OmniFlow: Any-to-Any Generation with Multi-Modal Rectified Flows

Shufan Li*[1], Konstantinos Kallidromitis*[2], Akash Gokul*[3], Zichun Liao[1]
Yusuke Kato[2], Kazuki Kozuka [2], Aditya Grover[1]
[1] UCLA [2]Panasonic AI Research [3]Salesforce AI Research
⋆Equal Contribution
Correspondence to jacklishufan@cs.ucla.edu

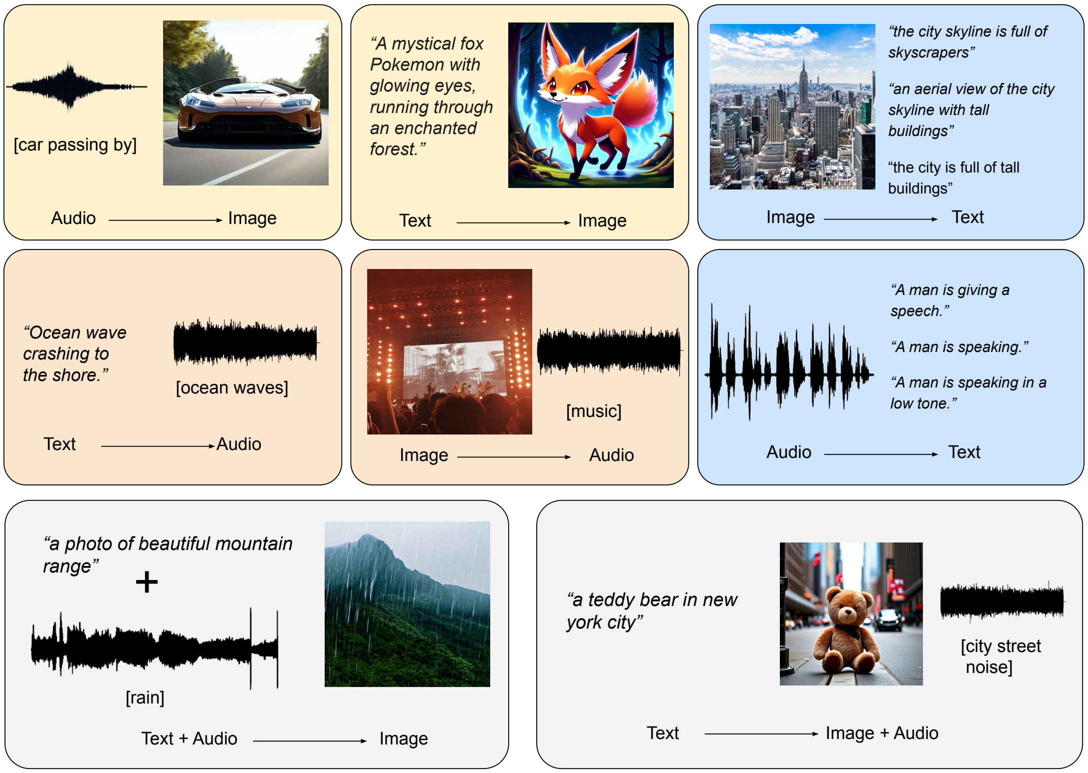

[car passing by]
Audio → Image
"A mystical fox Pokemon with glowing eyes, running through an enchanted forest."
Text → Image
"the city skyline is full of skyscrapers"
"an aerial view of the city skyline with tall buildings"
"the city is full of tall buildings"
Image → Text
"Ocean wave crashing to the shore."
[ocean waves]
Text → Audio
[music]
Image → Audio
"A man is giving a speech."
"A man is speaking."
"A man is speaking in a low tone."
Audio → Text
"a photo of beautiful mountain range"
+
[rain]
Text + Audio → Image
"a teddy bear in new york city"
[city street noise]
Text → Image + Audio

Figure 1. **OmniFlow is capable of a diverse range of any-to-any generation tasks**. OmniFlow supports generation of any output modalities given any input modality, such as text-to-image, text-to-audio, audio-to-image generations. It also supports tasks in multiple input modalities, such as text+audio-to-image.

Abstract

We introduce OmniFlow, a novel generative model designed for any-to-any generation tasks such as text-to-image, text-to-audio, and audio-to-image synthesis. OmniFlow advances the rectified flow (RF) framework used in text-to-image models to handle the joint distribution of multiple modalities. It outperforms previous any-to-any models on a wide range of tasks, such as text-to-image and text-to-audio synthesis. Our work offers three key contributions: First, we extend RF to a multi-modal setting and introduce a novel guidance mechanism, enabling users to flexibly control the alignment between different modalities in the generated outputs. Second, we propose a novel architecture that extends the text-to-image MMDiT architecture of Stable Diffusion 3 and enables audio and text generation. The extended modules can be efficiently pretrained individually and merged with the vanilla text-to-image MMDiT for fine-tuning. Lastly, we conduct a comprehensive study of the design choices of rectified flow transformers for large-scale audio and text *generation, providing valuable insights into optimizing performance across various modalities. Code is available at https://github.com/jacklishufan/OmniFlows.*

1. Introduction

Generative modeling has witnessed considerable advancements in recent years. Notably, diffusion models such as DALLE-3 [40], Stable Diffusion 3 [11], AudioLDM2 [33] achieves state-of-the art performance on text-to-image and text-to-audio tasks. However, these models can only perform a single task while requiring considerable computing resources and data for training. To achieve any-to-any generations, previous works such as CoDi [46] and UIO [36] typically combine a set of modality-specific encoders (*e.g.* ViT [1]) and decoders (*e.g.* Stable Diffusion [44]). However, this design limits these models' ability to integrate information across modalities and generate multi-modal outputs coherently. For example, to perform audio+text-to-image (A+T→I) generation, CoDi simply takes a weighted average of the audio embedding and text embedding to condition an image generator. However, there is no guarantee that the averaged embedding can faithfully represent the two input modalities, as arbitrarily many modality embeddings can average to the same embedding.

