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

Supplementary Material

	Size	Modality
LAION-Aesthetics-3M	2M*	T,I
CC12M	12M	T,I
COYO-700M(Subset)	5M	T,I
LAION-COCO	7M	T,I
SoundNet	2M	T,A,I†
VGGSound	0.2M	T,A,I†
T2I-2M	2M	T,I
AudioSet	2M	T,A
AudioCaps	46K	T,A
WavCaps	0.4M	T,A

Table 5. **List of all datasets used in training.** *Some image URLs are no longer accessible. † We generate synthetic captions using BLIP.

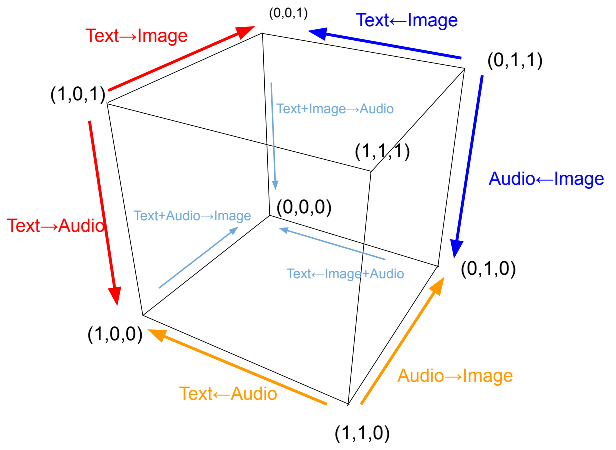


Figure 7. **Paths encoding different any-to-any generation tasks.** (t_1, t_2, t_3) represents the "noise level" of image, text and audio modalities. $(0, 0, 0)$ represents clean (image, text, audio) triplets, and $(1, 1, 1)$ represents pure Gaussian noise.

A. Implementation Details

A.1. Dataset

In Tab. 5, we list the size of all datasets used in training. We filter out all images whose shortest side is less than 256 pixels. To obtain data with all modalities (image, audio, text), we use BLIP to generate synthetic captions for images in the SoundNet[2] and VGGSound[7] dataset, which are extracted from videos. Since AudioSet only comes with class labels, we use synthetic captions generated by audio-language models provided by AudioSetCaps[3].

A.2. Schedules

Recall from Section 3 that we can represent different tasks with different paths in $[0, 1]^3$. We visualize this in Fig. 7. We adopted simple linear tasks for any-to-any generation tasks so that for simple cases like text-to-image and text-to-audio, our formulation matches the standard rectified flow.

A.3. Training Pipeline

We initialize our model with SD3 (Model 1 in Fig. 8). We first train the model on text-audio pairs to obtain Model 2. The text branch of Model 2 is initialized with weights of SD3, while the audio branch is randomly initialized. After the training, we merge Model 1, which contains a text branch and an image branch, and Model 2, which contains a text branch and an audio branch, to Model 3, which contains text, image, and audio branches. The text branch of Model 3 is obtained by averaging the weights of the text branches from Model 1 and 2. Finally, we train the Model 3 on all datasets mentioned in Suppl. A.1. This training pipeline is illustrated in Fig. 8.

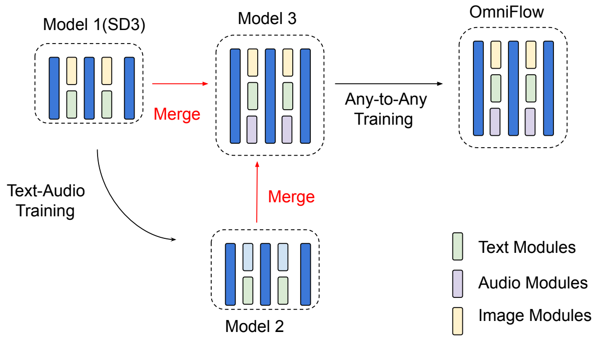


Figure 8. **Training Pipeline of OmniFlow.** We initialize our model with SD3 (Model 1). We then train the model on text-audio pairs to obtain Model 2. We merge Model 1 and Model 2 to obtain Model 3. The final model is obtained by further training Model 3 on any-to-any generation tasks.

We train Model 2 for 100k steps and Model 3 for 150k steps. We use 8 A6000 GPUs with a per GPU batch size of 8. We use AdamW optimizer with a learning rate of 1e-5 for Model 2 and 5e-6 for Model 3. The learning rate undergoes a linear warmup in the first 1000 steps and a cosine decay throughout the rest of the training. We adopt exponential moving average (EMA), which are updated every 100 training steps with a decay factor of 0.999.

