CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation

Zineng Tang[1,4*] Ziyi Yang[2†] Mahmoud Khademi[2] Yang Liu[2] Chenguang Zhu[3‡] Mohit Bansal[4†]

[1]UC Berkeley [2]Microsoft Azure AI [3]Zoom [4]UNC Chapel Hill

https://codi-2.github.io

Abstract

We present CoDi-2, a Multimodal Large Language Model (MLLM) for learning in-context interleaved multimodal representations. By aligning modalities with language for both encoding and generation, CoDi-2 empowers Large Language Models (LLMs) to understand modality-interleaved instructions and in-context examples and autoregressively generate grounded and coherent multimodal outputs in an any-to-any input-output modality paradigm. To train CoDi-2, we build a large-scale generation dataset encompassing in-context multimodal instructions across text, vision, and audio. CoDi-2 demonstrates a wide range of zero-shot and few-shot capabilities for tasks like editing, exemplar learning, composition, reasoning, etc. CoDi-2 surpasses previous domain-specific models on tasks such as subject-driven image generation, vision transformation, and audio editing and showcases a significant advancement for integrating diverse multimodal tasks with sequential generation.

1. Introduction

Multimodal generation has achieved remarkable advancements in recent years, e.g., generating high-fidelity image, video, audio and music samples from prompt provided by users. Recent advancements in AI-Generated Content (AIGC) highlight in-context generation [24, 38], concept learning [28], editing [2], and fine-grained control [46]. Recently, Tang et al. [32] proposed CoDi, the first model ever that can generate any combinations of modalities from any combinations of input ones. Building upon this foundational work, the subsequent study by [40] further advances CoDi by proposing a model that can facilitates conversational abilities and expansion to additional modalities.

Although remarkable advances have been made in multimodal generation, several critical challenges remain: (1) Zero-shot fine-grained and complex user-control of multimodal generation is difficult: current multimodal generative models (MGM) cannot follow in-context generation examples without finetuning on subtasks, such as replicating or transferring an editing effect via an 'analogy' setting or subject driven generation, as demonstrated in the prompt (as in the row "Exemplar Learning" and "'Subject Driven' of Table 3). Moreover, the reasoning ability of MGM is rather limited, e.g., the input prompts are usually descriptive where the generation do not require capabilities such as logical, compositional, and analytical intelligence. (2) The inputs in previous MGMs mostly only contain one or two modalities. The ability to understand modality-interleaved inputs, such as language instruction mixing with contextual visual and auditory inputs is critical to building a fundamental multimodality model. Hence, overall, a versatile any-to-any MGM, that can follow interleaved in-context multimodal instructions and interactive multi-round chatting is strongly needed. (3) The user-and-model interaction is usually constrained to single-round, or it is challenging for current models to follow multi-round instructions while ensuring the consistency and faithfulness of responses across the rounds, as shown in Figure 1.

To this end, we propose CoDi-2, a versatile Multimodal Large Language Model (MLLM) that understand in-context examples, follow modality-interleaved instructions, perform multi-round chatting and editing. Enabling in-context learning and following interleaved multimodal instructions in multimodal generation is challenging. In previous multimodal generative models, the backbone is mostly diffusion models (DMs) which are good at generation but intrinsically lack the capability to perform in-context understanding [39]. We therefore propose to harness a Large Language Model (LLM) as the "brain" to understand modality-interleaved human instructions and in-context examples, and output multimodal signals. LLMs have strong language reasoning capabilities for complex instructions in the lan-

[*] Work done while at Microsoft internship and UNC.
[†] Correspondence: ziyiyang@microsoft.com, mbansal@cs.unc.edu
[‡] Work done while at Microsoft.