An alternative approach for any-to-any generation is to use a single multi-modal model to learn the joint distribution of multiple modalities. This approach has often led to strong performance as it allows information to flow across modalities. However, existing single-model designs typically involve training from scratch, and thus require a considerable amount of data. Existing works in this area, such as UniDiffuser [4] and Chameleon [47] only experiment with text and image modalities. They also require considerable compute resources. To the best of our knowledge, there has yet to be a unified open-sourced multi-modal generative model that supports text, image, and audio simultaneously.

We propose OmniFlow, a unified multi-modal generative model for any-to-any generation. Unlike previous unified multi-modal models, OmniFlow does not need to be trained from scratch with a large amount of data because of its modular design, saving considerable computing resources for its training. OmniFlow is inspired by the MMDiT architecture used in Stable Diffusion 3 [11], which performs text-to-image generation using a two-stream network that combines a text-input stream and an image-output stream through a series of joint attention blocks. OmniFlow builds on MMDIT by incorporating additional input and output streams, extending its text-to-image capability to support any-to-any generation. Crucially, since the parameters for each stream are mostly independent, we can pretrain them separately or initialize them with a pretrained single-task expert model (*e.g.* SD3).

To effectively train OmniFlow, we propose a novel multi-modal rectified flow formulation that incorporates a diverse set of tasks, such as text-to-audio and audio-to-image, into a unified learning objective. Multi-modal rectified flow is built upon a decoupled, time-differentiable interpretation between the distribution of a multi-modal data pair and i.i.d. Gaussian noise. In this formulation, each of the any-to-any generation tasks can be represented by a path connecting two noise levels. For example, given text, image, and audio modalities, the task of text+audio-to-image (T+A→I) can be represented by a path between the distribution of (clean text, clean audio, Gaussian noise) to (clean text, clean audio, clean image).

We conducted extensive evaluations of OmniFlow. Experiment results show that OmniFlow outperforms previous any-to-any models on a wide range of tasks, including text-to-image and text-to-audio generation. Compared to single-task specialist models, OmniFlow achieves competitive performance with state-of-the-art methods.

In summary, our contributions are three-fold:

- First, we extend rectified flow formulation to the multi-modal setting and support flexible learning of any-to-any generation in a unified framework.
- Second, we proposed OmniFlow, a novel modular multi-modal architecture for any-to-any generation tasks. It allows multiple modalities to directly interact with each other while being modular enough to allow individual components to be pretrained independently or initialized from task-specific expert models.
- Lastly, to the best of our knowledge, we are the first work that provides a systematic investigation of the different ways of combining state-of-the-art flow-matching objectives with diffusion transformers for audio and text generation. We provide meaningful insights and hope to help the community develop future multi-modal diffusion models beyond text-to-image generation tasks.

2. Backgrounds

2.1. Flow-Based Generative Models

Flow-based generative models [23, 31, 34, 48], represent the coupling of data points x^0 and noise distribution x^1 using an ordinary differential equation (ODE):

$$dx^t = v_\theta(x^t, t)dt \tag{1}$$

where the velocity v is parameterized by a neural network. Directly solving this equation is expensive. However, we can define a forward process $x^t = a(t)x^0 + b(t)x^1$ to directly regress a conditional vector field using the Conditional Flow Matching (CFM) objective [48] as follows:

$$\mathcal{L}_{\text{CFM}} = (-\frac{b(t)\lambda'(t)}{2})\mathbb{E}_{t,x^1,x^t|x^1}\|\epsilon_\theta(x^t, t) - x^1\|^2 \tag{2}$$

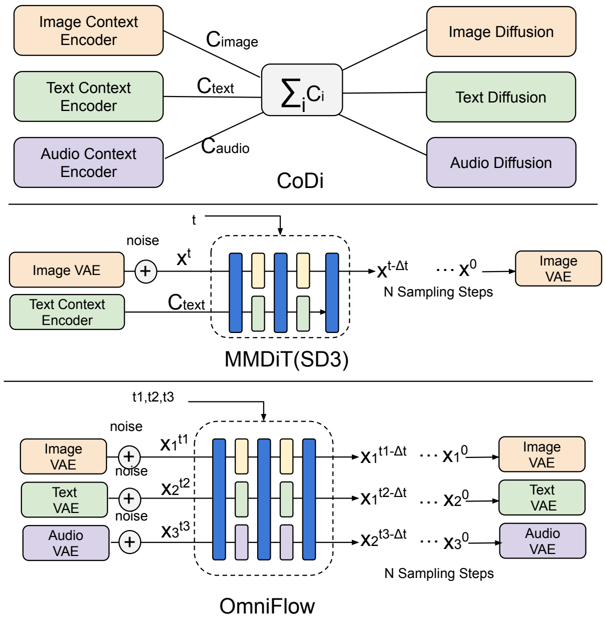


Figure 2. **Pipeline of OmniFlow**. Previous any-to-any models such as CoDi [46] (Top) concatenate multiple modality-specific encoders and decoders, and naively average the embedding of multiple modalities to achieve joint conditioning. By contrast, OmniFlow (Bottom) is a unified, modular multi-modal model, where features from different modalities directly interact with each other through joint attention layers. OmniFlow is inspired by the modular design of Stable Diffusion 3 [11] (Middle), a text-to-image model.