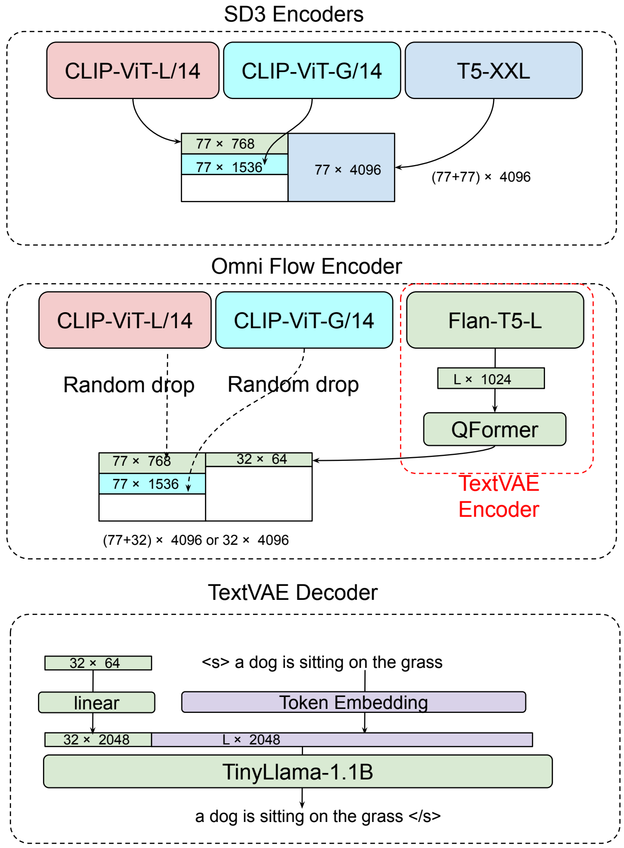


Figure 9. **Architecture of Text VAE and Text Encoders in OmniFlow.** SD3 (Top) uses three text encoders: CLIP-L, CLIP-G, and T5-XXL. OmniFlow (Middile) replaces the 4.7B T5-XXL with a VAE encoder based on Flan-T5-L. CLIP encoders become optional and are not used for tasks without clean text inputs. The decoder of VAE (Bottom) is based on TinyLlama-1.1B. The VAE embedding is used as the prefix for decoding.

A.4. Text VAE

We train a text VAE on caption data using Flan-T5-L [9]. Recall that SD3[11] makes use of three text encoders: CLIP-L, CLIP-G and T5-XXL. We replace the 4.7B T5-XXL with Flan-T5-L [27] to save computation cost and use it as part of a text VAE. Specifically, given an input caption of length L, it is first encoded by Flan-T5-L to obtain a vector of size $L \times 1024$. We then pass it to a QFormer[29] and obtain an output vector of size 32×64. This vector is used as the VAE embedding. In the decoding process, the VAE embedding is first processed by a linear projection layer to obtain a vector of size 32×2048. This is used as the prefix embedding for a TinyLlama-1.1B decoder [51]. These architecture designs are shown in Fig. 9. Note that while we introduced a 1.1B text-decoder, the overall system actually has fewer parameters since we replaced the 4.7B T5-XXL with a 783M Flan-T5-L.

We employ the auto-encoding training objective of OPTIMUS [27]. We freeze the Flan-T5-L encoder and fine-tune the QFormer and TinyLlama decoder end-to-end. We train the text VAE on all caption data mentioned in Suppl. A.1 for 2 epochs, with a learning rate of 1e-5, a global batch size of 256 using AdamW optimizer.

When using the VAE encoder as the text encoder of OmniFlow, we pad the embedding to 4096 with zeros to maintain the input dimension of SD3. Additionally, we also incorporate the CLIP-L and CLIP-G encoders of SD3 as auxiliary text encoders to stabilize the training. We apply random dropout to these encoders during the training. During the inference, the CLIP encoders are not used if the input does not contain clean texts (e.g. Image-to-Text task).

A.5. Audio VAE

We directly adapt the audio VAE used by AudioLDM [32]. In particular,we adopt the same vocoder and preprocessing pipeline as AudioLDM2. We use HiFiGen as VAE, which is used in AudioLDM and AudioLDM2. We use AudioLDM2's checkpoint. We also explored AudioMAE, but found it to perform significantly worse as measured by FAD (2.03 vs 1.79).