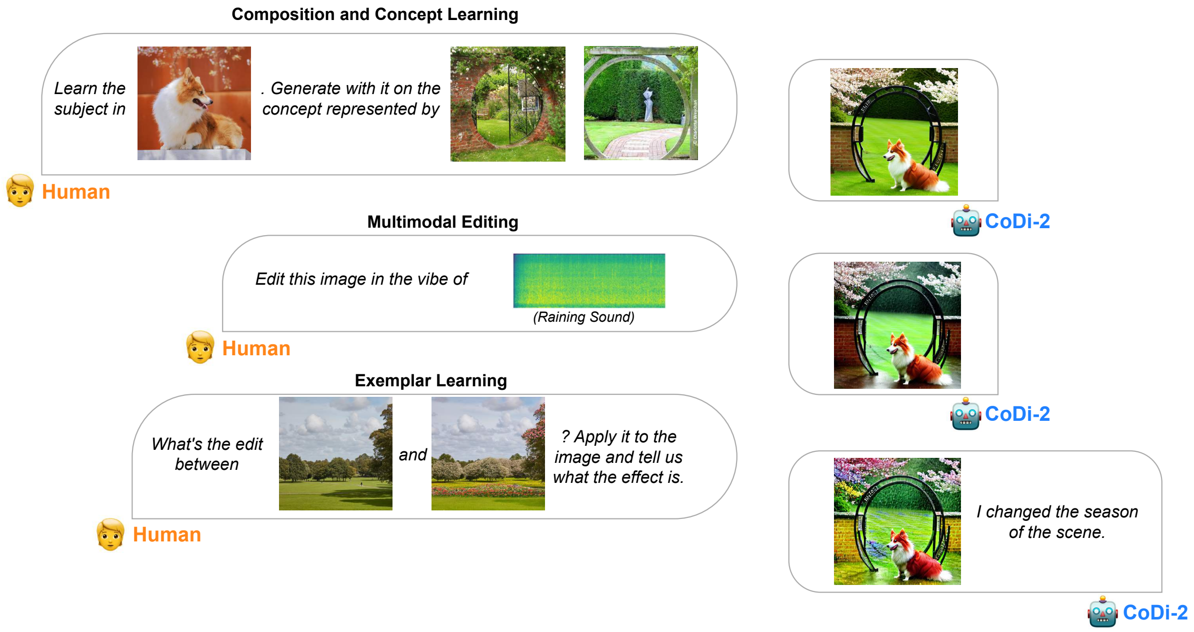

Composition and Concept Learning
Learn the subject in
. Generate with it on the concept represented by
Human
CoDi-2
Multimodal Editing
Edit this image in the vibe of
(Raining Sound)
Human
CoDi-2
Exemplar Learning
What's the edit between
and
? Apply it to the image and tell us what the effect is.
Human
I changed the season of the scene.
CoDi-2

Figure 1. Multi-round conversation between humans and CoDi-2 offering in-context multimodal instructions for image editing.

guage domain. By mapping all modalities to the space of language (as proposed in CoDi [32]) and connecting these modalities to LLM through encoder and synchronized decoders, CoDi-2 can process multimodal inputs within their context by aligning vision or audio features to the language model input and output space, and understand the delicate modality-interleaved instructions for zero-shot or few-shot generation.

For generation, we propose to train the MLLM to autogressively predict the features of the output modality. The predicted features are then input to (synchronized) diffusion models. This end-to-end any-to-any generative framework enables CoDi-2 to conduct elaborate reasoning for understanding and generate multiple modalities, and therefore allows diverse tasks such imitation, editing, compositional creation, etc. In training CoDi-2, the gradient obtained from the diffusion models' generation loss also directly backpropagates to the LLM which can enhance the perceptual faithfulness to the inputs including images or audio.

The development of the alignment data to train such model is also challenging, hindered by the scarcity of specialized data such as multimodal reasoning or in-context learning. To start with, we comprehensively collect latest instructional generation datasets across vision, audio, and language. We then propose to convert these instructional datasets to in-context generation ones, such that in the prompt (e.g., row "Exemplar Learning" of Table 1) and more can be referred in Tables 1 to 3. To further diversify the in-context learning datasets, we propose a novel method to build text-only datasets for multimodal in-context learning. Since language and other modalities (vision and audio) are mapped to the same space through the aligned encoders, we can flexibly build multimodal datasets with only language, where the multimodal components are represented by their respective textual descriptions (e.g., using image caption instead of the pixels to represent the image).

Empirical assessments of our multimodal generation tasks, which include a diverse array of complex and intertwined instructions, yield remarkable results. These tasks encompass audio fusion and editing, image generation with intricate composition, the use of in-context exemplars, and sophisticated reasoning, as well as understanding and generating videos. This wide range of tasks show strong capability in both zero-shot and few-shot prompting settings, showcasing our system's adaptability and robust performance across different scenarios.