where $\lambda(t) = \log \frac{\alpha(t)^2}{\beta(t)^2}$ is the signal-to-noise ratio (SNR), $\epsilon_\theta(x^t, t) = -\frac{2}{\lambda'(t)b(t)}(v_\theta(x^t, t) - \frac{\alpha'(t)}{\alpha(t)}x^t)$ is parameterized by v_θ. The optimum of this objective remains unchanged when introducing time-dependent weighting, and hence we can rewrite it following [22] as:

$$L_w(x_0) = -\frac{1}{2}\mathbb{E}_{t,\,x^1}\, w(t)\lambda'(t)\|\epsilon_\Theta(z_t, t) - \epsilon\|^2 \tag{3}$$

where, $w(t) = -\frac{1}{2}\lambda'(t)b(t)^2$ for CFM and $x^1 \sim \mathcal{N}(0, I)$ follows noise distribution. This formulation gives a unified representation for a variety of generative modeling approaches. For example, a rectified flow's forward process is defined as $x^t = (1-t)x^0 + tx^1$, which corresponds to $w^{\text{RF}} = \frac{t}{1-t}$. Esser et al. [11] summarized many configurations of common methods under this unified formulation, including (LDM)-Linear [44] and Cosine [39]. They also explored a logit-normal distribution of timestep t for text-to-image generation. We explore all these variants in the context of multi-modal generation, particularly for audio and text, as it is unclear if the results from the text-to-image domain can be directly generalized.

2.2. Any-to-Any Generation

Prior works have explored any-to-any generation. CoDi [46] achieved it first by combining multiple modality-specific encoders (*e.g.* ViT) and decoders (*e.g.* Stable Diffusion) through bridge alignment. However, its design has limited cross-modality interaction. For example, to achieve text+audio-to-image (T+A→I generation), it simply computes the weighted average of text embeddings and audio embedding. Unified-IO [36] models any-to-any generation as a sequence-to-sequence problem, and uses an autoregressive model to achieve any-to-any generation, such as text-to-image or text-to-audio. Our work is the first to use a multi-modal flow matching objective for any-to-any tasks.

Additional works focus exclusively on unifying text-to-image and image-to-text generation. Chameleon [47] uses an LLM-like large autoregressive model to handle multi-modal data. It represents images as VQGAN tokens [50]. Transfusion [52] adopted a similar design, but uses a non-autoregressive diffusion loss for image modeling, while maintaining an autoregressive loss for text generation. Despite their successes, these unified multi-modal models require considerable training resources, because they are less modular than previous works that combine multiple models. OmniFlow achieves a good balance by separating the parameters of each individual modality, while allowing the features of each modality to freely interact with each other at every layer.

3. Method

3.1. Multi-Modal Rectified Flow

We consider the joint distribution $(x_1^0, x_2^0, ..x_n^0) \sim \pi_{data}$ over the space of paired multi-modal data where $x_i \subseteq \mathbb{R}^{d_i}$ is a sample of modality i represented by a vector of d_i dimension. Let $(x_1^1, x_2^1, ..x_n^1) \sim \pi^1$ be the i.i.d Gaussian distribution where $x_i^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a Gaussian vector of d_i dimension. Given empirical observations $x^0 \sim \pi_{data}$, and $x^1 \sim \pi^1$, we consider the decoupled, continuous, time-differential interpolation given by:

$$\frac{\partial x_i^{t_i}}{\partial t_i} = v_i(x_1^{t_1}, x_2^{t_2}, \ldots, x_i^{t_i}, t_1, \ldots, t_i) \tag{4}$$

$$\frac{\partial x_i^{t_i}}{\partial t_j} = 0; i \neq j \tag{5}$$

$$x_i^{t_i} = (1-t_i)x_i^0 + t_i x_i^1 \tag{6}$$

where the independence condition of Eq (2) indicates $x_i^{t_i}$ only moves when t_i moves. Over this interpretation space, we can use a path $\tau : t \rightarrow (t_1, t_2..t_n); [0,1] \rightarrow [0,1]^n$ to model any-to-any generation tasks involving these modalities. For example, given $(x_1, x_2, x_3) \sim p_{data}$ where x_1, x_2, x_3 are image, text, and audio modalities. We can

Algorithm 1 Multi-Modal Rectified Flow

```
Input: Dataset $\mathcal{D}$ consists of modality $1, ...N$,where each sample $x = (x_{i1}^0, x_{i2}^0, ..)$ consists of a subset (or all) of modalities $i_1, i_2.. \in \{1, 2, ..N\}$.
Output: $v_{\theta,i} : (x_1^{t_1}, x_2^{t_2}, ...x_n^{t_n}) \rightarrow v_i^{t_i}$ for each $i = 1, 2..N$, parameterized by $\theta$
Initialize $\theta$
while not converged do
    Sample $x = (x_{i1}^0, x_{i2}^0, ..) \sim \mathcal{D}$
    $x_j^0 \leftarrow \mathbf{0}; \forall j \in \{1, 2..N\} \setminus \{i1, i2...\}$
    Sample path $\tau$.*
    Sample $t \sim$ Uniform$([0, 1])$
    $(t_1...t_N) \leftarrow \tau(t)$
    $x_i^{t_i} \leftarrow x_i^{t_i} = (1 - t_i)x_i^0 + t_i x_i^1; \forall i \in 1, 2..N$
    $\mathcal{L} = \sum_{i \in \{i_1, i_2..\}} \|v_i - v_{\theta,i}(x_1^{t_1}, ...x_n^{t_n}, t_1..t_n)\|^2$
    Perform optimizer step using $\nabla_\theta \mathcal{L}$
end while
Return $\theta$
    ▷ * $\tau$ encodes a task involving only modality $i_1, i_2..$, hence $t_j = 1; \forall j \notin \{i_1, i_2..\}$
```

model text-to-image(T→I) tasks as a path τ_{t2i} such that $\tau_{t2i}(0) = (0, 0, 1)$, which represents a clean text-image pair and $\tau_{t2i}(1) = (1, 0, 1)$, which represents clean text. We can similarly model the joint sampling of text, image and audio set as a path from $(0, 0, 0)$ to $(1, 1, 1)$ and text+image-to-audio ($T + I \rightarrow A$) as a path from $(0, 0, 0)$ to $(0, 0, 1)$.