A.6. Omni-Transformer

We followed the architectural design of SD3 for image and text modules and initialize them with SD3 weights. The audio modules are initialized with identical setup to the image modules. Specifically, it has 24 layers and a hidden size of 1536. The positional embedding layer has a patch size of 2. Since the audio VAE outputs a feature map of dimension 256×16, the positional embedding layer will convert each audio to a sequence of length $128 \times 8 = 1024$.

A.7. Pooled Conditional Embeddings

SD3 makes use of additional pooled embeddings from CLIP-ViT-L/14 and CLIP-ViT-G/14 in addition to the sequence embeddings. We maintain them as is, with additional dropout during the training. We additionally incorporate an Audio Encoder to create pooled embeddings for audio inputs [53]. These embeddings are not used when clean data of respective modality is not available.

A.8. Baselines

In this section, we describe the specific variants studied in Tab. 4. Except for the discrete text diffusions (SEDD and MDLM), these variants fit into the unified formulation of Eq. (3) by varying its parameters.

linear is a variant of DDPM used in LDM [44]. It discretizes the timesteps to $0, 1...T-1$ and uses the formulation $b_t = \sqrt{1-\alpha_t^2}$, where $a_t = \sqrt{\prod_{i=0}^{t}(1-\beta_i)}$, and $\beta_t = (\sqrt{\beta_0} + \frac{t}{T-1}(\sqrt{\beta_{T-1}} - \sqrt{\beta_0}))^2$. We explored ϵ-prediction and v-prediction objectives for this variant.

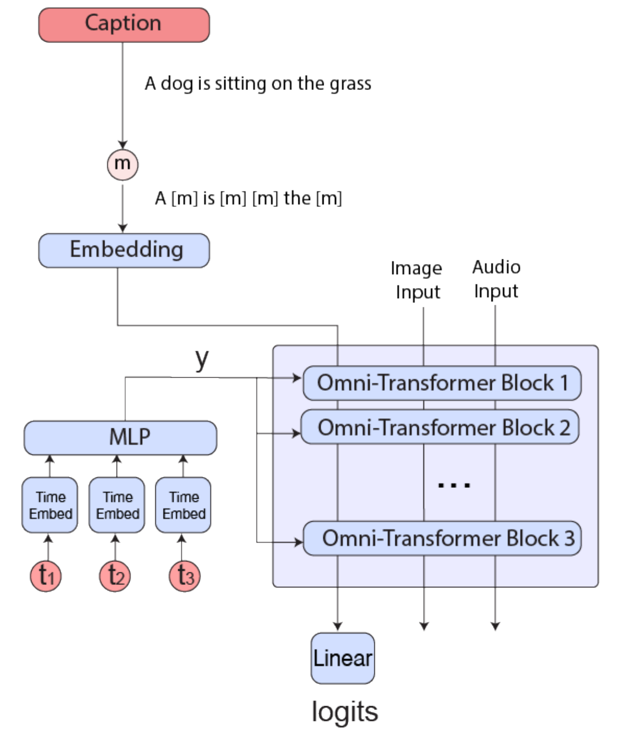


Figure 10. **Discrete Diffusion Variant of OmniFlow.** In this setup, we remove the text VAE and directly pass token embedding to the Omni-Transformer layers. “[m]” indicates a mask token.

cosine is defined by the forward process

$$x^t = \cos(\frac{\pi}{2}t)x^0 + = \sin(\frac{\pi}{2}t)x^1 \tag{12}$$

The weighting function is $w_t = e^{-\lambda_t/2}$ for v-prediction objectives[17].

SEDD and MDLM are recently proposed discrete text-diffusion models. We consider MDLM[45] and the absorbing state variants of SEDD[35] in our experiments.[1] These models directly define a forward process in the discrete token space, where clean text tokens are progressively replaced with a special “[MASK]” token. We adapt our implementation for these methods by removing the text VAE and introducing a token embedding layer. This design is shown in Fig. 10.

