TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation

Supplementary Material

A. Implementation Details

A.1. Motivation

Experimental Setup for Multimodal Understanding. To evaluate the multimodal understanding capabilities of current VQ tokenizers, we conduct experiments as detailed in Tab. 1. For LFQ [66], we utilize the open-source implementation [33], which demonstrates comparable performance to the original paper. The codebook size of LFQ is 262,144. For VQGAN-LC [76], we employ features before its projection layer, which is clustered from the pretrained CLIP image encoder, with a codebook size of 100,000.

Experimental Setup for Visual Comparison of VQKD, VQGAN and TokenFlow. To generate the visualizations in Fig. 4, we perform an experiment using 50,000 images from the ImageNet-1k validation set. We process these images through the encoders of VQKD, VQGAN and TokenFlow, applying average pooling to the extracted features to obtain a 1×1 representation. Subsequently, we identify the closest index in their respective codebooks using l_2 distance. We provide more visualizations in Fig. 11, and visualize the cluster size distribution in Fig. 7.

Experimental Setup for Image Reconstruction from Quantized Semantic Feature. We conducted an experiment to reconstruct original images from quantized features extracted by VQKD [35]. In this setup, we maintained the original encoder and quantizer of VQKD, while introducing an additional decoder aimed at reconstructing the input image. The architecture of this decoder is identical to the pixel decoder employed in our TokenFlow. We trained this decoder on the ImageNet-1K dataset for 100 epochs. Fig. 9 presents a visual comparison between the original and the reconstructed images. As observed, while the reconstructed images maintain the overall semantic content, they exhibit a noticeable loss of high-frequency details. This phenomenon suggests that the quantized semantic features cannot fully preserve fine-grained visual details, which is crucial for visual generation.

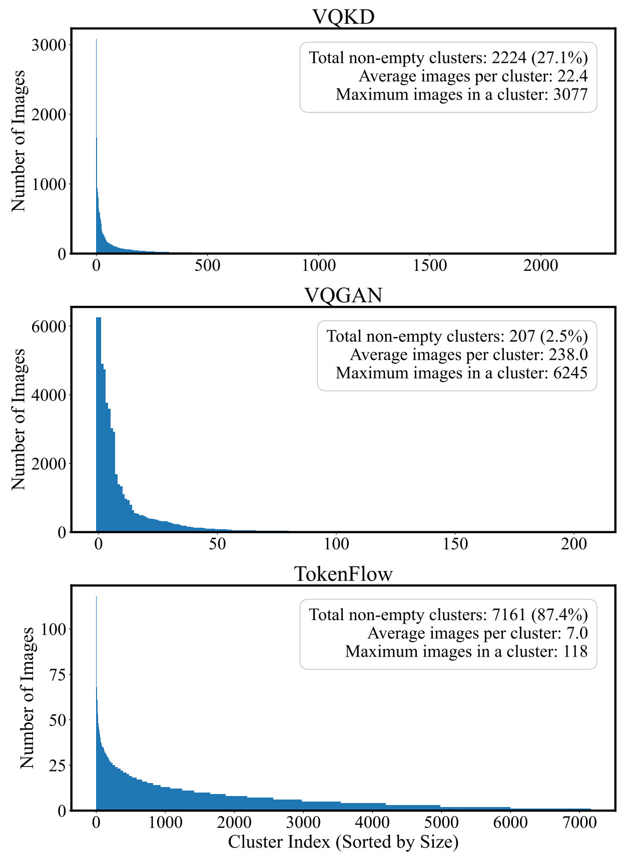


Figure 7. Comparison of cluster size distributions between VQKD [35], VQGAN [13], and TokenFlow (ours), with a fixed codebook size of 8,192. Analysis performed on 50,000 images from the ImageNet-1k validation set. TokenFlow exhibits significantly smoother distribution compared to others, attributed to our shared mapping design that learns joint distributions of semantic and pixel-level features. This joint learning approach helps maintain high codebook utilization (95%+) even with large-scale codebooks containing over 131K entries.

A.2. Tokenizer Training Details

We provide detailed training configurations for TokenFlow-B, TokenFlow-L, and TokenFlow-XL variants in Tab. 11. All models share common hyperparameters including learning rate, batch size, commitment loss factor, adversarial loss factor and distance balance weight. The models primarily differ in their input resolution (224, 256, and 384) and semantic teacher models, utilizing CLIP ViT-B/14 [37], ViTamin-XL [8], and SigLIP-SO400M [69].