2. Related Work

2.1. Multimodal Large Language Models

Recent years have witnessed the rapid evolution of LLMs, setting a new precedent in natural language understanding and generation [23, 34, 35]. Multimodal LLMs extend LLMs to multimodal learning [42, 45], enabling the processing of diverse input forms, not just limited to text but also incorporating visual and other sensory data [7, 16, 18, 20, 43, 44]. The innovation in this space has led to models that are not only capable of understanding multimodal inputs but also adept at generating multifaceted outputs, thereby pushing the boundaries of creative and contextual AI-generated content [32, 40]. Another notable line of work is using LLM to ground image generation [13, 31].

2.2. Multimodal In-Context

Multimodal in-context requires sometimes interleaved in-context understanding of multimodal inputs like images and text like Wikipedia (with images), documents, videos with narrations or QAs, etc. This domain has expanded yet facing its set of challenges. While there is a plethora of research focusing on the understanding aspect of multimodal data [16, 49], the generation of raw sensory perceptions such as images or audio remains a complex hurdle. The concept of treating images as a foreign language opened new avenues, particularly in in-context image generation [25]. However, these pioneering techniques are still in nascent stages, often constrained by their training regimes and lacking genuine in-context learning capabilities, which limits their performance and adaptability.

2.3. Multimodal Generation

Recent years have witnessed a significant growth in image editing and manipulation research, which can be broken into image editing [2, 21], exemplar learning [38] for image generation, image composition [14, 25, 28], and concept learning [11] from images.

Image editing [21] uses guidance control and edit the attributes of an image. To align the guidance with human instructions, InstructPix2Pix [2] takes in instructional image editing prompts to directly transform an image. The realm of image composition are tasks that compose one or more images into a single image and demand high fidelity to input images, which poses unique challenges. Techniques involved in subject-driven image generation [28] have shown promise in transforming a subject into a new scene. However, they often necessitate task-specific or subject-specific tuning. This specialization often confines the models within the boundaries of their training data, impeding their ability to generalize beyond learned tasks or subjects. Kosmos-G [25] furthers the efforts for zero-shot image generation with in-context interleaved image and text. But its efforts is limited to image composition. Lastly, learning visual concepts and apply them in image generation is also a growing direction [11, 14]. For example, multi-concept customization to text-to-image generation [14] requires the model to extract visual concept like a moon gate or a certain subject and apply them in image generation. The aspiration to develop a model with in-context multimodal reasoning abilities to transcend these limitations inspires our versatile framework that takes in task instructions and perform in-context zero-shot generation.

3. Model Architecture

CoDi-2 is designed to process in-context multimodal inputs, including text, images, and audio, utilizing specific instructions to facilitate in-context learning and generate corresponding text, images, or audio outputs. The model is distinguished by the several key features as introduced in following subsections.

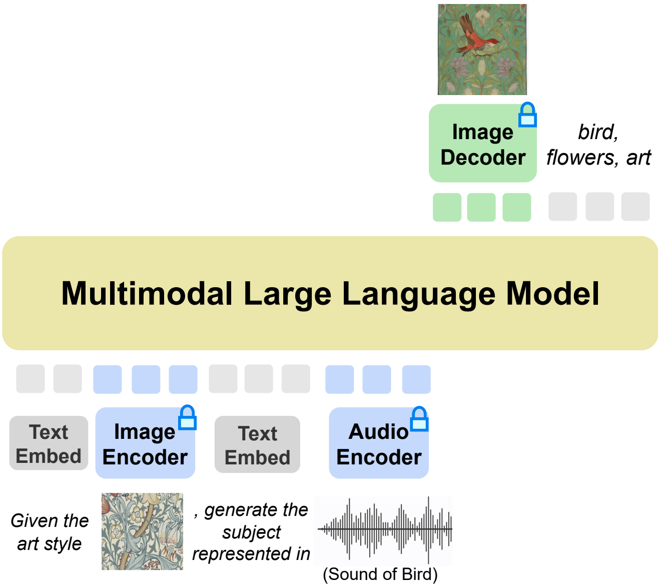


Figure 2. Model Architecture: CoDi-2 comprises a multimodal large language model that encompasses encoder and decoder for both audio and vision inputs, as well as a large language model. This architecture facilitates the decoding of image or audio inputs using diffusion models. In the training phase, our approach employs pixel loss obtained from the diffusion models alongside token loss, adhering to the standard causal generation loss.

