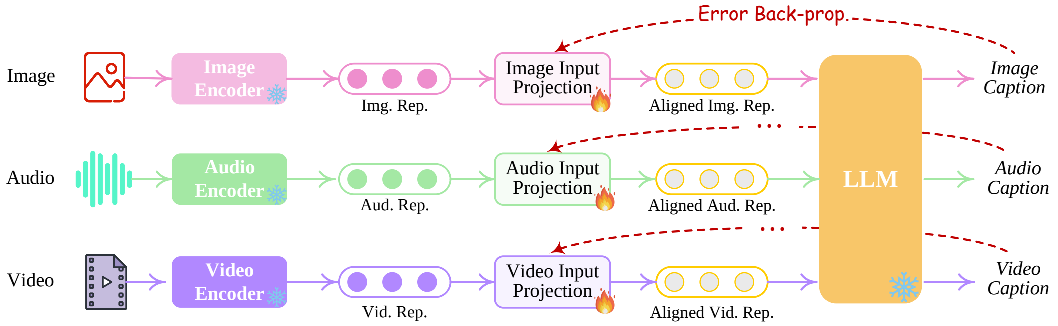
[2] `https://huggingface.co/lmsys/vicuna-7b-delta-v0`, 7B, version 0

[3] `https://huggingface.co/runwayml/stable-diffusion-v1-5`, version `1.5`.

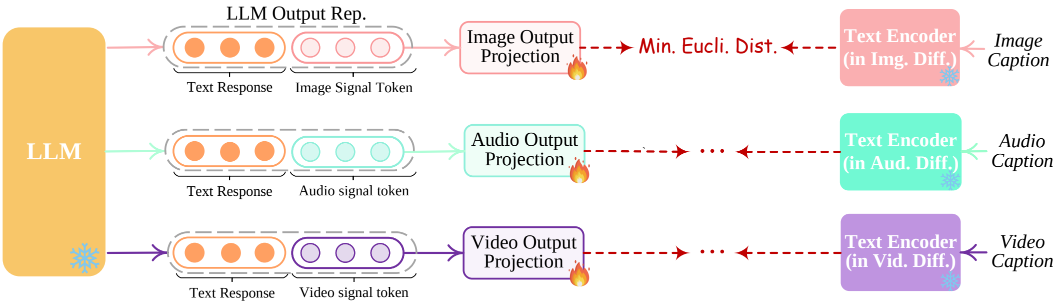
[4] `https://huggingface.co/cerspense/zeroscope_v2_576w`, version `zeroscope_v2_576w`.

[5] `https://audioldm.github.io/`, version `audioldm-l-full`.

[6] Except the text inputs, which will be directly fed into LLM.



(a) Encoding-side LLM-centric Alignment



(b) Decoding-side Instruction-following Alignment

Figure 3: Illustration of the lightweight multimodal alignment learning of encoding and decoding.

LLM. This is thus intuitively named the LLM-centric multimodal alignment learning. To accomplish the alignment, we prepare the 'X-caption' pair ('X' stands for image, audio, or video) data from existing corpus and benchmarks. We enforce LLM to produce the caption of each input modality against the gold caption. Figure 3(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 here explore a more efficient approach, decoding-side instruction-following alignment, as depicted in Figure 3(b). Specifically, since diffusion models of various modalities are conditioned solely on textual token inputs. This conditioning diverges significantly from the modal signal tokens from LLM in our system, which leads to a gap in the diffusion models' accurate interpretation of the instructions from LLM. Thus, we consider minimizing the distance between the LLM's modal signal token representations (after each Transformer-based project layer) and the conditional text representations of the diffusion models. Since only the textual condition encoders are used (with the diffusion backbone frozen), the learning is merely based on the purely captioning texts, i.e., without any visual or audio inputs. This also ensures highly lightweight training.

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 desired multimodal outputs. To address this, further instruction tuning (IT) [97, 77, 52] is deemed necessary to enhance the capabilities and controllability of LLM. 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 [32] 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 4, when an IT dialogue sample is fed into the system, the LLM reconstructs and generates the textual

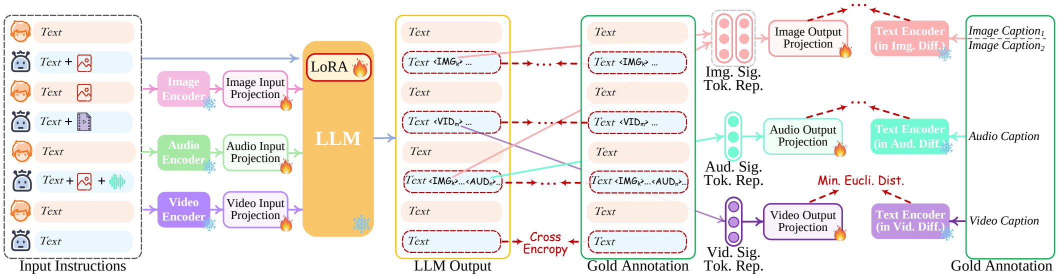


Figure 4: Illustration of modality-switching instruction tuning.

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 the LLM tuning, we also fine-tune the decoding end of NExT-GPT. We align the modal signal token 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 consider the following datasets.

'Text+X' — 'Text' Data The commonly used datasets for MM-LLM IT entail inputs of both texts and multimodal contents (i.e., 'X' could be the image, video, audio, or others), and the outputs are textual responses from LLM. There are well-established data of this type, e.g., LLaVA [52], miniGPT-4 [110], VideoChat [44], where we directly employ them for our tuning purpose.

