Democratizing Text-to-Image Masked Generative Models with Compact Text-Aware One-Dimensional Tokens

Supplementary Material

Appendix

We provide additional information in the supplementary material, as outlined below:

- Sec. A: Further implementation details, including dataset filtering, recaptioning, and MaskGen training hyperparameters.
- Sec. B: Ablation studies on the type and place of text guidance in TA-TiTok.
- Sec. C: Ablation studies on the number of tokens and aesthetic score condition in MaskGen.
- Sec. D: Comparisons between MaskGen using discrete and continuous tokens.
- Sec. E: Additional zero-shot text-to-image generation results on COCO validation set.
- Sec. F: Results demonstrating the performance of our KL variant of TiTok on ImageNet.
- Sec. G: More qualitative examples for latent code swapping with TA-TiTok.
- Sec. H: More qualitative examples generated by MaskGen.
- Sec. I: Discussions on limitations and future work of TA-TiTok and MaskGen.

A. More Implementation Details

Dataset Filtering. To prepare training data, we applied three filtering criteria: resolution, aesthetic, and watermark filtering. Details of applied filtering criteria and the total size of the dataset after filtering are presented in Tab. 7. Resolution filtering was applied to all datasets during the training of both TA-TiTok and MaskGen. This filtering ensured that the longer side of each image exceeded 256 pixels and maintained an aspect ratio below 2. For MaskGen training, we implemented aesthetic filtering using the LAION-aesthetic-v2 predictor [2] to retain only high-quality images. Images with scores above 5.0 were kept during the pre-training stage, while a stricter threshold of 6.0 was applied during fine-tuning to ensure even higher quality. Additionally, we employed watermark filtering for MaskGen using the LAION-WatermarkDetector [5], removing images with watermark probability exceeding 0.5 to prevent unwanted watermark artifacts in generated images. Synthetic datasets such as JourneyDB [24] and DALLE3-1M [14] were exempted from these filtering processes as they inherently met our high resolution and quality standards.

Dataset Recaptioning. To improve the text quality of DataComp [15], LAION-art [3], and LAION-pop [4], we utilize state-of-the-art vision-language models, Molmo-7B-D-0924 [11], to enhance captions based on both the image and its original caption. Specifically, we randomly sample one of four prompts as shown in Fig. 5 to generate updated captions. Since the augmented captions are often significantly longer than the original ones and frequently start with similar patterns (*e.g.*, "The image depicts/displays/showcases/shows/features..."), we apply prefix filtering to remove these repetitive prefixes, preventing information leakage. During training, we further address this by employing a 95:5 ratio, randomly sampling between augmented and original captions to ensure balanced learning, following [6]. A few recaption results are shown in Fig. 6, highlighting how the augmented captions provide richer details and align better with the image content.

Training. We adhere strictly to the hyperparameters used to train TiTok across all TA-TiTok variants. Specifically, TA-TiTok is trained with a batch size of 1024 for 1 epoch (650k steps) on the filtered DataComp dataset, using a maximum learning rate of 1e-4 and a cosine learning rate schedule. For MaskGen with discrete tokens, we employ a batch size of 4096, leveraging weight tying [22] to stabilize training, with a cosine learning rate schedule and a maximum learning rate of 4×10^{-4}. For MaskGen with continuous tokens, to accommodate the diffusion loss, we use a constant learning rate schedule with a maximum rate of 1×10^{-4} and a batch size of 2048. Masked tokens are sampled by randomly selecting the masking rate from [0, 1] on a cosine schedule, following MaskGIT [7], and text conditioning is randomly dropped with a 0.1 probability to enable classifier-free guidance [17]. Tab. 8 provides the complete list of hyper-parameters used for training MaskGen with both discrete and continuous tokenizers.

B. Ablation Studies for TA-TiTok

Text Guidance Type in TA-TiTok. In our text-aware de-tokenization design, we can use either the numerical IDs from the CLIP text tokenizer or the embeddings from the CLIP text encoder. We ablate this design choice in Tab. 9, where the latter option yields a marginal improvement.

Text Guidance Place in TA-TiTok In the design of TA-TiTok, we only incorporate the text guidance (*i.e.*, the text tokens from CLIP) into the tokenizer decoder to better capture high-level semantics and align with textual descriptions during both reconstruction and generation. In this study, we investigate whether injecting text guidance into both the encoder and decoder of TA-TiTok can further enhance the quality of the encoded latent tokens. This evaluation is

model	dataset	filtering			recaptioning	samples
		resolution	aesthetic	watermark		
TA-TiTok: tokenizer	DataComp [15]	✓				685.8M
MaskGen: pre-training	DataComp [15]	✓	✓ (5.0)	✓		219.8M
	CC12M [9]	✓	✓ (5.0)	✓		4.8M
	LAION-aesthetic [1]	✓		✓		28.3M
MaskGen: fine-tuning	DataComp [15]	✓	✓ (6.0)	✓	✓	3.6M
	LAION-art [3]	✓		✓	✓	4.2M
	LAION-pop [4]	✓		✓	✓	0.4M
	DALLE3-1M [14]					1.0M
	JourneyDB [24]					4.1M