3.1. Multimodal LLM as the Fundamental Engine

Building such an any-to-any foundation model that can digest interleaved inputs of modalities, understand and reason over complex instructions (e.g., multi-round conversation, in-context examples), and interact with multimodal diffusers requires a powerful fundamental "engine". We propose to leverage MLLM for this engine, which is built by empowering a text-only LLM with multimodal perceptions.

The motivation of harnessing LLM is intuitively inspired by the observation that LLMs exhibit exceptional ability such as chatting, zero-shot learning, instruction-following, etc, in language-only domain [48]. By leveraging projections from aligned multimodal encoders (e.g., [32]), we can seamlessly empower the LLM to perceive modality-interleaved input sequence. Specifically, in processing the multimodal input sequence, we first use the multimodal encoder to project the multimodal data into a feature sequence. Special tokens are prepended and appended to the features sequence, e.g. "⟨audio⟩ [audio feature sequence] ⟨/audio⟩". By such for instance, a modality-interleaved input sequence "*A cat sitting on* [image0:an image of a couch] *is making* *the sound of* `[audio0:audio of cat purring]`" is then transformed to "*A cat sitting on* ⟨`image`⟩ `[image feature sequence]` ⟨`/image`⟩ *is making the sound of* ⟨`audio`⟩ `[audio feature sequence]` ⟨`/audio`⟩, before inputting to the MLLM to process and generation.

3.2. Multimodal Generation with MLLM

To generate text, the MLLM can naturally generate text tokens autoregressively; for multimodal generation, one common way in previous works was to transform the multimodal target (e.g., the ground-truth image) into discrete tokens such that they can be generated autoregressively like text. However, the generation quality of this methodology is intrinsically constrained by the VAE-like generation decoder, while current SOTA multimodal generation frameworks generally adopt Diffusion Models (DMs) [27]. Therefore, we propose to integrate DMs into MLLM to generate multimodal outputs, following nuanced modality-interleaved instructions and prompts. Recall the training objective of a diffusion model is given as:

$$\mathcal{L}_{DM} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon},t}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(\boldsymbol{z}_t, t, C_y(\boldsymbol{y})\big)\|_2^2, \tag{1}$$

where $\boldsymbol{y}$ is the conditional modality, C_y is the conditional encoder for $\boldsymbol{y}$, $\boldsymbol{\epsilon}_\theta$ is the U-Net, and $\boldsymbol{z}_t$ is the noisy latent variable at time step t.

We propose to train the MLLM to generate the conditional feature $\boldsymbol{c} = C_y(\boldsymbol{y})$ that will be fed into DM to synthesize the target output $\boldsymbol{x}$. By such, the generative loss of DM can be used to train MLLM. To further provide a stronger and directer supervision signal for MLLM, and to retain the perceptual characteristics inherent in the original input, we explicitly induce that $\boldsymbol{c} = C_x(\boldsymbol{x})$, i.e., MLLM is trained to generate $C_x(\boldsymbol{x})$ and DM is expected to function as an autoencoder in this case[1]. The mean squared error between MLLM output feature $\boldsymbol{c}_{MLLM}$ and $C_x(\boldsymbol{x})$, together with $\mathcal{L}_{DM}$, and text token prediction loss $\mathcal{L}_t$ is the final training loss: $\mathcal{L} = \alpha\text{MSE}\big(\boldsymbol{c}_{MLLM}, C_x(\boldsymbol{x})\big) + \mathcal{L}_{DM} + \mathcal{L}_t$ controlled by weighting α.

4. Building Diverse Multimodal In-Context Generation Data

4.1. Dataset Construction

We construct and employ a variety of datasets to facilitate interleaved and in-context multimodal generation, enriching the capabilities of CoDi-2.

Multimodal In-Context Learning Datasets. Our approach leverages the strength of multimodal in-context understanding, and to bolster this aspect, we integrate MIMIC-IT [16] into our tasks. MIMIC-IT offers an extensive and diverse dataset comprising 2.8 million instruction-response pairs, specifically designed to elevate the performance of Vision-Language Models (VLMs) in real-world scenarios. This augmentation equips VLMs with abilities in perception, reasoning, and planning. Despite its output is text-only, it can help model's in-context understanding of multimodal inputs and overall instruction following. For example in perceptual understanding, given two images with only subtle differences, the instruction is to spot the different. By another example for reasoning, given video frames of football, the instruction is to predict what will happen next.