'Text' — 'Text+X' Data Significantly unlike existing MM-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' data, i.e., text-to-multimodal (namely T2M) data. Based on the rich volume of 'X-caption' pairs from the existing corpus and benchmarks [71, 50, 5, 38], with some templates, we borrow GPT-4 to produce varied textual instructions to wrap the captions, and result in the data.

`MosIT` **Data** 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 processing and understanding of complex user intentions is required. However, the above two types of data 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`). `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` data 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 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, 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 containing multimodal contents (e.g., image, audio, and video) in the conversations, we look for the best-matched contents from the external resources, including the retrieval systems, e.g., Youtube[7], and even AIGC tools, e.g., Stable-XL [63],

---

[7] https://www.youtube.com/

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
► ***Existing data***							
MiniGPT-4 [110]	CC [10], CC3M [71]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [47]	SD [68]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [105]	COCO [50]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [107]	MS-COCO [50], VG [41]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [105]	COCO [50], CC3M [71], LAION [70]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [44]	WebVid [5]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [54]	ActivityNet [28]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [104]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [15]	Multiple	T+I/V→T	Auto	✗	-	-	∼ 1.6M
MIMIC-IT [42]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [77]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [108]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M^3IT [45]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [97]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [109]	Clotho [20], VGGSS [11]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [96]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
► ***In this work***							
T2M	Webvid [5], CC3M [71], AudioCap [38]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
`MosIT`	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

Table 2: 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.

Method	FID (↓)
CogVideo [17]	27.10
GLIDE [58]	12.24
CoDi [78]	11.26
SD [68]	**11.21**
NExT-GPT	11.28

Table 3: Text-to-image generation results on COCO-caption data [50].

Method	FD (↓)	IS (↑)
DiffSound [95]	47.68	4.01
AudioLDM-S [51]	29.48	6.90
AudioLDM-L [51]	23.31	8.13
CoDi [78]	**22.90**	**8.77**
NExT-GPT	23.58	8.35

Table 4: Text-to-audio generation results on AudioCaps [38].

Method	FID (↓)	CLIPSIM (↑)
CogVideo [30]	23.59	0.2631
MakeVideo [74]	13.17	0.3049
Latent-VDM [68]	14.25	0.2756
Latent-Shift [2]	15.23	0.2773
CoDi [78]	—	0.2890
NExT-GPT	**13.04**	**0.3085**

Table 5: Text-to-video generation results (zero-shot) on MSR-VTT [92].

Method	B@4	METEOR	CIDEr
Oscar [46]	36.58	30.4	124.12
BLIP-2 [43]	43.7	—	145.8
OFA [86]	44.9	32.5	154.9
CoDi [78]	40.2	31.0	149.9
NExT-GPT	44.3	**32.9**	**156.7**

Table 6: Image-to-text generation (image captioning) results on COCO-caption data [50].

Method	SPIDEr	CIDEr
AudioCaps [38]	0.369	0.593
BART [26]	0.465	0.753
AL-MixGen [39]	0.466	0.755
CoDi [78]	0.480	0.789
NExT-GPT	**0.521**	**0.802**

Table 7: Audio-to-text generation (audio captioning) results on AudioCaps [38].

Method	B@4	METEOR
ORG-TRL [106]	43.6	28.8
GIT [85]	54.8	33.1
mPLUG-2 [91]	57.8	34.9
CoDi [78]	52.1	32.5
NExT-GPT	**58.4**	**38.5**

Table 8: Video-to-text generation (video captioning) results on MSR-VTT [92].

Method	Object		Background	
	CLIP (↑)	FID (↓)	CLIP (↑)	FID (↓)
PTP [29]	30.33	9.58	31.55	13.92
BLDM [4]	29.95	6.14	30.38	20.44
DiffEdit [14]	29.30	**3.78**	26.92	**1.74**
PFB-Diff [36]	**30.81**	5.93	**32.25**	13.77
NExT-GPT	29.31	6.52	27.29	15.20

Table 9: Text+image-to-image generation (text-conditioned image editing) results on COCO data [50].

Method	MCD (↓)
CampNet [87]	0.380
MakeAudio [33]	0.375
AudioLDM-L [51]	0.349
NExT-GPT	**0.302**

Table 10: Text+audio-to-audio generation (text-conditioned speech editing) results on VCTK data [83].

Method	CLIP-T (↑)	CLIP-I (↑)
CogVideo [30]	0.2391	0.9064
TuneVideo [89]	0.2758	0.9240
SDEdit [55]	0.2775	0.8731
Pix2Video [9]	**0.2891**	**0.9767**
NExT-GPT	0.2683	0.9645

Table 11: Text+video-to-video generation (text-conditioned video editing) results on DAVIS data [62].

Midjourney[8]. After human inspections and filtering of inappropriate instances, we obtain a total of 5K dialogues in low quality. In Table 2 we compare the existing multimodal IT datasets with our `MosIT` data.

6 Experiments

6.1 Any-to-any Multimodal Generation

We try to quantify the generation quality of NExT-GPT on certain benchmark datasets under some common settings, such as text-to-X generation, X-to-text generation, and Text-conditioned modality editing. We mimic the task by taking only one turn of interaction between the user and the model.

- **'Text' — 'X' Generation** represents the most frequent tasks of text-conditioned modal synthesis. Table 3, 4 and 5 present the comparisons between ours and some state-of-the-art systems. Overall NExT-GPT shows nice performance on par with the values from the best-performing baselines.

- **'X' — 'Text' Generation** represent the tasks of modal captioning. Table 6, 7 and 8 show the results on different tasks. Overall, we find that NExT-GPT can mostly achieve much better performance on the X-to-text generation than the CoDi baseline, owing to the direct generation of texts from LLM, which is inherently expertized by the LLM.