Table 7. **Training Data Details.** Filtering criteria applied to each publicly available dataset include resolution (aspect ratio < 2 and longer side $\geq$ 256), aesthetic score (predicted score exceeding the specified value in parentheses), and watermark detection (removal of images predicted to contain watermarks). For datasets with noisy web-crawled captions, Molmo [11] is used for recaptioning. The final column shows the number of text-image pairs remaining after filtering.

```
1. Describe the image in detail while
 considering the provided caption:
 '{original_caption}'. Correct any errors
 and improve the caption, ensuring the final
 description is in English and within 77
 tokens. Return only the corrected caption.

2. Analyze the image and the caption
 '{original_caption}'. Write a detailed and
 accurate description of the image in
 English, correcting any mistakes or
 low-quality aspects of the original
 caption. Keep the final caption under 77
 tokens, and return only the caption.

3. Use the caption '{original_caption}' as a
 reference to create a detailed and improved
 description of the image in English.
 Correct any errors and make sure the new
 caption is concise and within 77 tokens.
 Return only the revised caption.

4. Given the image and the original caption
 '{original_caption}', describe the image in
 a detailed and accurate way in English,
 improving upon the original caption where
 necessary. Ensure the description is within
 77 tokens. Return only the corrected caption.
```

Figure 5. **Prompts Used for Recaptioning.** One of four prompts is used to recaption each image, where `{original_caption}` is replaced with the original image caption.

conducted on the ImageNet validation set [12] using reconstruction metrics in a zero-shot setting, where the caption is represented as "A photo of *class*" without any prompt engineering. As shown in Tab. 10, injecting textual guidance into both the encoder and decoder has negligible impact on reconstruction quality. This finding suggests that incorporating text guidance in the decoder alone is sufficient to provide semantic information to the model.

hyper-parmeters	discrete	continuous
optimizer	AdamW	AdamW
β_1	0.9	0.9
β_2	0.96	0.95
weight decay	0.03	0.02
lr (pre-training)	0.0004	0.0002
lr (fine-tuning)	0.0001	0.0002
lr scheduling	cosine	constant
lr warmup steps	10K	50k
batch size	4096	2048
training steps (pre-training)	500K	1000k
training steps (fine-tuning)	250K	500k

Table 8. **Training Hyper-parameters for MaskGen.**

tokenizer	arch	#tokens	text guidance	rFID↓	IS↑
TA-TiTok	KL	32	ID	1.62	213.6
			Embedding	1.53	222.0

Table 9. **Ablation on Text Guidance Type.** Models are trained on DataComp and zero-shot evaulated on ImageNet validation set. ID refers to numerical IDs extracted by CLIP text tokenizer, Embedding denotes text features extracted by CLIP text encoder.

arch	tokens		Encoder + Decoder		Decoder Only	
	#	c	rFID↓	IS↑	rFID↓	IS↑
KL	32	16	1.65	218.4	1.53	222.0
	64	16	1.39	221.5	1.47	220.7
	128	16	0.92	227.1	0.90	227.7

Table 10. **Ablation on Text Guidance Place.** In the TA-TiTok design, we ablate on adding the text guidance to both encoder and decoder or just decoder. Adding text guidance to only the decoder results in similar reconstruction performances but enjoys a simpler structure. Models are trained on DataComp and zero-shot evaulated on ImageNet validation set.

C. Ablation Studies for MaskGen

Experimental Setup. For efficient ablation studies, we use the discrete version of MaskGen to analyze the impact of token count. Performance is evaluated using the FID metric on MJHQ-30K [19] and the overall GenEval [16] score.

Original caption: Pretty Dress with Matching Hat 3-6 Months Old

New caption: A charming white dress with red polka dot trim, featuring a large bow at the waist and smaller bows on the shoulders. The dress is adorned with a delightful pattern of blue and pink teacups, butterflies, and flowers. A matching bonnet with a red bow complements the outfit. The ensemble is displayed on a colorful wooden surface, with wooden puzzle pieces showing the numbers 3, 6, and M nearby.

Original caption: J.L. Jensen: Kamelier og rhododendron - Fineart photography by Art Classics

New caption: J.L. Jensen's "Kamelier og rhododendron" (Carnations and Rhododendrons) is a stunning fine art photograph showcasing a lush floral arrangement. The image features vibrant carnations in shades of pink and red, accompanied by elegant rhododendron blooms in white and pink. Dark green leaves provide a striking contrast against the colorful flowers. The composition is set against a deep black background, enhancing the vividness of the blooms. This exquisite piece captures the beauty and diversity of these two iconic flower species in a visually captivating manner.

Figure 6. **Re-captioning Results.** Captions augmented by Molmo [11] offer richer details and improved alignment with image content.

arch	generator	#tokens	T↓	I↑	MJHQ-30K FID↓	GenEval Overall↑
VQ	MaskGen-L	32	16.0	47.6	9.11	0.43
		64	17.5	40.2	7.85	0.50
		128	20.0	30.3	7.74	0.53

Table 11. **Zero-Shot Text-to-Image Generation Results on MJHQ-30K and GenEval with Varying Number of Tokens.** MaskGen achieves better generation quality with more tokens but incurs longer training times and slower inference speeds. T: Generator training cost, measured in 8 A100 days using float16 precision. I: Generator inference throughput, measured in samples per second on a single A100 with batch size 64 using float16 precision.

