A. Appendix Overview

This supplementary section delves deeper into model performance on unimodal generation tasks, the training methodologies, model details, and experimental setups, as well as the construction of datasets discussed in the main paper. The primary aim is to augment the comprehension of diverse multimodal in-context generation scenarios. Moreover, we discuss the potential biases of our model and datasets.

Table 6. Left: COCO-caption FID scores for text-to-image. Right: COCO image captioning. Note that CoDi and SCD-Net that are marked * are based on diffusion models.

Method	FID ↓
CogView [6]	27.10
Stable Diffusion-2.1 [27]	11.1
CoDi [32]	11.26
Next-GPT [40]	11.28
CoDi-2	**10.96**

Model	B@4	METEOR	CIDEr
ClipCap [22]	32.15	27.1	108.35
BLIP2 [17]	43.7	-	145.8
CoDi* [32]	40.2	31.0	149.9
Next-GPT [40]	44.3	**32.9**	156.7
CoDi-2	**45.2**	32.8	**161.0**

Table 7. Top: AudioCaps audio captioning scores comparison. Bottom: The comparison between our audio diffuser and baseline TTA generation models on AudioCaps test set.

Model	SPIDEr	CIDEr	SPICE
AudioCaps [12]	0.369	0.593	0.144
AL-MixGen [3]	0.466	0.755	0.177
CoDi [32]	0.480	0.789	0.182
Next-GPT [40]	0.521	0.802	-
CoDi-2	**0.531**	**0.806**	**0.189**

Model	KL ↓	FAD ↓	OVL ↑	REL ↑
DiffSound [41]	2.52	7.75	45.00	43.83
CoDi [32]	1.40	1.80	66.87	67.60
AudioLDM2 [19]	**0.98**	1.42	3.89	3.87
CoDi-2	**0.98**	**1.40**	3.85	**3.90**

Table 8. MSRVTT video captioning scores comparison.

Model	B@4	METEOR	CIDEr
ORG-TRL [47]	43.6	28.8	50.9
MV-GPT [30]	48.9	38.7	60.0
GIT [36]	54.8	33.1	75.9
CoDi [32]	52.1	32.5	74.4
Next-GPT [40]	58.4	38.5	-
CoDi-2	**58.9**	**39.8**	**82.2**

B. Unimodal Generation Evaluation

We further qualitatively evaluate the synthesis quality of text, image, audio, and video (multiple frames) with single modality as inputs. In Table 6, CoDi-2 achieves SOTA or near SOTA performance on text-to-image generation and image captioning. In Table 7, CoDi-2 achieves SOTA on audio captioning and text-to-audio generation. CoDi-2 can also perform video captioning and with very competitive near SOTA performance as revealed by Table 8.

C. Experiment Setups

C.1. Model Setups

To effectively condition the image diffusion model, we employ negative prompts as cross-attention conditions and utilize MLLM-generated features for embedding guidance [5]. The negative prompts used for image generation include: 'worst quality, normal quality, low quality, low res, blurry, watermark, logo, banner, extra digits, cropped, jpeg artifacts, signature, username, error, sketch, duplicate, ugly, monochrome, horror, geometry, mutation, disgusting'. For audio, the negative prompts are: distorted, muffled, static noise, background noise, interference, echo, low volume, inaudible, drowned out, screeching, piercing, off-key, out of tune, discordant, interrupted, choppy, glitches, overlapping voices, jumbled, incoherent, repetitive, monotonous, tedious, harsh, grating, abrasive, unbalanced, erratic levels, fluctuating volume, hissing. These negative prompts function as unconditioned input guidance, serving to enhance the quality of both the image and generated features.

C.2. Training Pipelines

The training pipeline involves simultaneous text, image, and audio generation via diffusion. To avoid the cumbersome and inefficient aspects of this multitasking approach, especially concerning model I/O, we alternate training phases between text, audio, and image generation. We apply LoRa [10] with a rank of 128 for fine-tuning the model. The fine-tuning process focuses only on the LoRa weights and the projection layers that map modality encoders into the LLM input space, as well as the decoder layer that projects LLM-generated features into the diffusion input space.

D. Extended Details of Multimodal In-Context

This section presents more details on the generation process in multimodal in-context generation datasets.

D.1. GPT-Generated Prompts

We crafted 100 distinct prompt templates for each task type, including instructional editing, multimodal paired datasets, and constructed in-context multimodal generation datasets. For instance, in instructional editing, prompts like '*Given the image* [Image0], *transform it into Van Gogh style*', or '*Presented with the visual* [Image0], *convert it into Van Gogh style*' are used. In multimodal paired datasets, we utilize prompts for text-to-image or audio tasks like, '*Generate an image based on the instruction: a cat on a couch*', or '*Produce audio of a person talking*'. For captioning image or audio, examples include '*Generate a caption for this image:* [Image0]' or '*Given* [Audio0], *produce its description*'. In exemplar learning, a typical prompt is '*Learn the transformation between* [Image0] *and* [Image1], *and apply it to* [Image2]'. For image composition, prompts like '*Create an image according to the description, combining* [Image0] *with* [Image1]' are used. Each task type includes 100 uniquely generated prompt prototypes, which are sampled uniformly during the data pipeline in training.

For training the model: we use 32 NVIDIA A100 to train the model for 5 days. The input sequence length is capped at 1024 for all batches. The minimum memory to fit model training will be 48GB (without partitioning the model). For inference: the I/O will be around 500ms and requires a minimum memory of 16GB for an input sequence of 128 with 16 floating point.

D.2. Addressing the Discrepancy Between Datasets and Practical Applications

In assembling our research, we have meticulously gathered or integrated a vast array of datasets from various sources. Despite this extensive collection, it is notable that several applications or use-cases remain insufficiently represented within our training datasets. A case in point is visual concept learning, which, although not extensively featured in our training data, is an area where our model excels. Overall, the architecture of our model and the design of our tasks are strategically formulated to leverage the innate in-context capabilities of large language models. This approach is complemented by the use of diverse, in-context, and interleaved datasets, sourced from an array of open-domain materials, thereby enhancing the model's applicability and versatility in addressing a broader spectrum of real-world scenarios.

E. Biases

The model, built upon the instruction-tuned LLM (Llama 2) for a broad range of tasks, has been exposed to various instructional styles. CoDi-2, based on CoDi-1, does not require its instructions to be overly descriptive. However, it's important to note potential biases in our datasets. Firstly, there is a selection bias: our datasets predominantly feature content from AIGC editing, which tends to emphasize personalized, stylistic, and aesthetic elements. Secondly, there's a language bias: the datasets include images accompanied by English text, optimizing for English's linguistic structure. This may limit the model's effectiveness in adapting to other languages.

Moreover, our model similar to other language models might not handle well adversarial inputs where there is a long tail distribution of modality features (like image/audio) and text, for instance, an image feature vector place at the end of a very long input sequence.