Multimodal Paired Datasets. Paired datasets like image-text are natural multimodal data for cross-modal generation. We use LAION-400M [29] that consists of 400 million image-text pairs filtered using CLIP. For audio paired dataset, we use AudioSet [8]. AudioSet offers a comprehensive ontology of 632 audio event classes. It also boasts a collection of 2,084,320 human-labeled 10-second sound clips sourced from YouTube videos. For video paired dataset, we use Webvid [1], featuring 10.7 million short video-caption pairs, totaling 52,000 hours, gathered from stock footage websites, showcasing diverse content. We construct two tasks with these datasets, 1) instructing to generate caption given an image or audio, and 2) instructing to generate the image or audio from caption.

Instructional Editing Datasets. Instructional Editing is a task structured as an input image, an editing instruction, and the resulting edited image. We use Instructpix2pix [2] for image instructional editing. For audio editing dataset, our approach is built on top of AudioSet [8]. We develop instructional editing versions of it, taking cues from AUDIT [37]. We have developed three versions of this dataset: audio addition (overlay), removal, and replacement, resulting in a dataset three times the size of AudioSet. By overlaying two distinct audio segments, `a, b`, we obtain a new combined audio `a+b`. This combined audio can also serve as an input for audio removal `a+b`→ `a`, or removing audio `b` from audio `a+b`. Audio replacement is constructed by integrating two different audio segments `b, c` into the same base audio `a`, and then we get `a+b`→ `a+c`, or replacing `b` with `c` for audio `a+b` to get `a+c`.

Constructed In-context Multimodal Generation Datasets. To further stimulate the multimodal in-context ability, we construct several in-context datasets for multimodal generation. *InstructPix2Pix* can be extended to interleaved in-context multimodal format, given its multiple image pairs corresponding to the same editing prompt.

[1] We leverage the condition encoder aligned across modalities from CoDi (Tang et al. [32], named prompt encoder in the original paper)

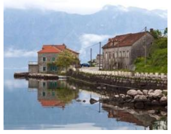
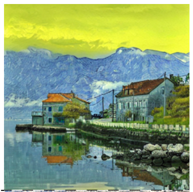
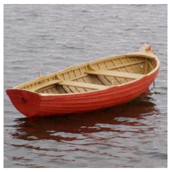
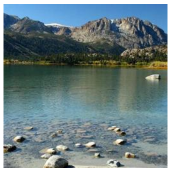
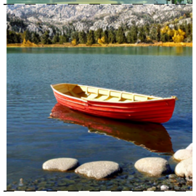
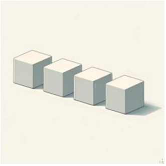
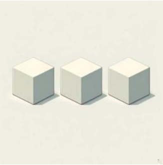
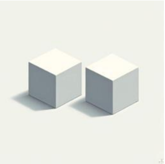
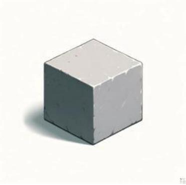
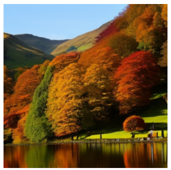
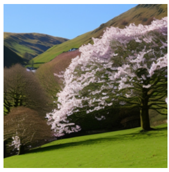
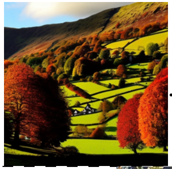
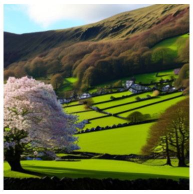
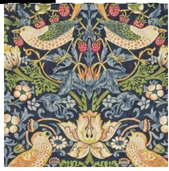
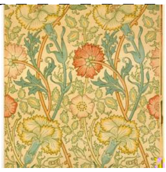
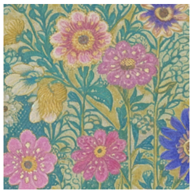
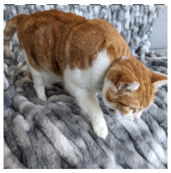
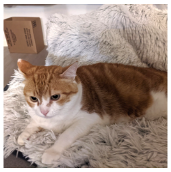
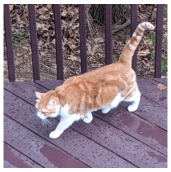
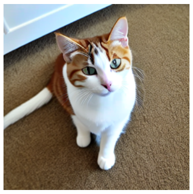
Task Type	Example Prompt	Output
	Zero-Shot Prompting	
Instruction Editing	Turn this  into Van Gogh style	
Composition	A  on  .	
Reasoning	Given   and  , what happens next?	
	One-Shot/Few-Shot Prompting	
Exemplar Learning	We apply a new concept to  and got . Apply the same concept to .	
Concept Learning	Given the artistic style represented in  and  , create a new artwork similar to it.	
Subject Driven	Given a set of pictures portraying your neighbor's cat   and  , create a new image of this cat.	