tokenizer	arch	generator	#params	open-data	FID-30K↓
MAGVIT-v2 [25]	VQ	Show-o [25]	1.3B	✓	9.24
TA-TiTok	VQ	MaskGen-L (ours)	568M	✓	13.62
TA-TiTok	VQ	MaskGen-XL (ours)	1.1B	✓	13.01
VAE [23]	KL	LDM [23]	1.4B	✓	12.64
VAE [23]	KL	Stable-Diffusion-1.5 [23]	860M	✓	9.62
VAE [23]	KL	PixArt-α [10]	630M	✗	7.32
TA-TiTok	KL	MaskGen-L (ours)	568M + 44M	✓	9.66
TA-TiTok	KL	MaskGen-XL (ours)	1.1B + 69M	✓	8.98

Table 12. **Zero-Shot Text-to-Image Generation Results on COCO-30K.** Comparison of MaskGen with state-of-the-art *open-weight* models.

tokenizer	arch	tokens #	tokens c	rFID↓	generator	gFID↓	IS↑	T↓	I↑
VAE [20]	KL	256	16	0.54	MAR [20]	2.45	275.5	8.0	1.0
TiTok (ours)	KL	64	16	1.54	MAR [20]	2.96	246.9	2.1	8.1
		128	16	1.31		2.70	252.9	3.2	3.2

Table 13. **Class-conditional ImageNet-1K** 256×256 **Generation Results Evaluated with ADM [13], using continuous tokens (*i.e.*, KL architecture).** #: Number of tokens. c: Channels of continuous tokens. T: Generator training cost, measured in 8 A100 days using float32 precision. I: Generator inference throughput, measured in samples per second on a single A100 with float32 precision.

Additionally, we provide visual comparisons to illustrate the effect of aesthetic score conditioning during sampling.

Number of Tokens. Tab. 11 presents an ablation study on the number of tokens used for text-to-image generation with MaskGen. As observed, increasing the token count improves generation quality but comes at the expense of longer training times and slower inference speeds.

Aesthetic Score Conditioning. Fig. 7 visualizes images generated with different aesthetic scores while keeping other hyperparameters and prompts constant. The results indicate a strong correlation between higher aesthetic scores and enhanced dramatic lighting and fine-grained details. For instance, in the third row, a higher aesthetic score yields richer depictions of trees and stars in the night sky, whereas a lower score results in simpler representations. This enables precise control over image generation based on user preferences.

D. Comparisons Between MaskGen Using Discrete and Continuous Tokens

Performance Comparisons Between VQ and KL Variants. The KL variant of MaskGen consistently outperforms the VQ variant on MJHQ-30K FID but performs slightly worse on GenEval's overall score. We hypothesize that the KL variant excels in generating diverse, high-aesthetic images, contributing to improved FID on MJHQ-30K. However, it falls behind on GenEval, which emphasizes object-focused compositional properties such as position, count, and color. In contrast, the VQ variant, constrained by a finite codebook, generates less diverse but more compositionally accurate images, leading to higher scores on GenEval. Fig. 9 visually compares generated samples, where the KL variant demonstrates slightly better overall generation quality.

Training and Inference Cost Comparisons Between VQ and KL Variants. The VQ variant of MaskGen benefits from faster training and significantly faster inference, primarily due to inherent differences in the diffusion process used in the KL variant. While the KL variant excels in generating more diverse and higher-aesthetic images, this advantage comes with increased computational demands. To address this gap in training and inference efficiency, we employ 128 tokens for the VQ variant and 32 tokens for the KL variant, effectively controlling the training cost to remain at a comparable level, as shown in Tab. 4 of the main paper.

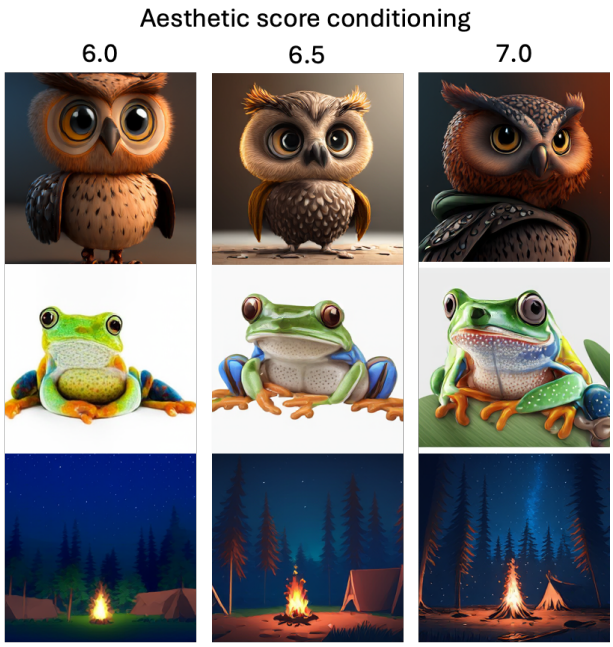


Figure 7. **Generated Images with Varying Aesthetic Score Conditioning.** Conditioning on higher aesthetic scores produces generated images with enhanced fine-grained details.