The flow matching objective would be solving n least squares regression problems for each modality of the form:

$$\min_{v_\theta^i} \mathbb{E}_\tau \int_\tau \mathbb{E}_{x^0, x^1} \|v_i - v_{\theta,i}(x_1^{t_1}, x_2^{t_2}, ...x_n^{t_n}, t_1..t_n)\|^2 ds \tag{7}$$

where $v_i = x_i^0 - x_i^1$, and $v_{\theta,i}$ is a neural network parameterized by θ. We use the same network θ to predict outputs for all modalities $1, 2..N$. The outer expectation is over some prior of paths encoding generation tasks which we are interested in. The integral is calculated over a path $\tau(t) = (t_1, ...t_n)$, and $ds = \frac{\partial t_i}{\partial t} dt$. Concretely, we consider three modalities: image, text, audio in our experiments as modalities: 1, 2 and 3 respectively. We consider the distribution of all possible linear paths $\tau(t) = (t_1, t_2, t_3)$ in $[0, 1]^3$ following the rectified flow formulation. They can encode a diverse set of tasks such as text-to-image or text+image-to-audio.

During training, we do not necessarily need all modalities for each data point. For data points that only contain a subset of three modalities (*e.g*. text-image pairs), we can set the time step of remaining modalities (*e.g*. audio) to 1, which corresponds to complete Gaussian noise. The full training algorithm is given as follows:

At inference, we simply pick a path and use the network prediction to solve for Eq. (5). Notably, for standard text-to-image generation with (x_1, x_2) pairs where x_1 is image and x_2 is text, and x_3 is the missing audio modality, picking a linear path from $(1, 0, 1)$ to $(0, 0, 1)$ is equivalent to the standard single-modality rectified flow (Text→Image) formulation used by Stable Diffusion 3 [11].

3.2. Multi-Modal Guidance

To flexibly control the multi-modal generation process, we extend the classifier free guidance (CFG)[16] to multi-modal rectified flow setting. Recall that CFG of single modalities are formulated as follows:

$$\hat{v}_\theta(x^t, c) = v_\theta(x^t, c) + (\alpha - 1)(v_\theta(x^t, c) - v_\theta(x^t)) \tag{8}$$

where c is a condition and x^t is the noised latent at timestep t of the single-modal output. We extend this formulation to multi-modal setting by defining $\delta_{ij} = v_\theta(x_i^t, x_j^0) - v_\theta(x_i^t)$, which represents the influence of input modality j to output modality i. In particular, we obtain $v_\theta(x_i^t, x_j^0)$ and $v_\theta(x_i^t)$ by setting inputs of modalities not present in the formula to Gaussian noise. For example, given three modalities x_1, x_2, x_3, we can obtain $v_\theta(x_1^t, x_2^0)$ by computing $v_\theta(x_1^t, x_2^0, x_3^1)$ and obtain $v_\theta(x_1^t)$ by computing $v_\theta(x_1^t, x_2^1, x_3^1)$. Note that x_2^1, x_3^1 is just Gaussian noise.

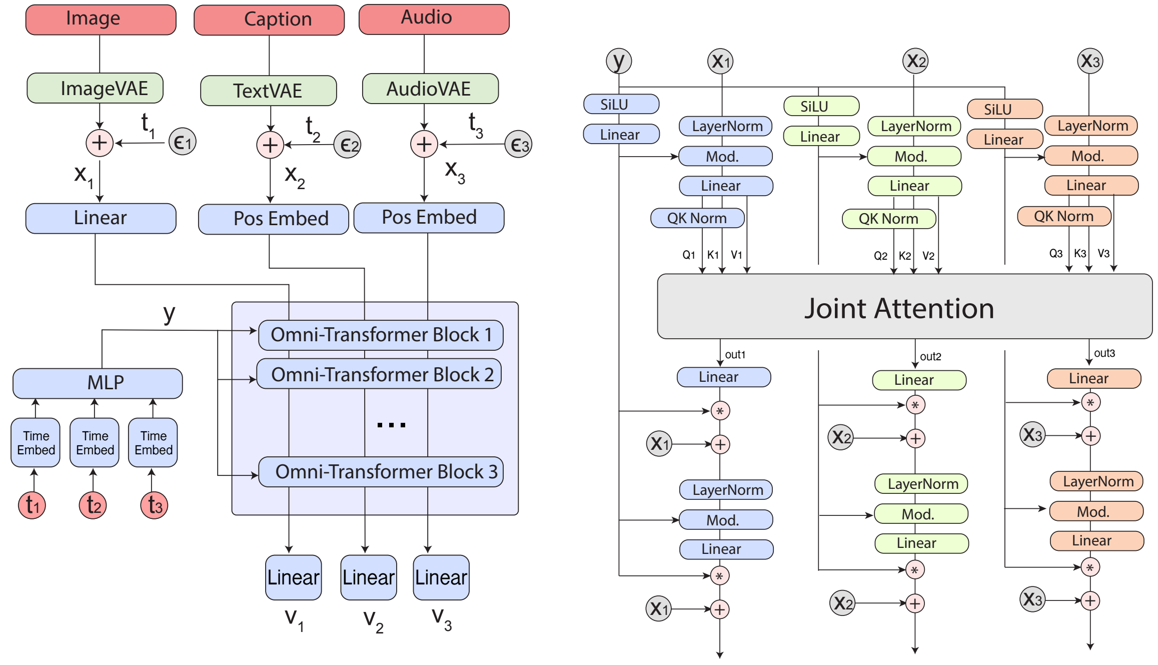
Given the set of δ_{ij}, we can guide the output generation of modality i by the following formula:

$$\hat{v}_\theta(x_1^{t_1}...x_n^{t_n}) = v_\theta(x_1^{t_1}...x_n^{t_n}) + \sum_{j \neq i} (\alpha_{ij} - 1)\delta_{ij} \tag{9}$$

where α_{ij} is the equivalent of α in a multi-modal setting. This scheme allows the user to precisely control the interaction between each of the input and output modalities. When there are only two modalities, our multi-modal guidance Eq. (9) is equivalent to the standard single-modal classifier-free guidance Eq. (8).

3.3. Model Architecture

We propose OmniFlow, a modular, effective extension to the MMDiT architecture used in Stable Diffusion 3. Concretely, given multi-modal inputs that consist of text, image, and audio, we first convert them to latents x_1, x_2, x_3 using modality-specific VAEs. We then add random Gaussian noise to the latents following the forward process defined in Eq. (6). We use the three sinusoidal embeddings to encode, t_1, t_2, t_3 which correlate to the noise scale for each modality. These three timestep embeddings are passed to an MLP to obtain y, a single embedding representing all modality-specific time steps. The final input to OmniFlow are the unified timestep embedding y, and noised latents (x_1, x_2, x_3). These four input vectors are passed to N consecutive Omni-Transformer



(a) Overall Pipeline of OmniFlow

(b) Design of Omni-Transformer Block

Figure 3. **Architecture of OmniFlow**. Left: We highlight the architecture of OmniFlow. Right: We show the design of an individual Omni-Transformer Block.

blocks. The final hidden states of each modality, are then processed by the linear output layer to obtain predictions of v.

Within each Omni-Transformer block, the inputs x_1, x_2, x_3 are processed by modality-specific projections to obtain $q_1, k_1, v_1, q_2, k_2, v_2, q_3, k_3, v_3$. We then concatenate the queries, keys, and values to obtain $Q = \text{Concat}(q_1, q_2, q_3), K = \text{Concat}(k_1, k_2, k_3), V = \text{Concat}(v_1, v_2, v_3)$. The joint attention output for i^{th} modality out_i is given by:

$$\text{out}_i = \text{SoftMax}(\frac{q_i^T K}{\sqrt{d}})V \tag{10}$$

where d is the dimension of each attention head. The output is passed to a feed forward network (FFN) to get the final output of the Omni-Transformer block. Following the design of DiT [41], we use the unified time embedding to modulate the qkv projection and FFN. We add skip connections after the joint attention operation and after the FFN.

We illustrate the model architecture in Fig. 3. Notably, different modalities are handled by different projection and feed-forward layers with independent parameters. The only multi-modal operation is the joint attention, with no trainable parameters of its own. This allows us to pretrain layers of different modalities individually and combine them for fine-tuning, which significantly improves the training efficiency.

4. Setup

4.1. Training Dataset

We use text-image pairs, text-audio pairs, and audio-image pairs during training. We also make use of a small amount of text-image-audio triplets. The text-image pairs include 5M images sampled from COYO-700M dataset [5], 2M images sampled from LAION-Aesthetic-3M subset [25], 7M images from LAION-COCO subset [26], the full CC12M dataset [6], and 2M high-quality image dataset generated by flux-dev and DALLE-3 [14]. We put high weights on images from LAION-Aesthetic-3M and the 2M high-quality images to maintain good aesthetic quality in the output. The text-audio pairs include the full training set of AudioSet [12], Audiocaps [21] and WavCaps [37]. The audio-image pairs include the training data of VGGSound [7] and SoundNet [2]. While SoundNet contains 2M images and is larger than VGGSound, we set the sample weight of VGGSound and SoundNet to 2:1 since SoundNet contains many improperly resized images with bad aspect ratios.

To generate text-image-audio triplets, we use BLIP [28] to generate synthetic captions for videos in VGGSound and SoundNet. We provide further details of the dataset construction in the Appendix.

4.2. Training Recipe

At a high level, we initialize OmniFlow with the text and image modules of Stable-Diffusion 3 (Model 1). We first train a separate text-to-audio model with text-audio pairs (Model 2). Then, we merge Model 1 and Model 2 to obtain a combined model with text, image, and audio modules (Model 3). Since Model 1 and Model 2 have separate text modules, we average their weights during the merge process. Finally, we fine-tune Model 3 on a diverse set of any-to-any tasks using the methods described in Sec. 3.1.

Due to our modular design, we can initialize and pretrain each module individually. This saves immense computational cost when compared to previous unified multi-modal models (*e.g*. UniDiffuser [4]) which are trained from scratch. We use a global batch size of 64 and train Model 2 and Model 3 for 100k, and 150k steps each. We provide further training and implementation details in the Appendix.