- **'Text+X' — 'X' Generation** represents a task category of text-conditioned modal editing. Table 9, 10 and 11 show the performances on different tasks. Compared with the above two types of tasks, NExT-GPT could be not that superior for the text-conditioned modal editing tasks. Yet, it still shows competitive performance.

---

[8] `https://www.midjourney.com/`

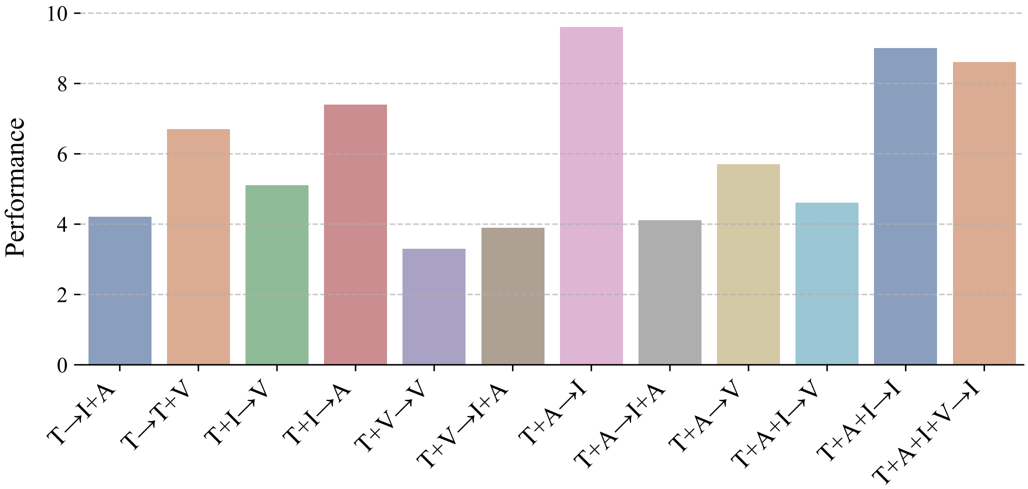


Figure 5: Comparative performance of NExT-GPT on various complex cross-modal conversions.

• **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 5 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.

6.2 Example Demonstrations

To demonstrate the effectiveness and potential of our proposed NExT-GPT in developing human-like conversational agents, here we further offer compelling examples that vividly illustrate the system's exceptional capacity to comprehend and reason contents across various modalities in any combination. Figure 6, 7, 8, 9, 10 and 11 show the examples from NExT-GPT. Go to the project page for more examples and access the dynamic video and audio contents.

7 Conclusion

In this work, we present 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, images, videos, 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 to future more potential modalities. To enable our NExT-GPT with complex cross-modal semantic understanding and content generation, we introduce 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.

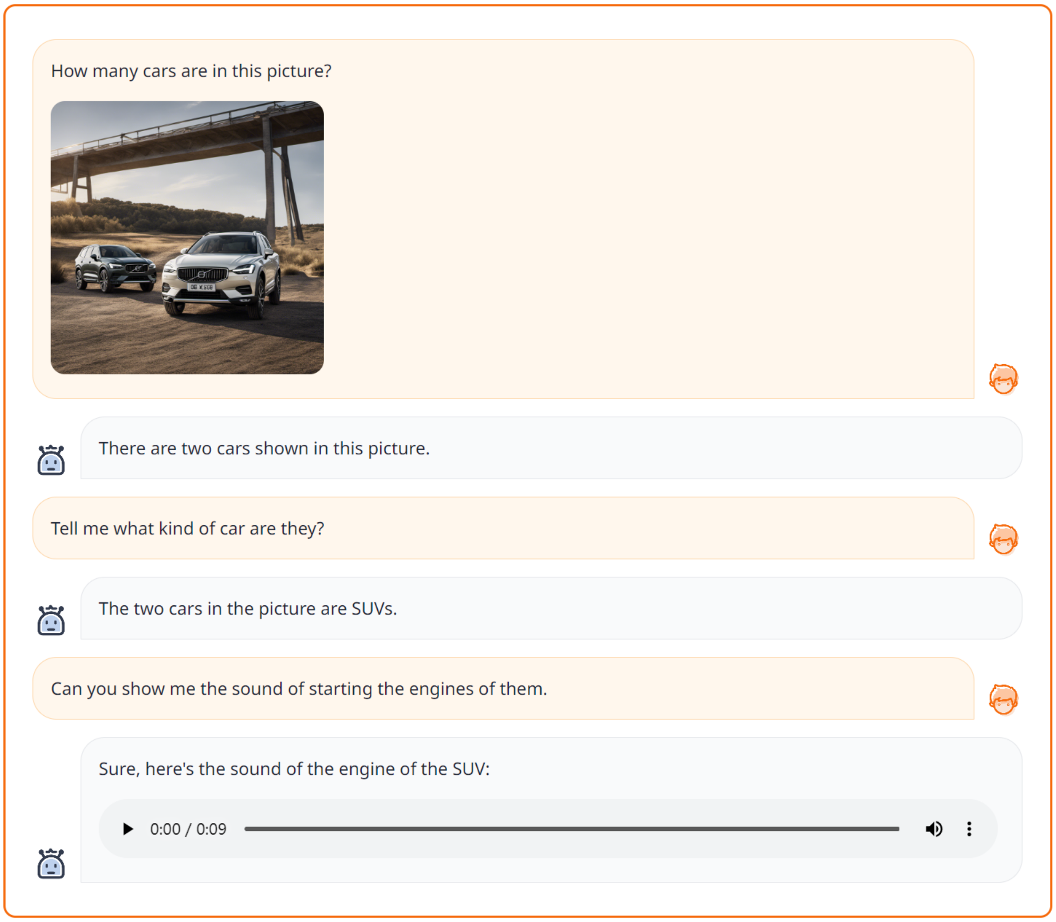
Limitation and Future work As future work, there are at least following four avenues to explore.