E. Zero-Shot Text-to-Image Generation Results on COCO

In Tab. 12, we evaluate zero-shot text-to-image generation on the COCO dataset [21] by randomly sampling 30K image-caption pairs from the COCO 2014 validation split and reporting the FID, as is standard in the literature. Since the fine-tuning stage of MaskGen often generates more aesthetically appealing images that deviate from the COCO dataset distribution, we perform the evaluation using MaskGen at the pre-training stage to ensure consistency with the dataset's characteristics. Notably, MaskGen-L (KL variant with continuous tokens) achieves an FID-30K of 9.66, while MaskGen-XL (KL variant) further improves to 8.98. These results demonstrate that MaskGen achieves performance comparable to other state-of-the-art text-to-image models, highlighting its effectiveness even in the zero-shot setting.

F. KL variant of TiTok on ImageNet

We evaluate the KL variant of TiTok as a drop-in replacement for standard 2D VAEs [18, 20] in class-conditional image generation on ImageNet [12]. Results, reported in Tab. 13, are based on the MAR [20] framework with its base model, after 400 epochs using unchanged MAR hyper-parameters. MAR with TiTok (our KL variant) achieves significant training time reductions (3.8× with 64 tokens, 2.7× with 128 tokens) and inference speedups (8.1× with 64 tokens, 3.2× with 128 tokens), thanks to its efficient 1D token design. Despite the substantial reduction in computational overhead, MAR with TiTok maintains performance comparable to MAR with conventional 2D VAEs using 256 tokens, highlighting TiTok's potential as an efficient and robust image tokenizer for class-conditional generation.

G. Qualitative Examples of Latent Code Swapping

Fig. 8 provides additional visualizations of latent code swapping in VQ variants of TA-TiTok. The results demonstrate that our proposed 1D tokenization encodes images such that each token captures meaningful semantic elements, enabling image manipulation through latent token swapping without even requiring the generator. This semantically rich tokenization explains why 1D tokenization achieves far more efficient encoding than its 2D counterpart while preserving competitive reconstruction fidelity: the encoder dynamically allocates tokens to perceptually informative regions in the image.

H. Qualitative Examples of MaskGen

Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, and Fig. 16 showcase additional qualitative examples of text-to-image generation using MaskGen. By utilizing the efficient and compact text-aware tokenizer TA-TiTok, MaskGen demonstrates its ability to produce high-fidelity and diverse images.

I. Limitations and Future Work

While MaskGen achieves competitive generation quality and benchmark scores comparable to recent text-to-image models, including those leveraging proprietary training data, we acknowledge several aspects for future exploration.

First, the current KL variant of MaskGen is designed to use 32 tokens. While increasing the token count improves tokenization quality, leading to better-reconstructed samples, it also significantly raises training costs due to longer convergence times. Additionally, scaling up the generator remains a challenge, as the current MaskGen-XL is constrained to 1.1B parameters due to limited computational resources.

Second, the current implementation of MaskGen operates at a resolution of 256×256. However, the scalability of its core architectural design—1-dimensional tokenization and masked generation—has been validated in high-resolution implementations like Muse [8].

This work emphasizes establishing a fully open-source, open-data text-to-image masked generative model using compact text-aware 1-dimensional tokenization. Future work will focus on optimizing convergence speed, model scaling up, and enabling high-resolution outputs.