5. Main Results

5.1. Evaluation Metrics

We perform extensive experiments on paired generation (text-to-image, text-to-audio) and generic any-to-any generation such as text-to-audio+image (T→I+A), audio-to-text+image (A→T+I). For text-to-image generation, we report FID [15] and CLIP [43] scores on MSCOCO-30K benchmark [30]. Following the official implementation, the cosine similarities between CLIP embeddings are multiplied by 100. We also report results on the GenEval benchmark [13]. For audio generation, we report FAD [20] and CLAP [10] score on AudioCaps. Results are reported with a 16kHz sampling rate. We also use CLAP scores for caption evaluations.

5.2. Text-to-Image Generation

Model	Param	FID↓	CLIP↑
UniDiffuser	0.9B	**9.71**	30.93
CoDi	4.3B	11.26	30.69
UIO-2XXL	6.8B	13.39	-
SDv1.5	0.9B	11.12	30.63
SDXL*	2.6B	16.49	31.36
SD3-Medium*	2B	20.94	30.65
OmniFlow*	3.4B	13.40	**31.54**

Table 1. **Text-to-Image Generation on MSCOCO-30K Benchmark.** *Indicates models pretraining data consists of high quality images and captions that do not follow the distribution of COCO dataset, which can negatively affect FID scores.

We report results on MSCOCO-30k in Tab. 1, and results on GenEval in table Tab. 2. On MSCOCO-30k, we achieve a lower FID than state-of-the-art models such as SDXL and SD3-Medium. While our FID number is higher than some previous models such as SDv1.5, it should be noted that more recent models such as SDXL and SD3 tend to have higher FID numbers because they are trained on high-quality text-image pairs that do not match the distribution of COCO images [42]. Notably, SD3 has a FID of 20.94 while SDv1.5 has 11.12, even though SD3 is considered a better model according to human evaluations. SDXL, which is widely recognized as the state-of-the-art open-source model before the release of SD3, also has a higher FID than SDv1.5.

Model	Param	Images	Gen.↑
Text-to-Image Specialist			
SD1.5	0.9B	4.0B	.43
SDv2.1	0.9B	2.3B	.50
SDXL	2.6B	1.6B	.55
DALL-E 2	4.2B	2.6B	.52
SD3-Medium	2B	1B	.62
SD3-Large	8B	2.0B	.68
Generalist			
CoDi	4.3B	400M*	.38
UniDiff.	0.9B	2B	.43
OmniFlow	3.4B	30M*	**.62**
Chameleon	7B	3.5B	.39
Transfusion	7B	3.5B	.63

Table 2. **Text-to-Image Generation on GenEval Benchmark.** We compare the model size, number of training images and GenEval benmark Score. * Indicates fine-tuning dataset. CoDi and MMDiT-O are both initialized with pretrained text-to-image diffusion models (SD and SD3).

In terms of CLIP scores, OmniFlow significantly outperforms previous models. In particular, when contrasted with generalist models UniDiffuser and CoDi, we achieve a gain of +0.61 and +0.85 respectively, showing superior text-to-image alignment. On GenEval Benchmark, which better measures the text-to-image capabilities, OmniFlow achieves a score of 0.62, a competitive score even when compared to the state-of-the-art specialist SD3-Medium. In addition, OmniFlow significantly outperforms previous any-to-any baselines at the same scale, such as CoDi (+.24) and UniDiffuser (+.19). Compared with larger models trained on a lot more images, OmniFlow outperforms Chameleon-7B and achieves competitive performance as Transfusion-7B.

Notably, unlike Chameleon, Transfusion, and UniDiffser which need to be trained from scratch, OmniFlow achieves high performance with only 30M training images, highlighting the effectiveness of our modular design. While the design of CoDi also allows it to make use of pretrained text-to-image model as its initialization, it is trained with considerably more images than OmniFlow while performing worse.

Model	**Param**	**FAD↓**	**CLAP↑**
Text-to-Audio Specialist			
AudioGen-L[24]	1B	1.82	-
Make-an-Audio[19]	0.4B	2.66	-
AudioLDM-L[32]	0.7B	1.96	.141
Make-an-Audio 2[18]	0.9B	2.05	.173
AudioLDM 2-Full-L[33]	0.7B	1.86	.182
Generalist			
CoDi	3.4B	1.80	.053*
OmniFlow	3.4B	**1.75**	**.183**
UIO-2XXL	6.7B	2.64	-

Table 3. **Text-to-Audio Generation on AudioCaps Evaluation Set.** Comparison of FAD and CLAP scores for various audio generators. *Reproduced from official checkpoint, see Appendix for details.

5.3. Text-to-Audio Generation

We report text-to-audio generation results on AudioCaps in Tab. 3. Compared with previous state-of-the-art, OmniFlow achieves strong performance on FAD and CLAP scores. It outperforms AudioLDM2 on FAD (-0.11) and achieves equivalent performance on CLAP (+0.001). When compared with generalist models, OmniFlow significantly outperforms CoDi on both FAD (-0.05) and CLAP (+.13) metrics.

5.4. Recipes for Audio and Text Diffusions

	Audio Gen.	**Text Gen.**
	FAD↓	**CLAP↑**
Continuous Flow Matching		
eps/linear	2.08	.141
v/cos	2.01	.203
v/linear	1.86	.126
rf/uniform	1.82	.227
rf/lognorm	**1.79**	**.254**
Discrete Text Diffusion		
SEDD[35]	-	.180
MDLM[45]	-	.163

Table 4. **Various Formulations for Audio and Text Generation.** We report FAD for audio generation and CLAP for text generation on AudioCaps dataset.