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 such that it becomes 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.



How many cars are in this picture?

There are two cars shown in this picture.

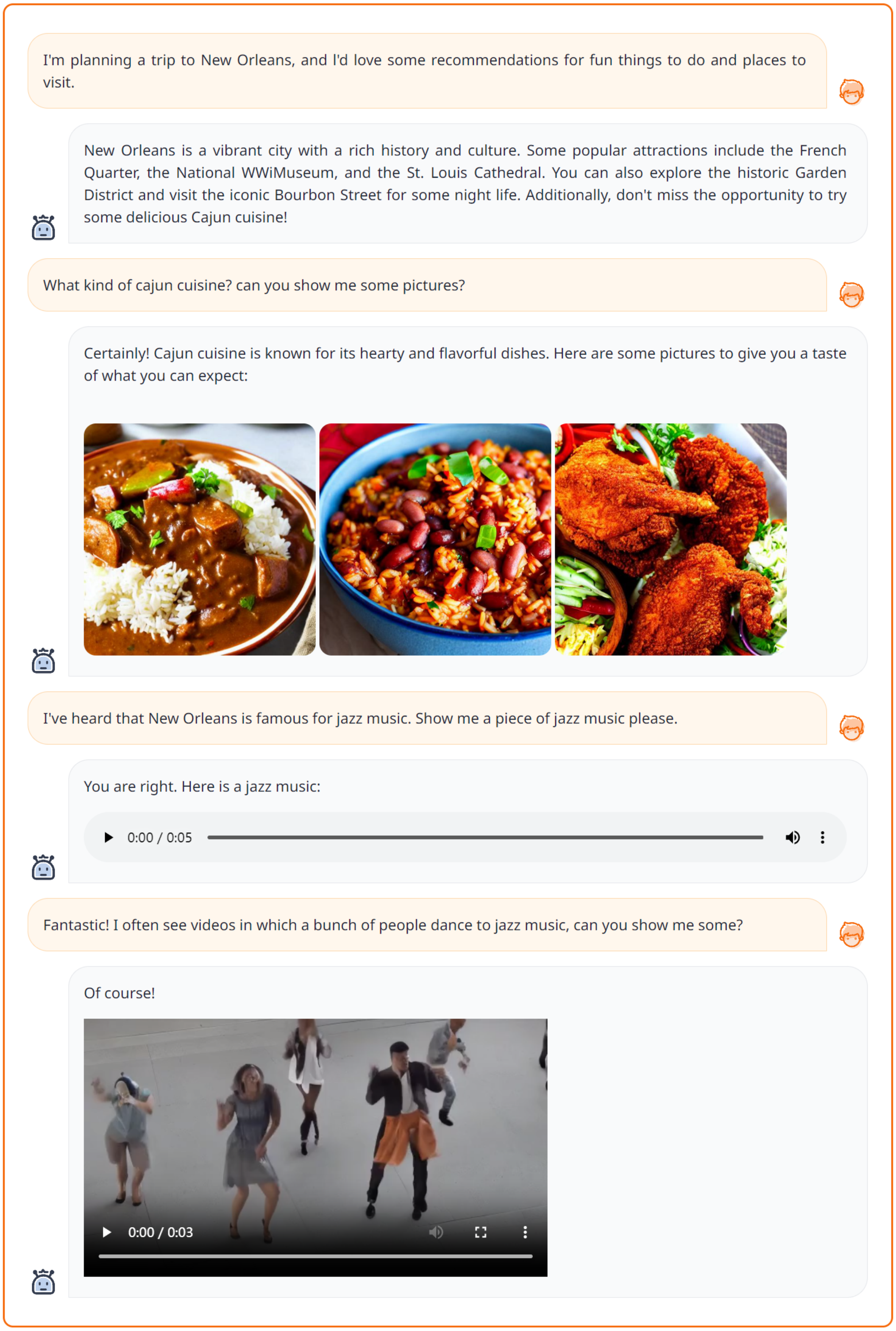
Tell me what kind of car are they?

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:

Figure 6: 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 WWiMuseum, 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:

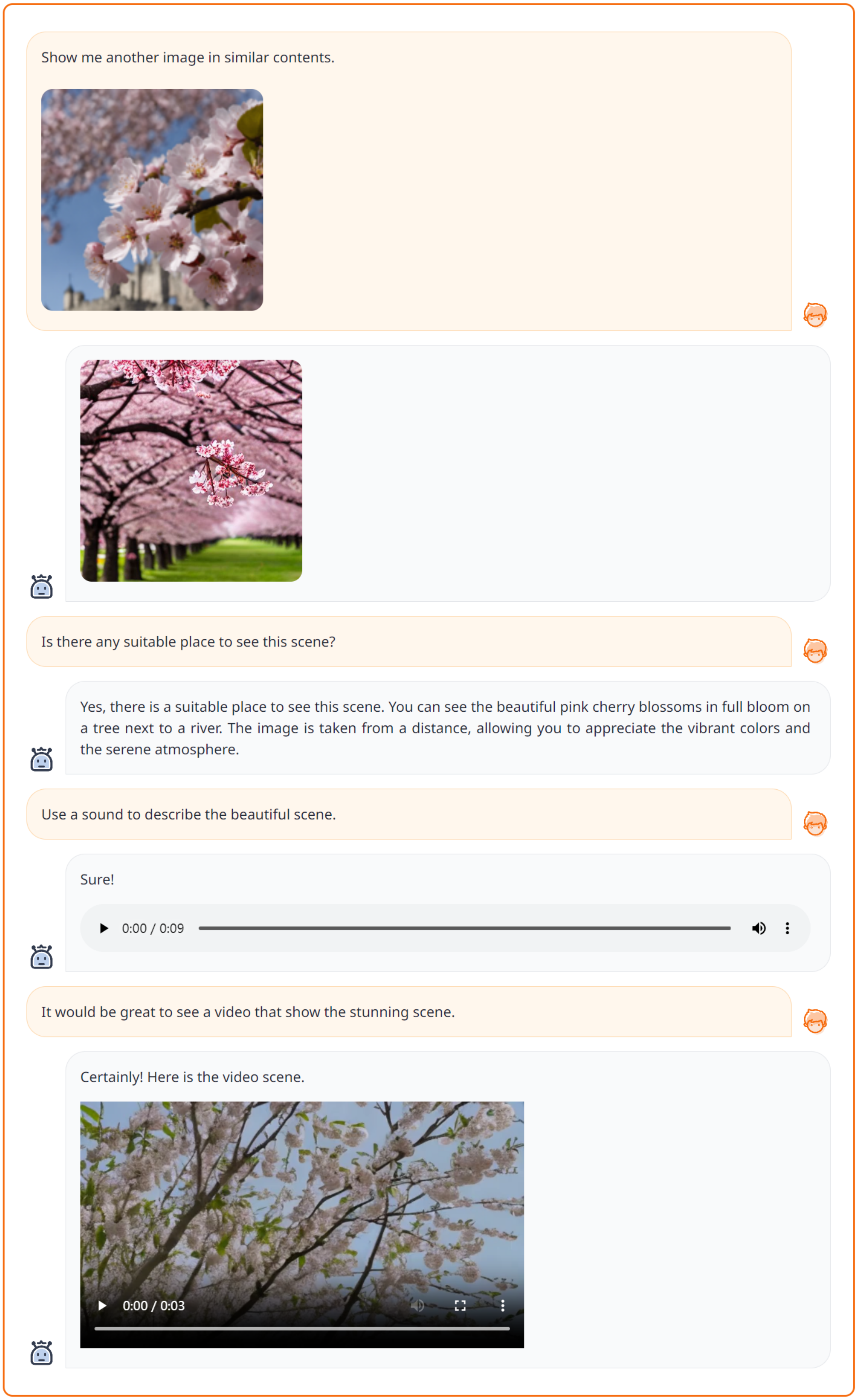
I've heard that New Orleans is famous for jazz music. Show me a piece of jazz music please.

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 7: 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.

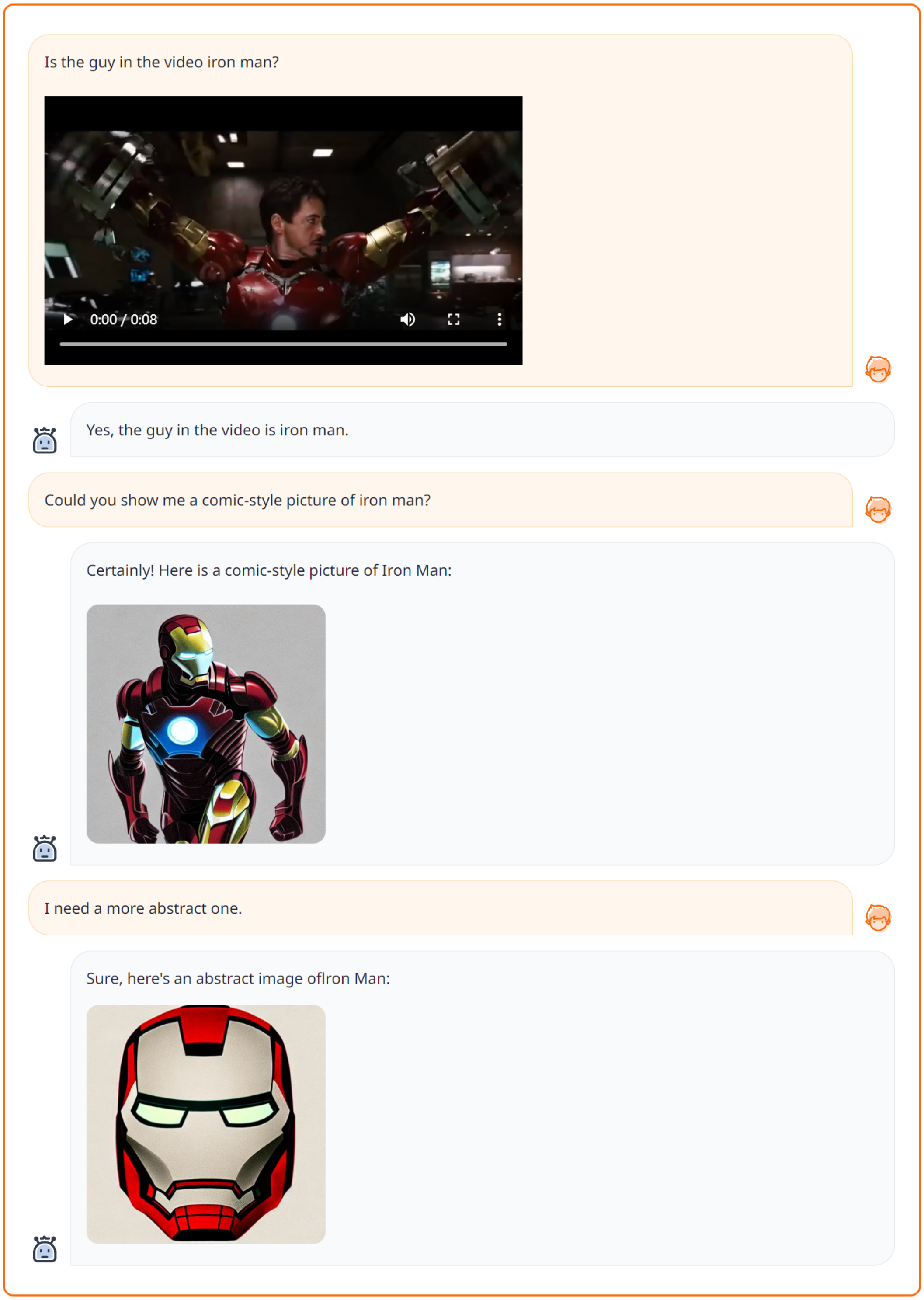
Use a sound to describe the beautiful scene.