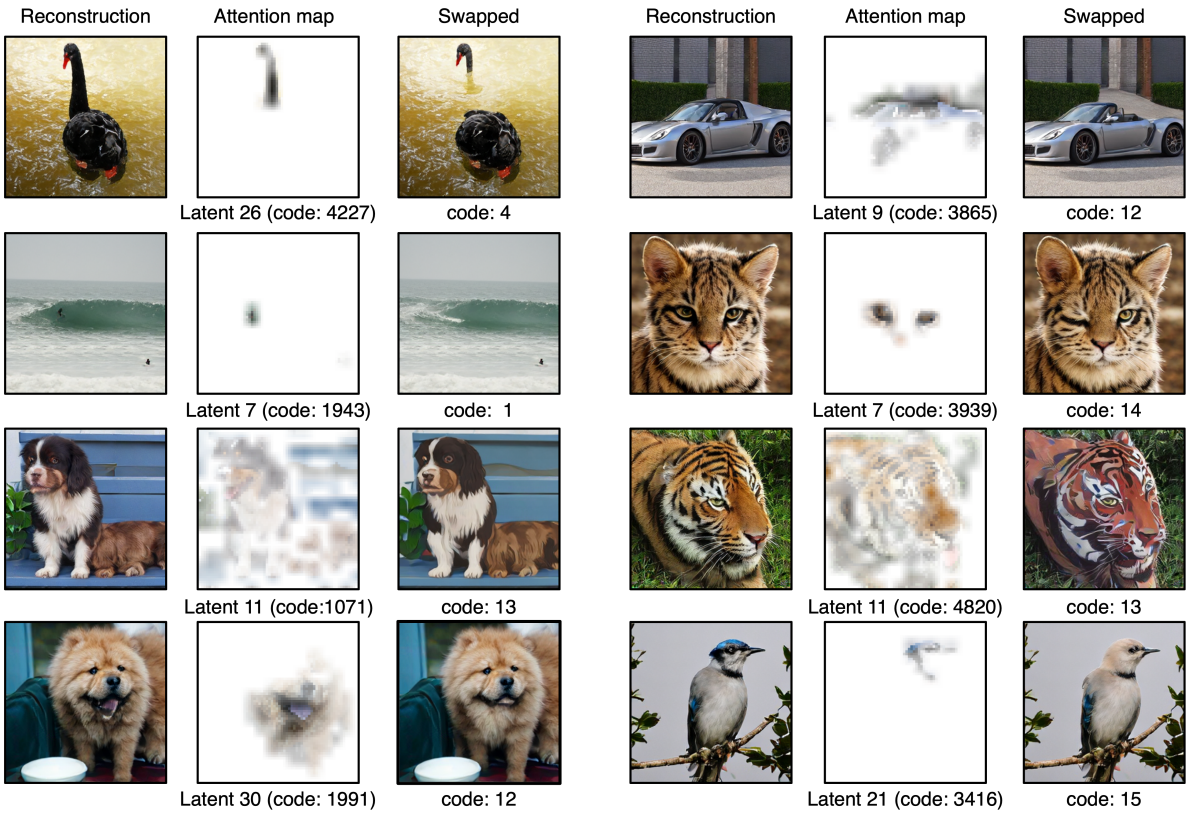


Figure 8. **Visualization of Latent Token Attention Map and Latent Code Swapping.** The results are from VQ variant of TA-TiTok with 32 tokens. Each latent token attends to prominent semantic and swapping the code leads to appearance changes in the corresponding semantic entity that the latent token focuses on.

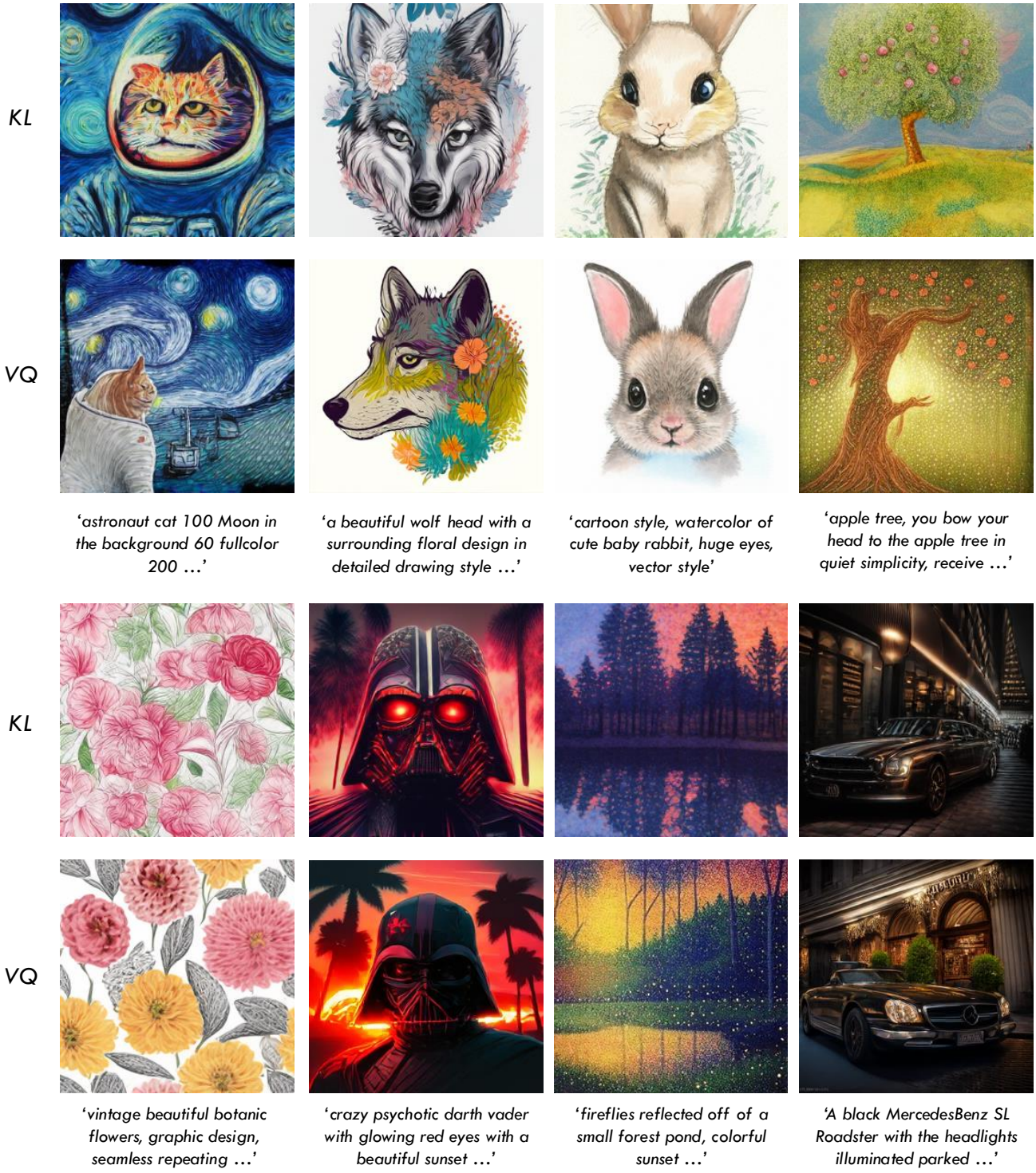


Figure 9. **Generated Images by MaskGen with Different Tokenizer Types.** For each caption, the top row displays images generated using continuous tokens (KL), while the bottom row shows images generated using discrete tokens (VQ). Long prompts are truncated for brevity.