Table 1. Zero-shot, one-shot, and few-shot image generation examples by CoDi-2.

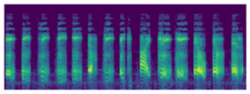
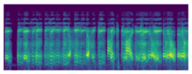
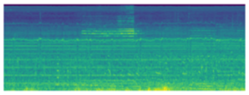
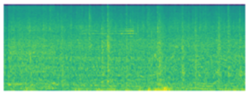
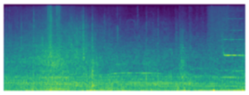
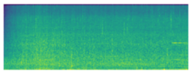
Task Type	Example Prompt	Output
	Zero-Shot Prompting	
Instruction Editing	Add echoing to this audio  *(person speaking)*.	 *(person speaks in echoes)*
	One-Shot/Few-Shot Prompting	
Exemplar Learning	We overlaid a new sound to  *(street noise)* and got  *(street noise, raining)*. Apply this same new sound to  *(train noise)*.	 *(train noise, raining)*

Table 2. Audio generation examples of CoDi-2.

Consequently, we define the in-context learning template as: **Input:** "*Given the transformation between* `[image0]` *and* `[image1]`*, apply the same editing to* `[image2]`*.*" **Target:** `[image3]`. This approach can also be adapted for audio editing datasets using the same template structure.

In addition, we utilize the *Kosmos-G* dataset [25], constructed using *Open Images V7* [15] with 9M images for image composition. Here, entities from captions are extracted to produce image segmentation for each identity. For instance, for descriptions like 'A cat on a couch', we obtain: **Input:** `[image0]` on `[image1]` **Target:** `[image2]`, where `[image0]` and `[image1]` represent the segmented cat and couch images derived from `[image2]`, respectively.

Text-Only Datasets Repurposed for Interleaved Multimodal In-Context. We propose to employ text-only datasets for enhancing generation with multimodal reasoning. Since the encoder features for all modalities are aligned, replacing text tokens with text encoder features can enhance interleaved multimodal understanding. Concretely, we randomly select phrases or words in the sentence and encode them with text feature encoder to swap out the original text embeddings. This innovative strategy bolsters the model's proficiency in understanding complex, interleaved multimodal scenarios by aligning the encoder features and the original language model text embeddings. We employ instructional dataset alpaca [33]. For example, we convert " *A cat typically has a compact, flexible body, covered in soft fur that can come in a variety of colors and patterns.*" to "`[text0]` *typically has* `[text1]`*, covered in* `[text2]` *that can come in a variety of colors and patterns.*" where `[text0]` `[text1]` `[text2]` are respectively text features from "*a cat*", "*a compact, flexible body*", and "*soft fur*".

4.2. In-Context Instruction Task Types

Tables 1 to 3 offer a comprehensive overview of the task types utilized in in-context multimodal generation. Each task type presents a unique approach to prompting models to generate or transform in-context multimodal content, including images, audio, and combinations thereof.

Zero-Shot Prompting. Zero-shot prompting tasks require the model to reason and generate new content without any prior examples. For instance, in Table 1, model transforms an image to match Van Gogh's style or compositing two separate images to form a coherent scene exemplifies the model's capacity to understand and apply complex instructions directly. Model also can perform reasoning and predict the next image in a sequence, which is one cube in consistent style. Table 2 shows adding echoes to an audio. Table 3 shows visual editing with sound vibe and frame+sound prediction of a video sequence.

One-Shot/Few-Shot Prompting. One-shot or few-shot prompting provides the model with one or a few examples to learn from before performing a similar task. This method is evident in tasks where the model adapts a learned concept from one image to another or creates a new piece of artwork by understanding the styles depicted in provided exemplars.

