Appendix

The supplementary material provides additional information:

- Sec. A: More implementation details, including dataset filtering, and FlowTok training hyperparameters.
- Sec. B: Additional qualitative text-to-image and image-to-text generation samples produced by FlowTok.
- Sec. C: Discussions on limitations and future work of FlowTok.

A. More Implementation Details

model	dataset	filtering		
		resolution	aesthetic	watermark
Text Decoder	COCO [12]			
Image Tokenizer	DataComp [10]	✓		
FlowTok: pre-training	DataComp [10]	✓	✓ (5.0)	✓
	CC12M [6]	✓	✓ (5.0)	✓
	LAION-aesthetic [1]	✓		✓
FlowTok: fine-tuning	DataComp† [10]	✓	✓ (6.0)	✓
	LAION-art† [3]	✓		✓
	LAION-pop† [4]	✓		✓
	DALLE3-1M [8]			
	JourneyDB [14]			

Table 1. **Training Data Details.** The filtering criteria applied include resolution (aspect ratio < 2 and longer side ≥ 256), aesthetic score (predicted score exceeding the specified value in parentheses), and watermark detection (removal of images predicted to contain watermarks). †: We use the re-captioned version released by MaskGen [11], which contains improved captions.

Dataset Filtering. In line with previous works [11], we apply three filtering criteria to curate open data for training the image tokenizer and FlowTok: resolution, aesthetic quality, and watermark filtering. The COCO [7, 12] dataset is used directly to train the text decoder without any filtering. Details of the applied filtering criteria are shown in Tab. 1.

Specifically, resolution filtering is applied during the training of the image tokenizer and for text-to-image generation. This ensures that the longer side of each image is at least 256 pixels and the aspect ratio is below 2. For text-to-image training, we further apply aesthetic filtering using the LAION-aesthetic-v2 predictor [2] to retain only high-quality images. Images with aesthetic scores above 5.0 are retained during the pre-training stage, while a stricter threshold of 6.0 is used during fine-tuning to ensure even higher image quality.

Additionally, watermark filtering is implemented for FlowTok's text-to-image generation by using the LAION-WatermarkDetector [5], removing images with watermark probabilities exceeding 0.5. Synthetic datasets such as JourneyDB [14] and DALLE3-1M [8] are exempt from these filtering steps, as they inherently meet our high resolution and quality standards.

Training Hyper-parameters. Tab. 2 provides the complete list of hyper-parameters used for training FlowTok.

hyper-parmeters	pre-training	fine-tuning
optimizer	AdamW	AdamW
optimizer-β_1	0.9	0.9
optimizer-β_2	0.95	0.95
weight decay	0.03	0.03
lr	0.0004	0.0002
lr scheduling	constant	constant
lr warmup steps	10K	0
batch size	4096	4096
training steps	250K	150k

Table 2. **Training Hyper-parameters for FlowTok.**

B. Qualitative Examples of FlowTok

Additional Generation Results. Fig. 1, Fig. 2, and Fig. 3 present additional text-to-image generation results produced by FlowTok, demonstrating its ability to generate diverse, high-fidelity images. Meanwhile, Fig. 4 displays the image-to-text generation results, showcasing FlowTok's capability to produce accurate and descriptive captions.

C. Limitations and Future Work

The primary limitation of FlowTok arises during text-to-image generation. To match the compact dimensionality of image latents (*e.g.*, 16), FlowTok projects CLIP text embeddings into the same low-dimensional latent space. While the text alignment loss helps preserve semantic information, some degree of information loss is inevitable during this projection. Consequently, the alignment between text and generated images may be weaker compared to state-of-the-art models employing cross-attention mechanisms. To address this, one potential solution is to introduce a stronger alignment loss that better retains textual semantics. A more fundamental approach, however, involves increasing the channel dimensionality of image latents by aligning them with vision foundation models [13] during image tokenizer training, as suggested by VA-VAE [15]. This strategy aims to identify an optimal channel dimension for both image and text tokens, achieving a balance between preserving semantic information and maintaining efficiency in training and inference.

Additionally, FlowTok currently utilizes only the vanilla flow matching technique to validate the framework's effectiveness. However, many recent advancements in flow matching, such as logit-normal sampling [9], have not yet been explored in our model. Incorporating these techniques could accelerate convergence and enhance performance.

Finally, FlowTok serves as a starting point for exploring efficient direct evolution between text and image modalities. In the future, we aim to extend to a more general framework that can accommodate a broader range of modalities, supporting additional tasks under the same unified formulation.