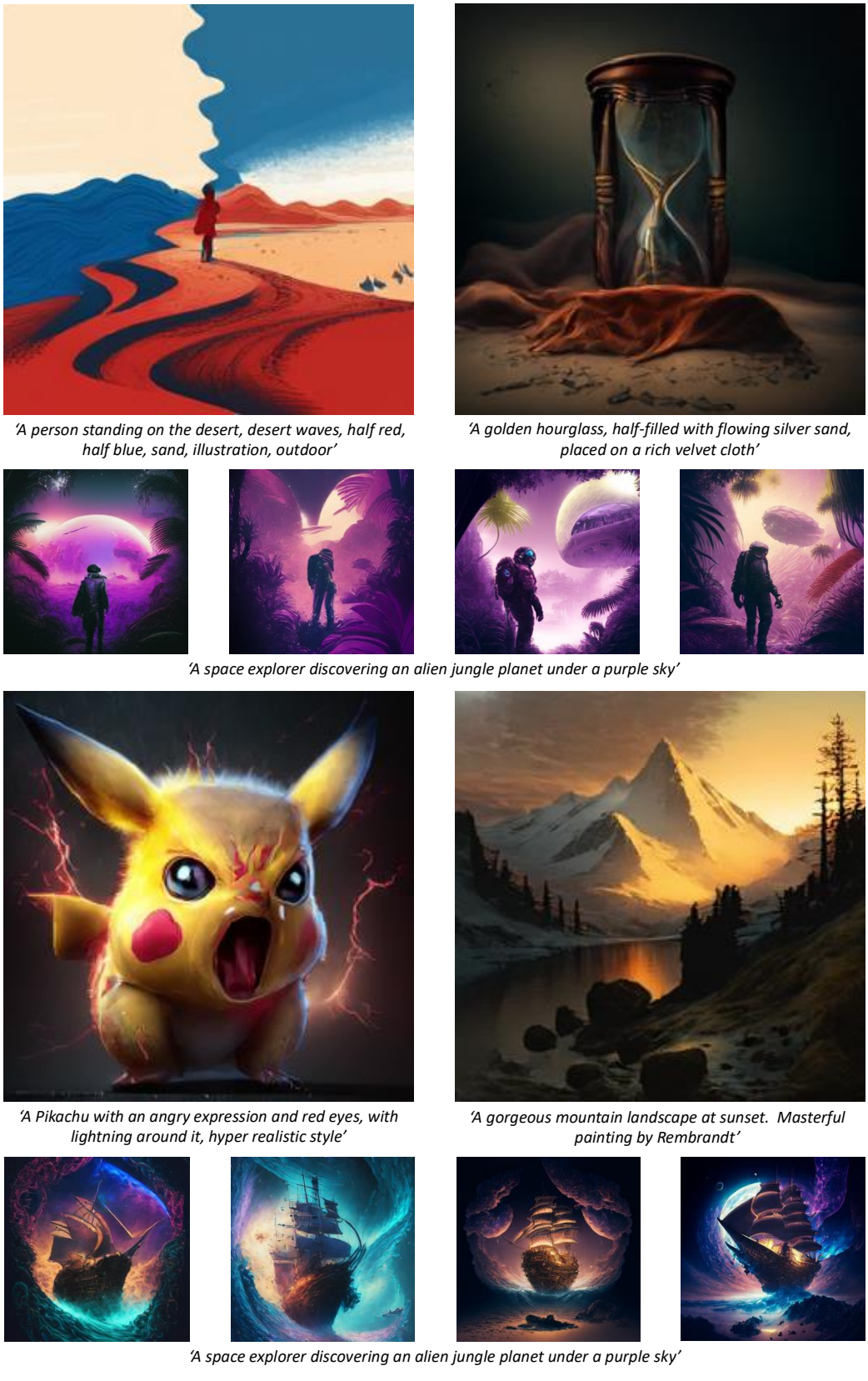


Figure 10. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

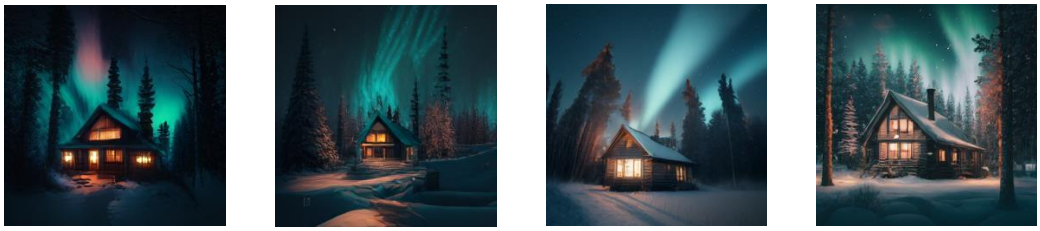
'A cloud dragon flying over mountains, its body swirling with the wind'

'A window with raindrops trickling down, overlooking a blurry city'

'A mountain village built into the cliffs of a canyon, where bridges connect houses carved into rock, and waterfalls flow down into the valley below'

'A vintage typewriter with paper spewing out like a waterfall'

'A medieval knight standing on a cliff overlooking a vast battlefield'