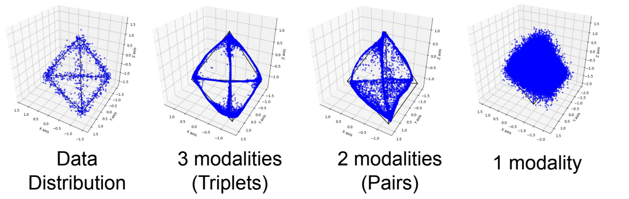


Figure 11. **Synthetic Experiments on three 1D-modalities.** We consider the joint distribution of three toy modalities (x_1, x_2, x_3), each represented by a vector of dimension 1. Hence, a triplet consisting of three modalities be represented by a point in $\mathbb{R}^3$ We assume the joint distribution is a uniform distribution in the neighborhood of tetrahedron (Left). We experiment with training OmniFlowusing triplets, pairs, and only individual modalities. Models trained with triplets of three modalities best represent the original distribution.

B. Additional Discussions

B.1. Sampling

OmniFlow does not directly model the marginals of two modalities. For example, given three modalities (x_1, x_2, x_3), it does not directly model $p(x_1^0|x_2^0) = \int_{x_3^1 \in \mathbb{R}^{d_3}} p(x_1^0, x_3^1|x_2^0)dA$, where d_3 is the dimension of x_3^1. Integrating over x_3^1 is infeasible. Instead, we sample $p(x_1^0, x_3^1|x_2^0)$ by first sample $x_3^1 \sim q(x_3^1|x_2^0) = \mathcal{N}(0, I)$ and sample $p(x_1^0|x_3^1, x_2^0)$ using path from (1,0,1) to (0,0,1).

B.2. Necessity of text, image, audio triplets.

Compared with previous works such as CoDi[46] which uses weighted average of embeddings to mix multiple input modalities, OmniFlow requires directly training on triplets consisting of all modalities (image, text, audio). To study the necessity of this requirement, we conduct synthetic toy examples on three modalities (x_1, x_2, x_3), each represented by a one-dimensional vector. A triplet of three modalities can then be represented by a point (X, Y, Z) in 3D space. We show this experiments in Fig. 11. We assume the ground truth data distribution follows a uniform distribution in a small neighborhood adjacent to a tetrahedron (Leftmost Figure). We experiment with training an 8-layer MLP with triplets (x_1, x_2, x_3) (Second-Left Figure), pairs of (x_1, x_2),(x_1, x_3),(x_2, x_3) (Second-Right Figure), and only individual modalities (x_1), (x_2), (x_3) (Rightmost Figure). For each model, we plot 50k samples generated by the model. Qualitatively, models trained on triplets best represent the data distribution. This makes sense as pairs are essentially projections on XY, XZ, YZ planes and individual modalities are projections on X, Y, Z axis. These projections are not sufficient to recover the original distribution represented in this 3D space.

[1] SEDD also has a uniform variant, where the tokens are not replaced with a ”[MASK]” token, but a randomly token sampled from the vocabulary.

			AudioCaps		COCO-Karpathy	
	Images	**Parms.**	**CLAP↑**	**CIDEr↑**	**CLIP↑**	**CIDEr↑**
Specialist						
BLIP-2[29]	129M	2.7B	-	-	-	145.8 ‡
SLAM-AAC[8]	-	7B	-	84.1‡	-	-
Generalist						
OmniFlow	30M	3.4B	**0.254**	**48.0**	26.8	**47.3**
CoDi †	400M	4.3B	0.206	7.9	25.9	17.2
Unidiffuser †	2B	0.9B	-	-	**29.3**	20.5
UIO2-XXL	1B*	6.8B	-	48.9	-	125.4*
Transfusion	3.5B	7B	-	-	-	35.2

Table 6. **X-to-Text Performance comparison on AudioCaps and COCO Captions.** * UIO2's training data includes COCO. The fine-tuning dataset also includes 53M image understanding data, including 14 image captioning datasets. † evaluated with official checkpoints. ‡ fine-tuned on respective datasets (COCO and Audiocaps).

C. Quantative Text Evaluation

We report quantitative results of image captioning on COCO-Karpathy-Test dataset and audio captioning on Audiocaps dataset. We report CLIP score, CLAP score, and CIDEr[49] on these two benchmarks. We compare against generalist models such as CoDi and Uni-Diffuser. Uni-Diffuser, released two checkpoints v0 and v1, where v1 is fine-tuned on internal data. We compare against v0 for a fairness. OmniFlow outperforms CoDi on both tasks, and outperforms UniDiffuser in CIDEr score (+26.8). It has a lower CLIP score (-2.5). We consider the performance of OmniFlow as competitive, considering OmniFlow is trained on significantly less data than UniDiffuser and can also perform audio captioning task. We note that the performance of generalist models significantly lags behind specialist models that are fine-tuned of respective datasets, suggesting rooms for further improvements. We provide further discussion in the limitation section.