Sure!

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

Certainly! Here is the video scene.

Figure 8: Example of Text+Image → Text+Image+Video+Audio.



Is the guy in the video iron man?

Yes, the guy in the video is iron man.

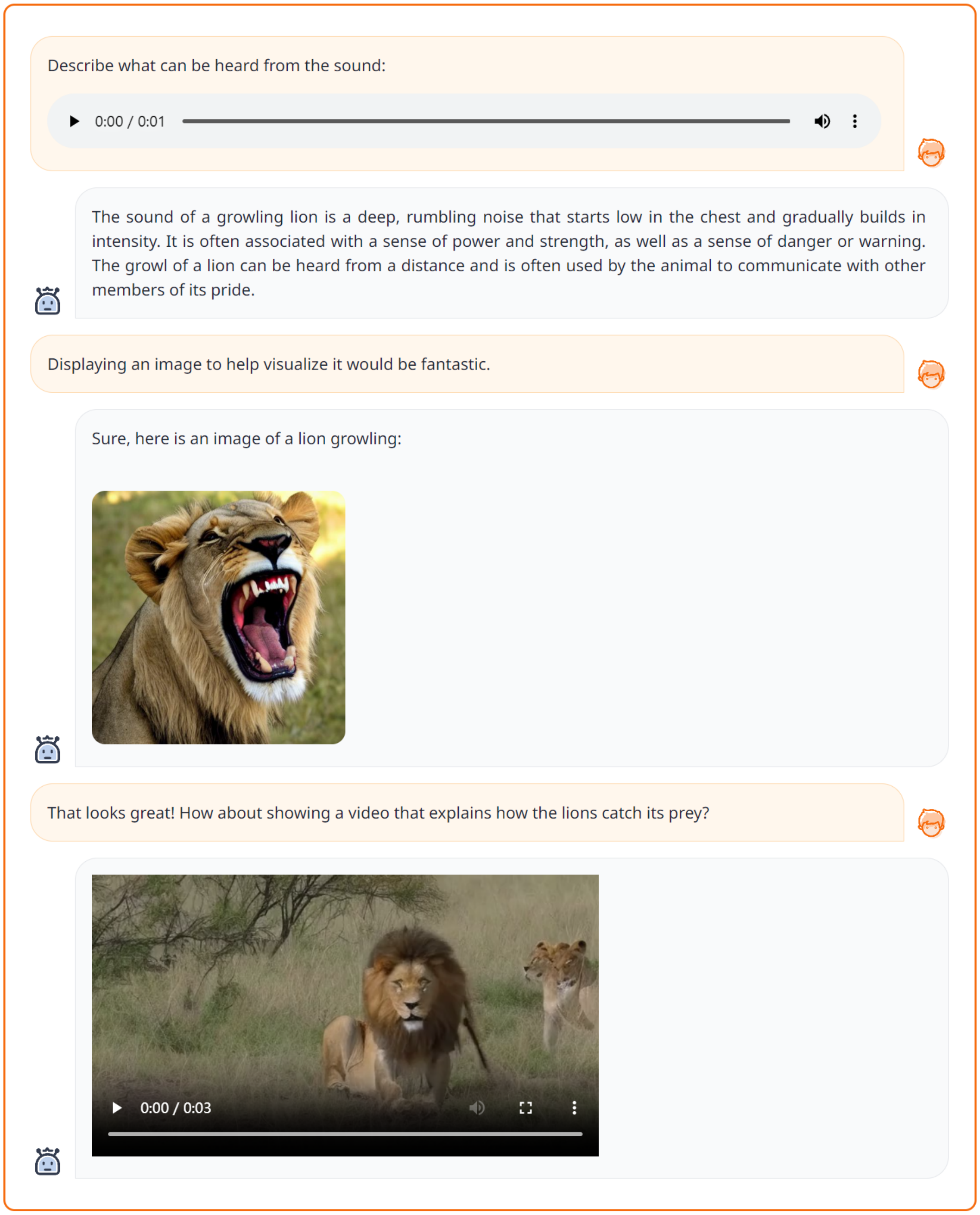
Could you show me a comic-style picture of iron man?

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

I need a more abstract one.

Sure, here's an abstract image ofIron Man:

Figure 9: Example of Text+Video → Text+Image.