We explore various recipes for training audio and text diffusion transformers for multi-modal generation, which is a relatively under-explored area. Concretely, we explored five formulations mentioned in the section Sec. 2.1. For these experiments, we used a model with only audio and text modules (Model 2 in Sec. 4.2) and trained for 50k steps. We report FAD score for text-to-audio generation and CLAP score for audio-to-text generation. Amongst all five formulations, rf/lognorm performs the best with the lowest FAD (1.79) and highest CLAP score (.254). We also explored two discrete space diffusion models, SEDD [35] and MDLM [45] which showed advantages over continuous-space diffusion models in recent literature. Specifically, we use the absorbing state version of SEDD. For these experiments, the text-vae encoder is replaced with a token-embedding layer, and, text-vae decoder is replaced with a simple linear output layer to predict token logits. We also replace the flow-matching loss on the text-embedding with the loss function of SEDD and MDLM respectively, which operates on token logits instead of continuous embeddings. We report the CLAP score on audio-to-text generation. We do not see considerable advantages over continuous alternatives.

6. Sampling

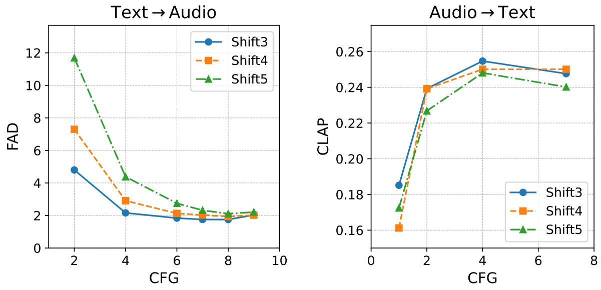
On the sampling side, we explored the effect of guidance and timestep shift. The timestep shift was originally introduced by SD3 to balance the sampling process of images at different resolutions. Concretely, it augments the inference schedule as:

$$\hat{t} = \frac{\gamma t}{1 + (1 - \gamma)t} \tag{11}$$

where $\gamma = \sqrt{\frac{m}{n}}$, with m being the target sample resolution and n being a reference resolution. For audio and text generation, there is no concept of varying resolution, as the input audio spectrogram and text embedding have fixed resolutions. However, we empirically observe applying a shift can improve the generation quality. Concretely, incorporating the shift term $\gamma > 1$ will lead to a concave schedule, where the denoising process progresses slowly at the beginning and accelerates towards the end. We find that this improves sample quality for text-to-audio and audio-to-text generation tasks.

We employ the multi-modal guidance mentioned in Sec. 3.2. For simple audio-to-text and text-to-audio generation, our formulation is reduced to standard classifier-free guidance. We show the effect of guidance and timestep shift in Fig. 4. Generally, we find that shift=3.0 works well for both tasks. For audio generation, a guidance scale of 8 achieves the highest performance. For text generation, a guidance scale of 4 achieves the best result.

To explore the effect of multi-modal guidance in Sec. 3.2, we provide qualitative results for audio+image-to-text (A+I→T) task. Recall that we use x_1, x_2, x_3 to denote image, text, and audio modalities. The multi-modal guidance for this task can be controlled by α_{21} and α_{23} where α_{21} controls text-image alignment and α_{23} controls text-audio alignment. For simplicity, we denote α_{21} as α_{im} and α_{23} as α_{au}. We vary α_{im}, α_{au} between the interval $[1.0, 2.0]$ such that $\alpha_{\text{im}} + \alpha_{\text{au}} = 3.0$. We show the results in Fig. 5. Qualitatively, higher α_{au} will make the model's output resemble more the audio captions, and α_{im} will make the model's output resembles more the image captions. Interestingly, we observe that it also reflects the subtle differences in the



(a) Text-to-Audio Generation.

(b) Audio-to-Text Generation.

Figure 4. **Effect of CFG and Shift for audio and text generation**. We evaluate the impact of guidance and timestep shift on text-to-audio and audio-to-text tasks.

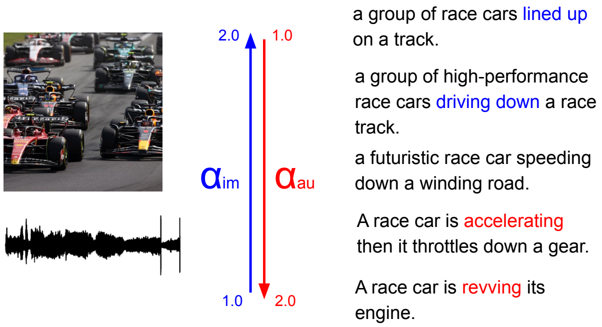

2.0
1.0
α_{im}
α_{au}
1.0
2.0
a group of race cars lined up on a track.
a group of high-performance race cars driving down a race track.
a futuristic race car speeding down a winding road.
A race car is accelerating then it throttles down a gear.
A race car is revving its engine.

Figure 5. **Effect of Multi-Modal Guidance.** In this example, the user can flexibly control the alignment between output text and input image, audio independently by varying α_{au} and α_{im}. Higher α_{im} will make the output texts resemble image captions, with visual descriptions such as lined up, driving down. Higher α_{au} will make the output texts resemble audio captions, with descriptions such as accelerating, revving.

style of audio and image captions in the training data (*e.g.* whether the first letter is capitalized). By varying these two parameters, users can achieve flexible control of generation.