'A cozy cabin in the middle of a snowy forest, surrounded by tall trees with lights glowing through the windows, a northern lights display visible in the sky'

Figure 11. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

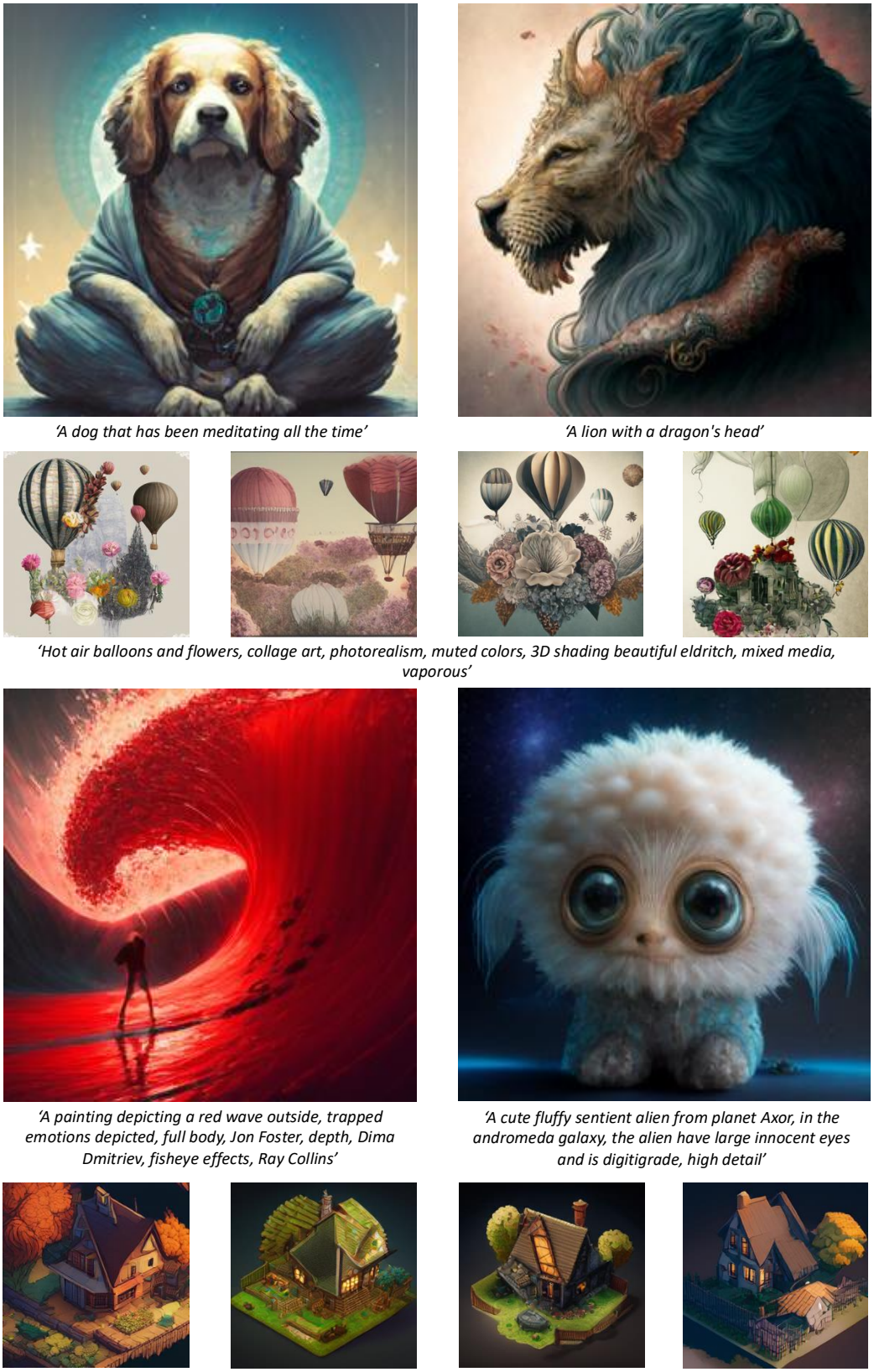


Figure 12. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

'A snowy mountain'

'A man looks up at the starry sky, lonely'

'A still life of a vase overflowing with vibrant flowers, painted in bold colors and textured brushstrokes, reminiscent of van Gogh's iconic style'

'A deep forest clearing with a mirrored pond reflecting a galaxy-filled night sky'

'knolling of a drawing tools and books, knowledge, white background'

'A vibrant yellow banana-shaped couch sits in a cozy living room, its curve cradling a pile of colorful cushions. on the wooden floor, a patterned rug adds a touch of eclectic charm, and a potted plant sits in the corner, reaching towards the sunlight filtering through the window'