Describe what can be heard from the sound:

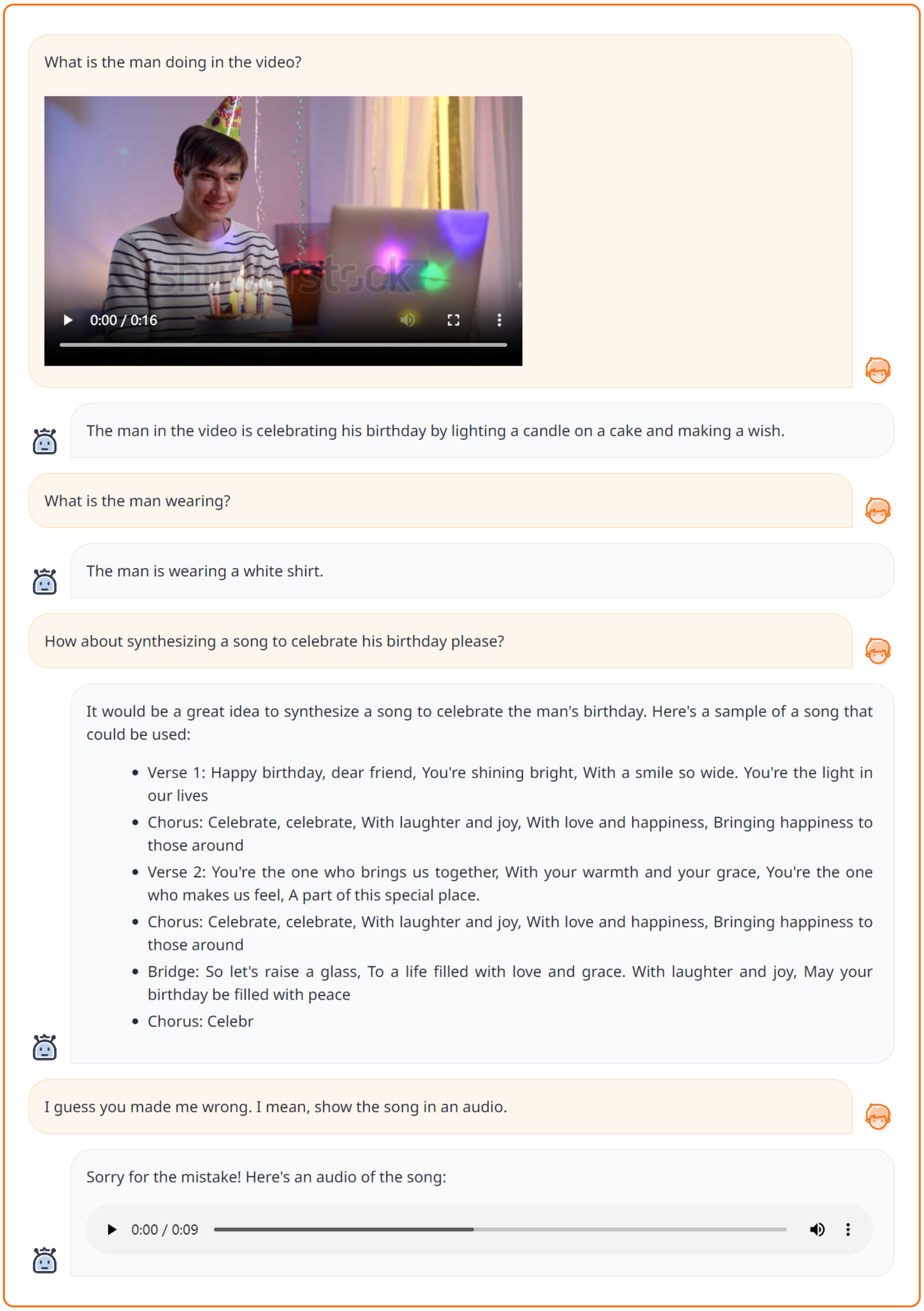
The sound of a growling lion is a deep, rumbling noise that starts low in the chest and gradually builds in intensity. It is often associated with a sense of power and strength, as well as a sense of danger or warning. The growl of a lion can be heard from a distance and is often used by the animal to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

Figure 10: 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: Celebr

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 11: Example of Text+Video → Text+Audio.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the NeurIPS*, 2022.

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

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the CVPR*, pages 6077–6086, 2018.

[4] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4): 149:1–149:11, 2023.

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the ICCV*, pages 1708–1718, 2021.

[6] Mohammad Bashiri, Edgar Y. Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas S. Tolias, and Fabian H. Sinz. A flow-based latent state generative model of neural population responses to natural images. In *Proceedings of the NeurIPS*, pages 15801–15815, 2021.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proceedings of the ICLR*, 2019.

[8] Cerspense. Zeroscope: Diffusion-based text-to-video synthesis. 2023. URL `https://huggingface.co/cerspense`.

[9] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. *CoRR*, abs/2303.12688, 2023.

[10] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the CVPR*, pages 3558–3568, 2021.

[11] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the CVPR*, pages 16867–16876, 2021.

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

[13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[14] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *Proceedings of the ICLR*, 2023.

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023.

[16] Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonirina, Francesca Franzon, and Marco Baroni. Cross-domain image captioning with discriminative finetuning. In *Proceedings of the CVPR*, pages 6935–6944, 2023.

[17] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Proceedings of the NeurIPS*, pages 19822–19835, 2021.

[18] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Björn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the CVPR*, pages 3742–3753, 2021.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the ICLR*, 2021.

[20] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In *Proceedings of the ICASSP*, pages 736–740, 2020.

[21] Wan-Cyuan Fan, Yen-Chun Chen, Dongdong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *CoRR*, abs/2208.13753, 2022.

[22] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *CoRR*, abs/2212.05032, 2022.

[23] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *CoRR*, abs/2208.01618, 2022.

[24] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *Proceedings of the ECCV*, pages 102–118, 2022.