6.1. Qualitative Comparison

We directly compare OmniFlow with two recent any-to-any generation methods: CoDi [46] and UniDiffuser [4]. In addition to the quantitative results, we present qualitative text-to-image comparisons in Fig. 6. These examples demonstrate that OmniFlow achieves a significant improvement in generation quality compared to previous any-to-any models. Specifically, in the first example (top), our model successfully follows the prompt while maintaining high aesthetic quality, accurately capturing both the cat's features and its mirrored reflection. In contrast, CoDi is unable to change the cat's eyes, and UniDiffuser fails to depict the cat looking at the mirror. A similar trend is evident in the third example: OmniFlow correctly positions lanterns tied to a rope, while UniDiffuser places them on the river. Finally, in the lighthouse example, CoDi fails to incorporate seagulls, and UniDiffuser ignores the adjective "gentle," instead producing an image with rough waves and an out-of-focus lighthouse.

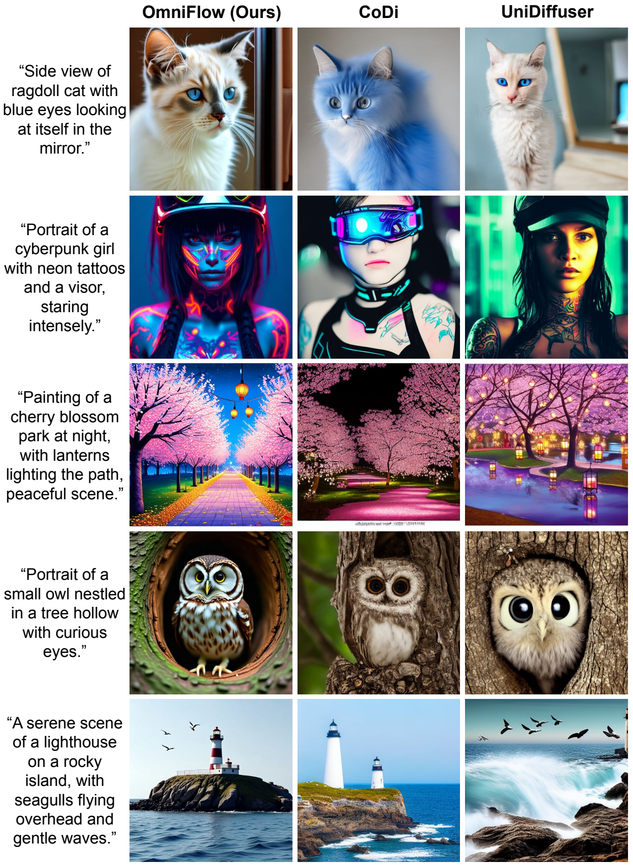

OmniFlow (Ours)
CoDi
UniDiffuser
"Side view of ragdoll cat with blue eyes looking at itself in the mirror."
"Portrait of a cyberpunk girl with neon tattoos and a visor, staring intensely."
"Painting of a cherry blossom park at night, with lanterns lighting the path, peaceful scene."
"Portrait of a small owl nestled in a tree hollow with curious eyes."
"A serene scene of a lighthouse on a rocky island, with seagulls flying overhead and gentle waves."

Figure 6. **Qualitative Comparison with baselines on text-to-image generation.** OmniFlow achieves better image quality and prompt alignment when compared to previous generalist models.

Our results show that OmniFlow achieves a much higher generation quality compared with previous any-to-any models, both in terms of image-text alignment and image fidelity.

7. Conclusion

We present OmniFlow, a unified early-fusion multi-modal generative model for any-to-any generation tasks. OmniFlow adapts a modular design that enables individual components to be pretrained separately, while allowing features from different modalities to directly interact with each other, through a joint attention mechanism. We conduct extensive experiments to show that OmniFlow outperforms previous any-to-any models on a wide range of challenging generation tasks, including text-to-image and text-to-audio generation. We provide further analysis on the limitation of OmniFlow in the Appendix.

8. Acknowledgments

This research was supported by NSF Career #2341040 and a Schmidt Science Fellowship.

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 2

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 5, 12

[3] JISHENG BAI, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. Audiosetcaps: Enriched audio captioning dataset generation using large audio language models. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. 12

[4] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023. 2, 6, 8, 15

[5] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 5

[6] 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 *CVPR*, 2021. 5

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 5, 12

[8] Wenxi Chen, Ziyang Ma, Xiquan Li, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Kai Yu, and Xie Chen. Slam-aac: Enhancing audio captioning with paraphrasing augmentation and clap-refine through llms. *arXiv preprint arXiv:2410.09503*, 2024. 15

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 13

[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 4, 13

[12] 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. 5

[13] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[14] Jacky Hate. Text-to-image-2m dataset, 2024. Accessed: 2024-11-14. 5

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 14

[18] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023. 7

[19] 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 *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 7

[20] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018. 6

[21] 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. 5

[22] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[23] Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[24] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 7

[25] LAION. Aesthetics for open source, 2023. Accessed: 2024-11-14. 5

[26] LAION. Laion coco: 600m synthetic captions from laion2b-en, 2023. Accessed: 2024-11-14. 5

[27] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. *arXiv preprint arXiv:2004.04092*, 2020. 13

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 13, 15

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[31] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[32] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 7, 13

[33] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 2, 7

[34] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[35] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024. 7, 14

[36] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 2, 3

[37] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 5

[38] MidJourney AI. Image generated using midjourney ai, 2024. Accessed on November 21, 2024. URL: https://www.midjourney.com/. 15, 17

[39] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 3

[40] OpenAI. Dall-e 3, 2023. 2

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[42] 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. *arXiv preprint arXiv:2307.01952*, 2023. 6

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[44] 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. 2, 3, 13

[45] Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7, 14

[46] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 8, 14, 15, 18

[47] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3

[48] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2(3), 2023. 2

[49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 15

[50] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[51] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. 13

[52] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma,

Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 3

[53] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2023. 13