Exemplar learning is a subset of few-shot prompting where the model is explicitly shown an example of the desired output before being asked to apply this learning to a

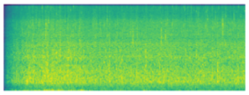
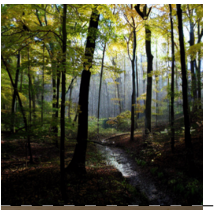
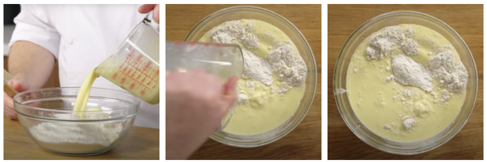
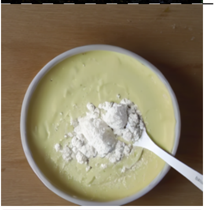
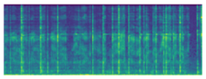
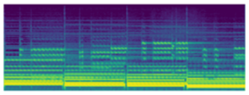
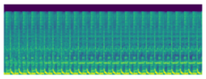
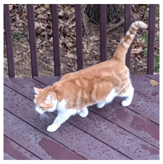
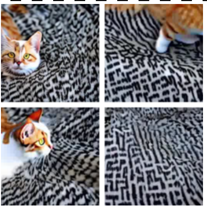
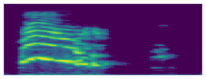
Task Type	Example Prompt	Output
	Zero-Shot Prompting	
Instruction Editing	 *(raining)* took place in .	
Reasoning	Given video frames, , what will happen next? Generate the sound and image for it.	  *(stirring the batter)*
	One-Shot/Few-Shot Prompting	
Exemplar Learning	For image , music  *(soft jazz music)* captures its vibe. What's the right music to ?	 *(dance music)*
Subject Driven	Given a set of pictures portraying your neighbor's cat and , create a video and sound of this cat.	  *(cat meowing)*

Table 3. Example generation with multimodal inputs and outputs. The instructions, image, and audio are interleaved demanding in-context understanding of the inputs. The outputs are either unimodal and multimodal requiring model's synchronized generation abilities.

new instance. This technique is particularly useful when trying to generalize a concept from a specific instance to a new, but related, context. In Table 1, model is shown a set of images with season change and then asked to apply the same to a similar image. In Table 2, model is shown a set of audio where the latter one has raining sound on top of it, and then asked to apply the same to a new audio. In Table 3, model is shown the audio caption of an image and asked to generate audio for an image with different vibe.

Concept learning involves the model learning from shared concept/attributes of given examples, such as artistic styles or patterns, and then creating new content that exhibits similar concept/attributes. The model's ability to discern and replicate complex patterns indicates a sophisticated understanding of visual styles. In Table 1, model learns the intricate floral patterns and then draws a new im-

Model	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images (Oracle)	0.774	0.885	-
Fine-Tuning			
DreamBooth [28]	0.668	0.803	0.305
Test Time Tuning Free			
Re-Imagen [4]	0.600	0.740	0.270
KoSMOS-G [25]	0.694	0.847	0.287
Ours	**0.703**	**0.852**	**0.311**

Model	CLIPSIM ↑
Diffusion Model Only	
SDEdit-1/2T [21]	0.134
InstructPix2Pix [2]	0.151
Diffusion Model + LLM	
Ours	0.147

Table 4. Left: Comparisons on DreamBench. Right: Image Editing on MS-COCO.

		Adding			Dropping			Replacement		
Model	Text	LSD (↓)	KL(↓)	FD(↓)	LSD (↓)	KL(↓)	FD(↓)	LSD (↓)	KL(↓)	FD(↓)
SDEdit-3/4T [21]	caption	1.54	1.68	28.87	1.54	1.14	29.66	1.63	1.58	28.78
SDEdit-1/2T [21]	caption	1.43	1.38	28.75	1.43	1.05	28.19	1.52	1.27	27.71
SDEdit-1/4T [21]	caption	1.38	1.30	28.25	1.40	1.30	31.31	1.46	1.15	26.72
AUDIT [37]	instruction	1.35	0.92	21.80	1.37	0.95	22.40	1.37	0.84	21.65
Ours	instruction	**1.21**	**0.88**	**19.72**	**1.26**	**0.90**	**18.06**	**1.25**	**0.80**	**17.32**

Table 5. Evaluation results of the adding, dropping, and replacement tasks on auditory data.

age that reflect the same intricate styles.