[25] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *CoRR*, abs/2305.05665, 2023.

[26] Félix Gontier, Romain Serizel, and Christophe Cerisara. Automated audio captioning by fine-tuning BART with audioset tags. In *Proceedings of the DCASE*, pages 170–174, 2021.

[27] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the CVPR*, pages 18941–18951, 2023.

[28] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the CVPR*, pages 961–970, 2015.

[29] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *Proceedings of the ICLR*, 2023.

[30] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *CoRR*, abs/2205.15868, 2022.

[31] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *CoRR*, 2021.

[32] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR*, 2022.

[33] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the ICML*, pages 13916–13932, 2023.

[34] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head. *CoRR*, abs/2304.12995, 2023.

[35] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *CoRR*, abs/2302.14045, 2023.

[36] Wenjing Huang, Shikui Tu, and Lei Xu. Pfb-diff: Progressive feature blending diffusion for text-driven image editing. *CoRR*, abs/2306.16894, 2023.

[37] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *CoRR*, abs/2304.06025, 2023.

[38] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the NAACL*, pages 119–132, 2019.

[39] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, Jaeheon Sim, Jinwoo Lee, and Kyogu Lee. Improving audio-language learning with mixgen and multi-level test-time augmentation. *CoRR*, abs/2210.17143, 2022.

[40] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *CoRR*, abs/2305.17216, 2023.

[41] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[42] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425, 2023.

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, pages 19730–19742, 2023.

[44] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023.

[45] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387, 2023.

[46] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the ECCV*, pages 121–137, 2020.

[47] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *CoRR*, abs/2308.10253, 2023.

[48] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the CVPR*, pages 2918–2927, 2022.

[49] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the CVPR*, pages 17928–17937, 2022.

[50] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of the ECCV*, pages 740–755, 2014.

[51] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *Proceedings of the ICML*, pages 21450–21474, 2023.

[52] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023.

[53] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the CVPR*, pages 14274–14283, 2020.

[54] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424, 2023.

[55] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *Proceedings of the ICLR*, 2022.

[56] Victor Siemen Janusz Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? In *Proceedings of the AACL*, pages 504–515, 2020.

[57] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[58] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the ICML*, pages 16784–16804, 2022.

[59] OpenAI. Introducing chatgpt. 2022.

[60] OpenAI. Gpt-4 technical report. 2022.

[61] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the NeurIPS*, 2022.

[62] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the CVPR*, pages 724–732, 2016.

[63] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023.

[64] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from LLM for text-to-image generation. *CoRR*, abs/2308.05095, 2023.

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the ICML*, pages 8748–8763, 2021.

[66] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the ICML*, pages 8821–8831, 2021.

[67] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Proceedings of the NeurIPS*, pages 14837–14847, 2019.

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the CVPR*, pages 10674–10685, 2022.

[69] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CoRR*, abs/2208.12242, 2022.

[70] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Proceedings of the NeurIPS*, 2022.

[71] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the ACL*, pages 2556–2565, 2018.

[72] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.

[73] Hisaichi Shibata, Shouhei Hanaoka, Yang Cao, Masatoshi Yoshikawa, Tomomi Takenaga, Yukihiro Nomura, Naoto Hayashi, and Osamu Abe. Local differential privacy image generation using flow-based deep generative models. *CoRR*, abs/2212.10688, 2022.

[74] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *CoRR*, abs/2209.14792, 2022.

[75] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Proceedings of the NeurIPS*, 2020.

[76] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *CoRR*, abs/2205.02655, 2022.

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

[78] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *CoRR*, abs/2305.11846, 2023.

[79] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. 2023. URL `https://github.com/tatsu-lab/stanford_alpaca`.

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

[81] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Proceedings of the NeurIPS*, 2020.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the NeurIPS*, pages 5998–6008, 2017.

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

[84] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: extended textual conditioning in text-to-image generation. *CoRR*, abs/2303.09522, 2023.

[85] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.*, 2022, 2022.

[86] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the ICML*, volume 162, 2022.

[87] Tao Wang, Jiangyan Yi, Ruibo Fu, Jianhua Tao, and Zhengqi Wen. Campnet: Context-aware mask prediction for end-to-end text-based speech editing. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 2241–2254, 2022.

[88] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.

[89] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *CoRR*, abs/2212.11565, 2022.

[90] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI*, pages 2804–2812, 2022.

[91] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In *Proceedings of the ICML*, pages 38728–38748, 2023.

[92] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*, pages 5288–5296, 2016.

[93] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the CVPR*, pages 1316–1324, 2018.

[94] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the ICCV*, pages 1666–1677, 2021.

[95] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:1720–1733, 2023.

[96] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499, 2023.

[97] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687, 2023.

[98] Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. Unified language representation for question answering over text, tables, and images. *CoRR*, abs/2306.16762, 2023.

[99] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *CoRR*, abs/2206.10789, 2022.

[100] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the CVPR*, pages 23465–23476, 2023.

[101] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *CoRR*, abs/2305.01278, 2023.

[102] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based GAN for high-resolution image generation. In *Proceedings of the CVPR*, pages 11294–11304, 2022.

[103] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *CoRR*, abs/2305.11000, 2023.

[104] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858, 2023.

[105] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*, abs/2306.17107, 2023.

[106] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the CVPR*, pages 13275–13285, 2020.

[107] Bo Zhao, Boya Wu, and Tiejun Huang. SVIT: scaling up visual instruction tuning. *CoRR*, abs/2307.04087, 2023.

[108] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, and Xiangyu Zhang. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *CoRR*, abs/2307.09474, 2023.

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

[110] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.

[111] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the CVPR*, pages 5802–5810, 2019.