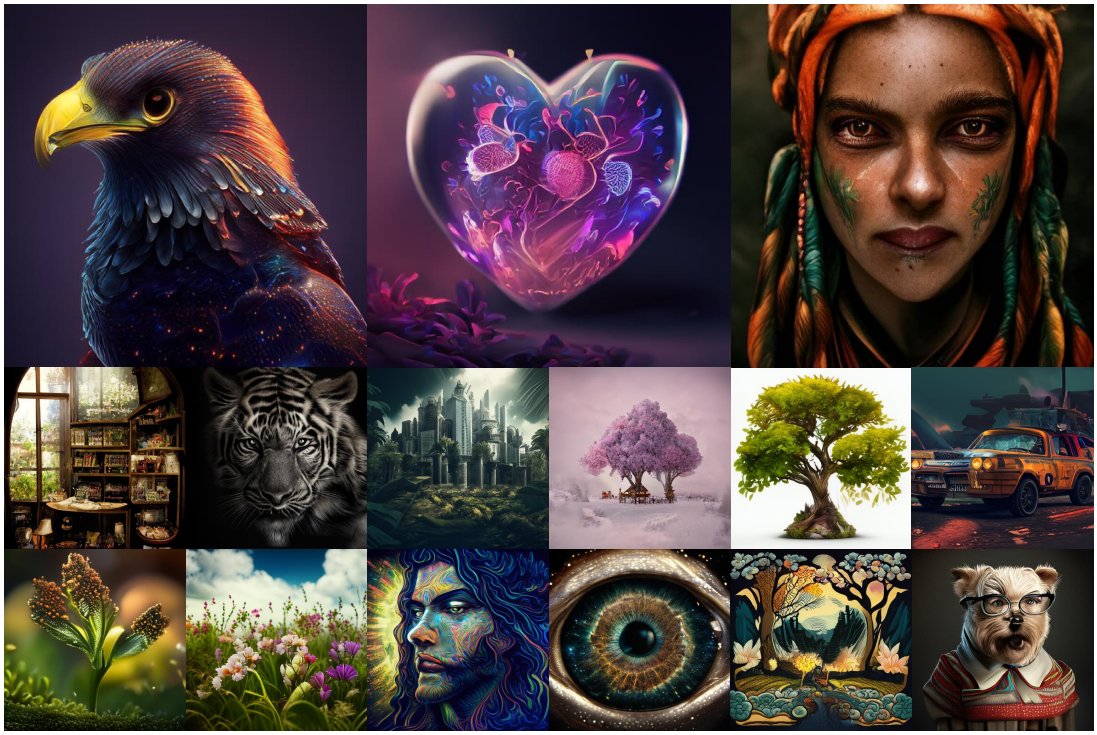

Figure 1. **Text-to-Image Generation Results by FlowTok.** FlowTok generates diverse, high-fidelity images.

Figure 2. **Text-to-Image Generation Results by FlowTok.** FlowTok generates diverse, high-fidelity images.

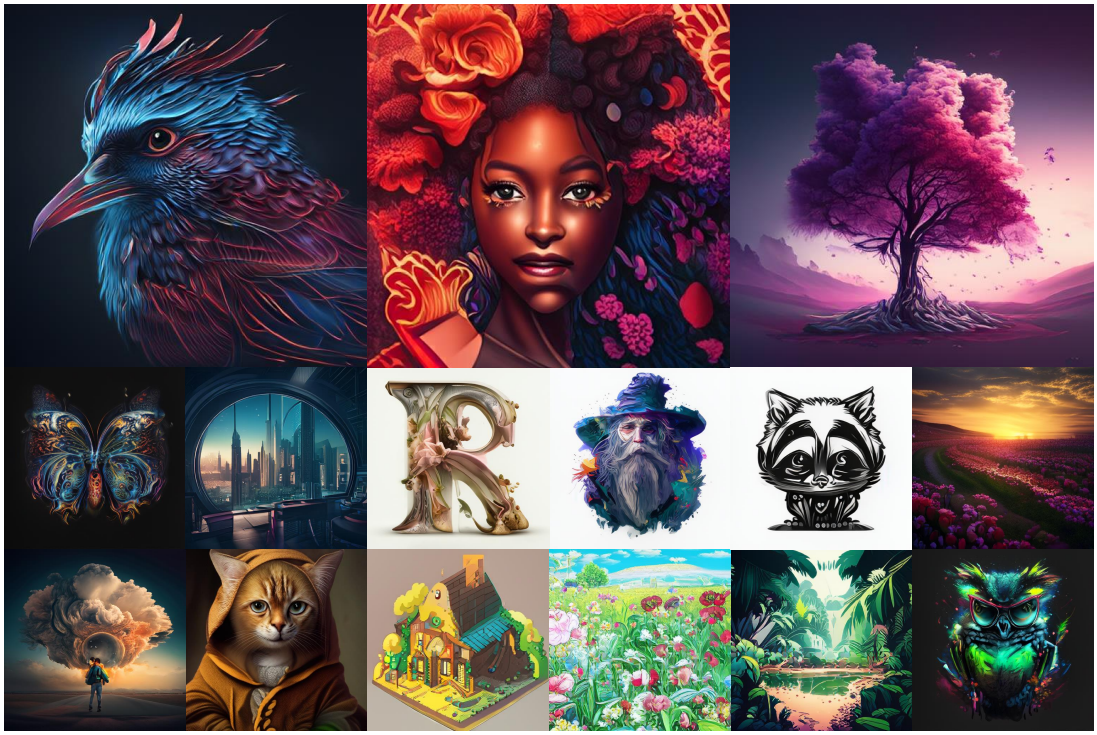

Figure 3. **Text-to-Image Generation Results by FlowTok.** FlowTok generates diverse, high-fidelity images.

Figure 4. **Image-to-Text Generation Results by FlowTok.** FlowTok generates precise captions.

References

[1] LAION2B-en-aesthetic. https://huggingface.co/datasets/laion/laion2B-en-aesthetic, . 1

[2] LAION-aesthetics predictor V2. https://github.com/christophschuhmann/improved-aesthetic-predictor, . 1

[3] LAION-art. https://huggingface.co/datasets/laion/laion-art, . 1

[4] LAION-pop. https://huggingface.co/datasets/laion/laion-pop, . 1

[5] LAION-5B-WatermarkDetection. https://github.com/LAION-AI/LAION-5B-WatermarkDetection, . 1

[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. 1

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[8] Ben Egan, Alex Redden, XWAVE, and SilentAntagonist. Dalle3 1 Million+ High Quality Captions. https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions, 2024. 1

[9] 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 *ICML*, 2024. 1

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *NeurIPS*, 2023. 1

[11] Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *arXiv preprint arXiv:2501.07730*, 2025. 1

[12] 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 *ECCV*, 2014. 1

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[14] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *NeurIPS*, 2023. 1

[15] Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. 1