D. Benefits of Multi-Task Training

In this section, we discuss if joint training on multiple modalities benefit single-task performance. We provide additional results in Tab. 7 by comparing a model trained on all tasks with a model only trained on a subset of tasks. We show that Omniflow was able to leverage the training data of related tasks (e.g. T2A, I2A) and boost individual performance. In additional to results presented in Tab.R1, we also observe improvements in image generation, where OmniFlow generate high fidelity A2I outputs even though A2I datasets consist of low-res videos (+1.22 Aesthetic score), thanks to high-fidelity T2I data.

Training Data	**FAD ↓**
OmniFlow	**1.83**
I2A,A2I,T2A,A2T	1.89
I2A,A2I	2.03
I2A-Only	2.05

Table 7. **Performance of Various Training Data Compositions.** We compare FAD scores for Image to Audio (I2A) under different setup on VGGSound.

E. Additional Qualitative Results

E.1. Text-to-Image

Fig. 14 demonstrates a range of qualitative text-to-image examples for OmniFlow. We depict a wide variety of people, scenes and objects to demonstrate the robustness of our approach.

E.2. Image-to-Text

We provide a side-by-side image-to-text comparison between OmniFlow , CoDi [46] and UniDiffuser [4] using synthetic high quality images from the Midjourney Explore page [38]in Fig. 12.

E.3. Audio-to-Text

In Tab. 8, we show qualitative results on Audiocaps audio-to-text task. OmniFlow can generation captions that match the ground truth. While CoDi can correctly grasp the main objects in the audio such as "car", "bird", "sheep", "computer", it struggles with generating captions that accurately reflect the scene.

E.4. Text-VAE AutoEncoding

In Tab. 9, we show reconstruction examples of Text VAE. The reconstruction mostly adheres to the semantics of the ground truth, with minor differences. For example, it may change "well-furnished" to "well-decorated".

ID	CoDi	OmniFlow	GT
yVjivgsU2aA	Four car driver trying forcoming for a speeding car.	A race car engine revs and tires squeal.	An engine running followed by the engine revving and tires screeching.
8F-ndyrEWJ8	Fire police cars stop and red traffic on different highway.	A fire siren goes off loudly as a man shouts and a low hum of an engine is running throughout the whole time.	A distant police siren, then racing car engine noise, and a man calling in police code over his radio.
350OCezayrk	Four motor car driving for completing an automobile service.	A vehicle engine is revving and idling.	A motor vehicle engine starter grinds, and a mid-size engine starts up and idles smoothly.
LCwSUVuTyvg	Door, a blue hat and winter jacket.	A door is being slammed.	Glass doors slamming and sliding shut.
7XUt6sQS7nM	The sheep of the woman are the sheep of the sheep.	Multiple sheep bleat nearby.	A sheep is bleating and a crowd is murmuring.
PVvi2SDOjVc	Car going for a car coming home. Three cars coming for a blue car coming down a road after the highway.	A car horn beeps.	A car engine idles and then the horn blows.
Z_smJ66Tb3c	Men in the bird while the man in the boat.	Two men talk over blowing wind and bird chirps.	A man is speaking with bird sounds in the background followed by a whistling sound.
CMNlIW6Lkwc	Two men in the fire and two men are coming towards the other man in the game.	A man speaks, followed by a loud bang and people laughing.	A man talking as a camera muffles followed by a loud explosion then a group of people laughing and talking.
JQz40TkjymY	Writing computers for people in writing.	Typing on a computer keyboard.	Typing on a computer keyboard.
U90e2P9jy30	A man shouts the word to the person on the sidewalk to walk to get him to the door the hand to fall down on the sidewalk in.	Basketballs being dribbled and people talking.	Several basketballs bouncing and shoes squeaking on a hardwood surface as a man yells in the distance.
5I8lmN8rwDM	Stationary fire drill technician drilling down a hose pipe while wearing safety gear. Railroad safety drill for motorcycle with hose or oil checking equipment.	A drill runs continuously.	Drilling noise loud and continues.
NlKlRKz8OKI	Birds on blue birds.	A woman talks and then an animal chewing.	A woman speaks with flapping wings and chirping birds.