Figure 13. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

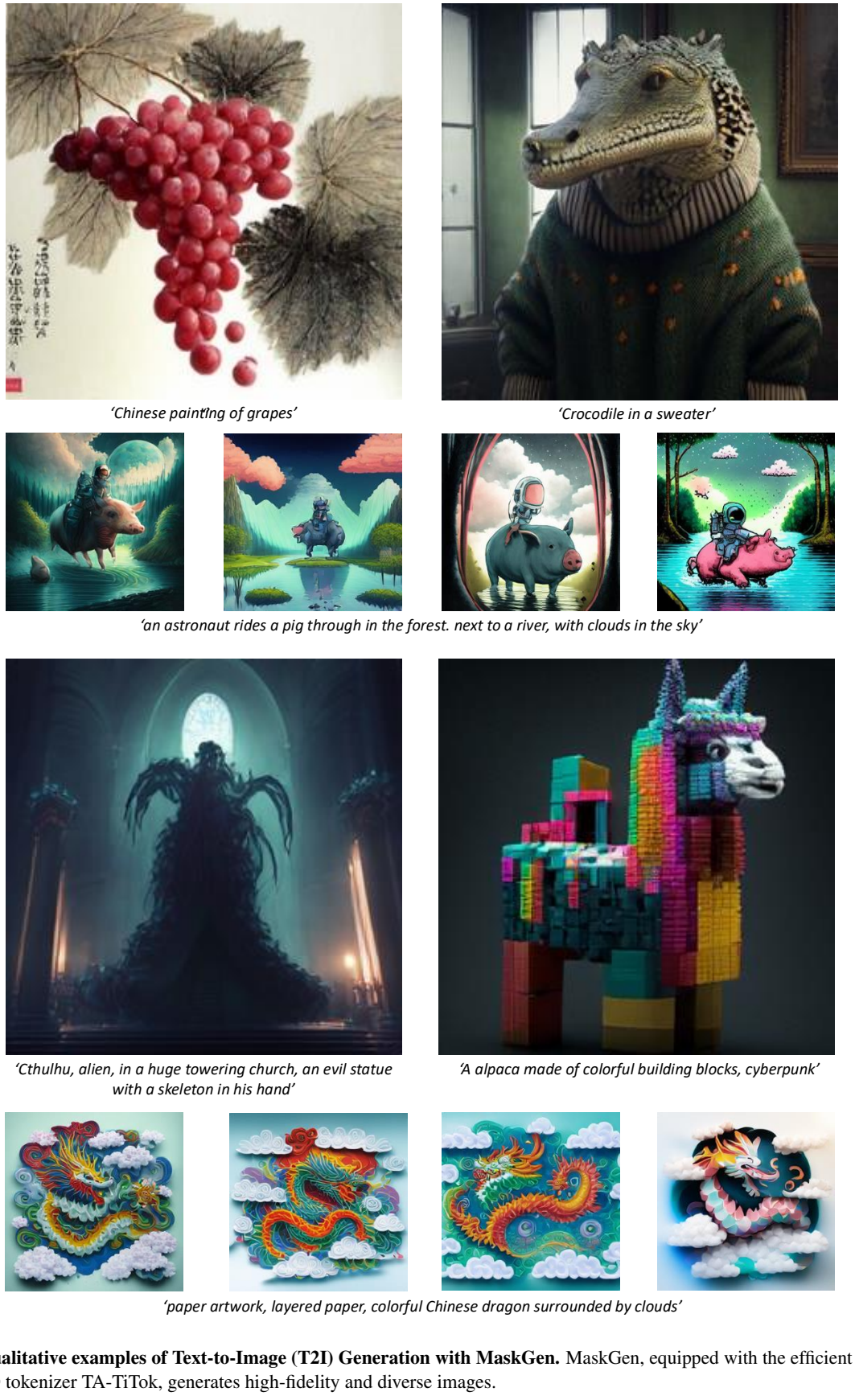


Figure 14. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

'An image of a chrome sphere reflecting a vibrant city skyline at sunset'

'how world looks like in 100 A car made out of vegetables years, intricate, detailed'

'Chinese architecture, ancient style,mountain, bird, lotus, pond, big tree, Unity, octane rendering'

'stars, water, brilliantly, gorgeous large scale scene, a little girl, in the style of dreamy realism'

'a Emu, focused yet playful, ready for a competitive matchup, photorealistic quality with cartoon vibes'

'a traveler navigating via a boat in countless mountains, Chinese ink painting'

Figure 15. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

'an illustration of an stylish swordsma'

'Futurist painting of the building'

'beautiful scene with mountains and rivers in a small village'

'A tranquil scene of a Japanese garden with a koi pond, painted in delicate brushstrokes and a harmonious blend of warm and cool colors'

'A dark forest under a full moon, with twisted, gnarled trees, shadows lurking behind every branch, and a lone figure holding a glowing lantern'

'A silhouette of a grand piano overlooking a dusky cityscape viewed from a top-floor penthouse, rendered in the bold and vivid style of a vintage travel poster'

Figure 16. **Qualitative examples of Text-to-Image (T2I) Generation with MaskGen.** MaskGen, equipped with the efficient and compact text-aware 1D tokenizer TA-TiTok, generates high-fidelity and diverse images.

References

[1] LAION2B-en-aesthetic. https://huggingface.co/datasets/laion/laion2B-en-aesthetic, . 2

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

[3] LAION-art. https://huggingface.co/datasets/laion/laion-art, . 1, 2

[4] LAION-pop. https://huggingface.co/datasets/laion/laion-pop, . 1, 2

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

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 1

[8] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 4

[9] 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. 2

[10] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 3

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 2, 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 4

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 3

[14] 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, 2

[15] 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, 2

[16] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023. 2

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[19] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 2

[20] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024. 3, 4

[21] 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. 4

[22] Ofir Press and Lior Wolf. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*, 2016. 1

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[24] 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, 2

[25] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 3