Subject-driven learning focus on generating new content based on a set of provided images. This approach tests the model's ability to understand and recreate the subject with variations in pose, lighting, or context, while maintaining the subject's distinct features. In Table 1, given several pictures of a specific cat, the model will create a new image of the same cat in new poses. In Table 3, given the cat images, the model can create a video+sound of the same cat.

5. Experiments

5.1. Model Setups

Our implementation is based on Llama2 [35], specifically Llama-2-7b-chat-hf. We use ImageBind [9] which has aligned image, video, audio, text, depth, thermal, and IMU modality encoders. We use ImageBind to encode the image and audio features and project it to the input dimension of the LLM (Llama-2-7b-chat-hf) with a multilayer perceptron (MLP) that consists of a linear mapping, activation, normalization, and one more linear mapping. When the LLM generates the image or audio features, we project them back to ImageBind feature dimension with another MLP. Our image diffusion model is based on StableDiffusion-2.1 [27] (stabilityai/stable-diffusion-2-1-unclip [26]), AudioLDM2 [19], and zeroscope_v2[2].

For images or audio that require higher fidelity to the original input, we additionally feed the original image or audio to the diffusion model alongside the generated features by concatenation of the diffusion noise [2, 19, 27]. This approach is particularly effective in preserving the most perceptual features of the input including instruction editing like adding new content or changing style.

5.2. Image Editing Evaluation

Section 4 shows the evaluation results of subject driven image generation on Dreambench [28] and CLIPSIM scores on MSCOCO. Our method achieves very competitive zero-shot performance, showing our model's generalization to new unseen tasks.

5.3. Audio Editing Evaluation

Table 5 provides an overview of our evaluation results concerning audio manipulation tasks—namely, adding, dropping, and replacing elements within audio tracks. These results are pivotal in understanding the effectiveness of the proposed methods. It is evident from this table that our approach demonstrates superior performance in comparison to previous methodologies. Notably, it has achieved the lowest scores across all metrics—Log Spectral Distance (LSD), Kullback-Leibler (KL) divergence, and Fréchet Distance (FD)—across all three editing tasks.

6. Conclusion

We introduced CoDi-2, a model for multimodal generation with groundbreaking abilities such as modality-interleaved instruction following, in-context generation, user-model interaction through multi-round conversations. CoDi-2 is able to processes complex modality-interleaved input and instructions by MLLM, and then autoregressively produce the latent features that is fed to diffusers for multimodal generation. The evaluations show that CoDi-2 has exceptional zero-shot and few-shot ability on tasks including style adaptation, subject-driven generation, and editing across modalities. CoDi-2 represents a remarkable exploration to build the GPT-like fundamental multimodal system.

[2] https://huggingface.co/cerspense/zeroscope_v2_576w

References

[1] 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 IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 4

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 3, 4, 8

[3] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 1

[4] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 8

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[6] 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. *arXiv preprint arXiv:2105.13290*, 2021. 1

[7] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 4

[9] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 8

[10] 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. *arXiv preprint arXiv:2106.09685*, 2021. 1

[11] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 3

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

[13] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023. 2

[14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 6

[16] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2, 3, 4

[17] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[19] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023. 8, 1

[20] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023. 2

[21] 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. *arXiv preprint arXiv:2108.01073*, 2021. 3, 8

[22] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1

[23] OpenAI. Gpt-4 technical report, 2023. 2

[24] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 1

[25] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models, 2023. 3, 6, 8

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 8

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 4, 8, 1

[28] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 3, 8

[29] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 4

[30] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 1

[31] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2

[32] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 4

[33] 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. 6

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 8

[36] 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. *arXiv preprint arXiv:2205.14100*, 2022. 1

[37] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. Audit: Audio editing by following instructions with latent diffusion models. *arXiv preprint arXiv:2304.00830*, 2023. 4, 8

[38] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023. 1, 3

[39] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. The generative ai paradox:" what it can create, it may not understand". *arXiv preprint arXiv:2311.00059*, 2023. 1

[40] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 1, 2

[41] D Yang, J Yu, H Wang, W Wang, C Weng, Y Zou, and D Diffsound Yu. Discrete diffusion model for text-to-sound generation. arxiv 2022. *arXiv preprint arXiv:2207.09983*. 1

[42] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. i-code: An integrative and composable multimodal learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10880–10890, 2023. 2

[43] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[44] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. 2

[45] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[47] 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 IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020. 1

[48] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 3

[49] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 3