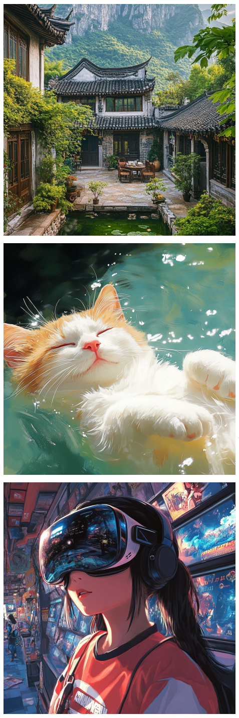
Table 8. **Qualitative comparisons of CoDi and OmniFlow on Audiocaps audio captioning task.** Audios are randomly sampled. Audiocaps provide five ground truth captions per audio. For better presentation, we only list one in this table.

F. Limitations

On text generation tasks, our model's performance is not state-of-the-art and has considerable room for improvements. We believe this is the side effect of incorporating large-scale data with many noisy texts of different styles (e.g. alt texts, human written prompts) that differs from the distribution of standard benchmark datasets such as MSCOCO. Additionaly, for image-to-text task specifically, OmniFlow is exposed to considerably less image-text pairs (30M) during the training compared with previous generalist models such as CoDi(400M) and UniDiffuser(2B). There is also the question of balancing datasets of different caption qualities. For example, WavCaps is a weakly-labeled dataset, but is 10x larger than higher quality AudioCaps. Additional consideration is required in order to generate captions that can achieve high scores on audiocaps benchmark. Despite these limitations, we show that OmniFlow can generate reasonable image and audio captions through quantitative and qualitative experi-

Reconstruction	GT
Crispy chicken tenders alongside a portion of a bbq sauce.	Crispy chicken tenders alongside a portion of bbq sauce.
A well-furnished living room with a patterned curtain rod, a small white side table holding a vase of flowers, and a tufted gray sofa.	A well-decorated living room with a patterned curtain panel hanging from the window, a small white side table holding a vase of flowers, and a tufted gray sofa.
A young man wearing a black shirt and holding an American flag.	A young man wearing a black shirt and holding an American flag.
An artistic painting of a futuristic city by the water.	An artistic painting of a futuristic city by the water.
Cozy and well-designed living room with a green velvet sofa, glass coffee table displaying potted plants, and a large skylight overhead.	Cozy and stylish living room with a green velvet sofa, glass coffee table displaying potted plants, and a large skylight overhead.
A silver Audi Rs4 sedan driving on the passenger side near a mountainous coastline.	A silver Acura RLX sedan driving on the passenger side near a mountainous coastline.

Table 9. **Text VAE reconstruction results. We show reconstruction results (Left) and the ground truth text (Right).**The reconstruction mostly adheres to the semantics of the ground truth, with minor differences.



OmniFlow (Ours): “A building with lots of plants and trees surrounding it, and a small pond in the middle of the building.”

CoDi: “Beautiful evening villas and small country ornate gardens.”

UniDiffuser: “Beautiful chinese gardens in mountains.”

OmniFlow (Ours): “Watercolor painting of a cat lounging in the water under the sunshine.”

CoDi: “A cat sleepy on beautiful day in the swimming pool.”

UniDiffuser: “Cute cat swimming on water wallpaper.”

OmniFlow (Ours): “A girl in futuristic looking gear is standing in a video game store.”

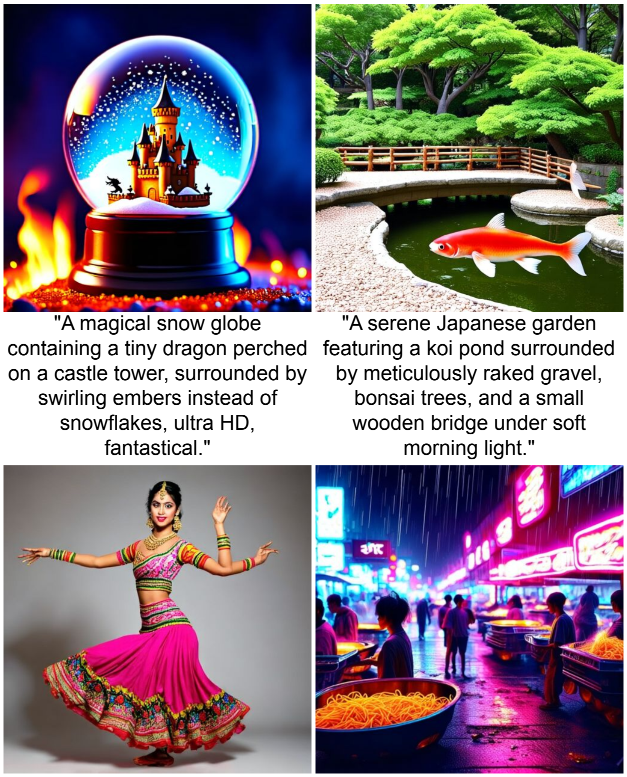
CoDi: “A woman wearing glasses and observing images.”

UniDiffuser: “Anime girl with gaming headset”

Figure 12. **Qualitative comparison of OmniFlow with baselines on image-to-text generation.** Images are provided from the Midjourney Explore page [38].

ments. Our work focuses on develop an effective recipe for any-to-any generalist models. We leave optimizing for text generations to future works.

On Image generation tasks, while OmniFlow can generate high quality images, it has the same limitations as any text-to-image models. For example, it may inherit unintended biases from the training dataset. It may also struggle in prompts that the vanilla SD3 model also struggles with.



"A magical snow globe containing a tiny dragon perched on a castle tower, surrounded by swirling embers instead of snowflakes, ultra HD, fantastical."

"A serene Japanese garden featuring a koi pond surrounded by meticulously raked gravel, bonsai trees, and a small wooden bridge under soft morning light."

"A portrait of a Bollywood dancer mid-spin, wearing a brightly colored lehenga with embroidery, her movements blurred with motion."

"A bustling cyberpunk market filled with vendors selling gadgets and steaming bowls of noodles with rain softly falling under vibrant neon lights, ultra HD, lively atmosphere."

Figure 13. **Examples of failure cases encountered during the text-to-image generation process of OmniFlow.**

G. Miscellaneous

G.1. Reproducibility of CoDi

To accurately reproduce the results of CoDi [46], we follow the weights and instructions as indicated in the i-Code-V3 GitHub repository [2]. However, we encounter reproducibility issues, similar to open issues reported by others, which have remained unresolved [3].

H. Reproducibility Statement

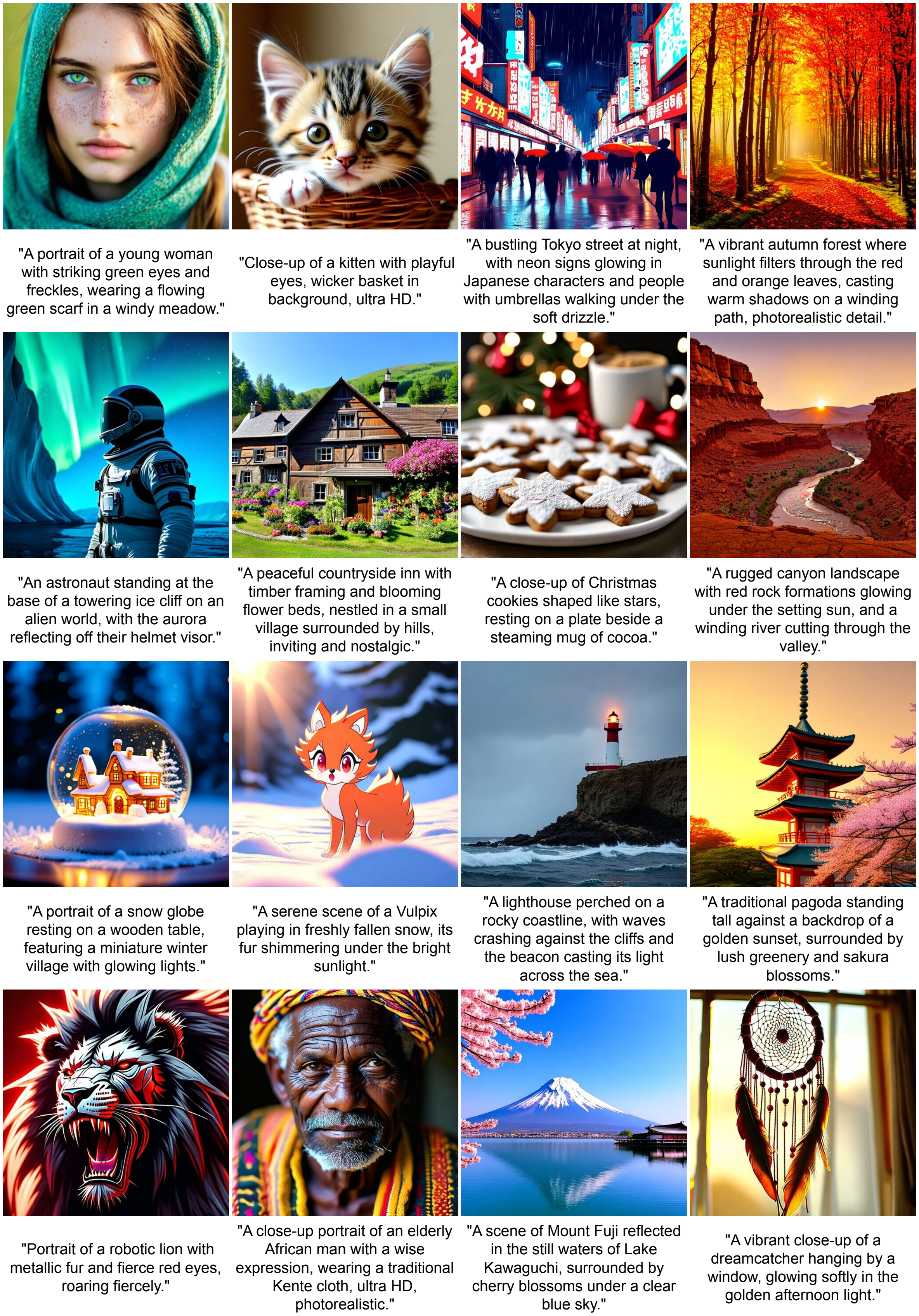
All dataset used in this work are public and accessible from the Internet, except for synthetic captions of SoundNet and VGGSound we generated. We have release the code, checkpoints, and generated captions for these two dataset.

I. Failure Cases

In Fig. 13 we present several failure cases of Omni-Flow when performing text-to-image generation. In the snow globe example, the model fails to interpret the prompt specifying ”swirling embers instead of snowflakes,” mistakenly generating snow instead. Another issue arises with the dancer, where the prompt ”movements blurred with motion” is inaccurately represented as an additional arm. Lastly, the Koi pond and ramen examples highlight unnatural outputs, with the former resembling a poorly edited image of a fish in a pond and the latter depicting oversized bowls of noodles placed unnaturally on the street.

[2] https://github.com/microsoft/i-Code/tree/main/i-Code-V3

[3] https://github.com/microsoft/i-Code/issues/134



"A portrait of a young woman with striking green eyes and freckles, wearing a flowing green scarf in a windy meadow."

"Close-up of a kitten with playful eyes, wicker basket in background, ultra HD."

"A bustling Tokyo street at night, with neon signs glowing in Japanese characters and people with umbrellas walking under the soft drizzle."

"A vibrant autumn forest where sunlight filters through the red and orange leaves, casting warm shadows on a winding path, photorealistic detail."

"An astronaut standing at the base of a towering ice cliff on an alien world, with the aurora reflecting off their helmet visor."

"A peaceful countryside inn with timber framing and blooming flower beds, nestled in a small village surrounded by hills, inviting and nostalgic."

"A close-up of Christmas cookies shaped like stars, resting on a plate beside a steaming mug of cocoa."

"A rugged canyon landscape with red rock formations glowing under the setting sun, and a winding river cutting through the valley."

"A portrait of a snow globe resting on a wooden table, featuring a miniature winter village with glowing lights."

"A serene scene of a Vulpix playing in freshly fallen snow, its fur shimmering under the bright sunlight."

"A lighthouse perched on a rocky coastline, with waves crashing against the cliffs and the beacon casting its light across the sea."

"A traditional pagoda standing tall against a backdrop of a golden sunset, surrounded by lush greenery and sakura blossoms."

"Portrait of a robotic lion with metallic fur and fierce red eyes, roaring fiercely."

"A close-up portrait of an elderly African man with a wise expression, wearing a traditional Kente cloth, ultra HD, photorealistic."

"A scene of Mount Fuji reflected in the still waters of Lake Kawaguchi, surrounded by cherry blossoms under a clear blue sky."

"A vibrant close-up of a dreamcatcher hanging by a window, glowing softly in the golden afternoon light."

Figure 14. **Qualitative examples of the text-to-image capability of OmniFlow.**

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