B. Additional Results

B.1. Additional Ablation Study

Effect of Sampling Strategy to Visual Generation. We conduct comprehensive ablation studies to analyze the im-

Figure 8. Qualitative comparison of visual generation capabilities between 1B and 7B models. Prompts (from left to right): (1) "A pizza sitting on top of a wooden cutting board", (2) "Television set being held by a hand", (3) "The guy is nicely dressed in a suit and tie", and (4) "A sailing ship rests on waters". The 7B model demonstrates enhanced quality compared to its 1B counterpart.

pact of different sampling strategies on generation quality. As shown in Table 6, we evaluate various configurations using GenEval [15] and ImageReward [63] metrics. We choose ImageReward for ablation due to its strong correlation with human preferences, particularly in capturing local artifacts and overall visual quality. The ImageReward is average over 10k prompts from the MS-COCO validation set. For multi-step configurations, we denote the top-p and top-k values for each step using bracket notation $[x_1, ..., x_n]$.

Our multi-step approach with a two-step strategy (top-k=[1200, 1], top-p=[0.8, 0]) significantly improves generation quality, yielding gains of +0.039 in GenEval and +0.084 in ImageReward compared to single-step sampling. This validates our hypothesis that progressive refinement helps maintain global consistency. When increasing the second-step k value to 10 or 100 while maintaining top-p, we observe slightly degraded performance. This degradation suggests that excessive sampling freedom in refinement steps can lead to increased artifacts and local inconsistencies.

Most notably, three-step strategy (top-k=[1200, 100, 1], top-p=[0.8, 0.8, 0]) achieves the best performance across both metrics. This represents substantial improvements of 10.2% and 14.3% over traditional single-step sampling, respectively. The gradual narrowing of sampling space (1200→100→1) strikes a balance between generation diversity and local consistency. As illustrated in Figure 5, our multi-step approach produces more coherent and visually appealing results. These quantitative and qualitative results demonstrates that progressive refinement in top-p top-k sampling is crucial for high-quality generation in next-scale prediction frameworks.

Effect of Model Size to Visual Generation. We conduct ablation studies to investigate the impact of model size on our decoder-only visual generation architecture.

Table 6. Impact of sampling strategy to visual generation. We compare single-step *v.s.* multi-step sampling strategy using GenEval and ImageReward. For multi-step approaches, values in brackets indicate parameters for successive sampling steps.

Strategy	Top-k	Top-p	GenEval ↑	ImageReward ↑
Single Step	1200	0.8	0.502	0.722
Multi Step	[1200, 1]	[0.8, 0]	0.541	0.806
	[1200, 10]	[0.8, 0.8]	0.531	0.799
	[1200, 100]	[0.8, 0.8]	0.529	0.745
	[1200, 100, 1]	[0.8, 0.8, 1]	**0.553**	**0.825**

Table 7. Impact of model size to visual generation.

Model size	Training epoches	GenEval ↑	ImageReward ↑
1B	4	0.485	0.677
7B	2	**0.553**	**0.825**

Table 8. Impact of different input strategies on multimodal understanding. Best results for each metric are highlighted in bold.

Input strategy	MME ↑	MME-P ↑	SEEDB ↑	TQA ↑
Full scale	1610.1	1315.1	59.6	49.5
Full scale residual	1527.5	1216.5	57.0	48.1
Last scale semantic feat. only	1580.3	1315.6	**60.1**	**49.7**
Last scale	**1634.3**	**1356.5**	59.9	49.1

Specifically, we initialize our framework with two different backbone models: TinyLlama-1B [72] and Llama-2-7B [53]. Experiments demonstrate that model size plays a crucial role in generation performance. As shown in Tab. 7 and Fig. 8, under identical sampling strategies and training dataset configurations, the 1B model significantly underperforms compared to its 7B counterpart, even with doubled training epochs.

Effect of Input Strategy to Multimodal Understanding. We validate different feature input strategies for multimodal understanding with TokenFlow. As shown in Tab. 8, final-scale features consistently outperform both full-scale features and full-scale residual features across all benchmarks. This suggests that the final scale captures the most relevant semantic information for multimodal understanding, while additional scale features or residual features may introduce noise that compromises performance. Our experiments also reveal that utilizing semantic features only does not improve the overall understanding performance.

Effect of Tokenizer Decoder Finetuning. To further improve our model's ability to generate fine details, we follow [6] and double both the number of residual layers and channel dimensions in the decoder. We exclusively finetune these enhanced decoder layers while keeping all other

Figure 9. Comparison of original images and their reconstructions from quantized semantic features extracted by VQKD [35]. The reconstructed images preserve the semantic content but exhibit significant loss of high-frequency details.

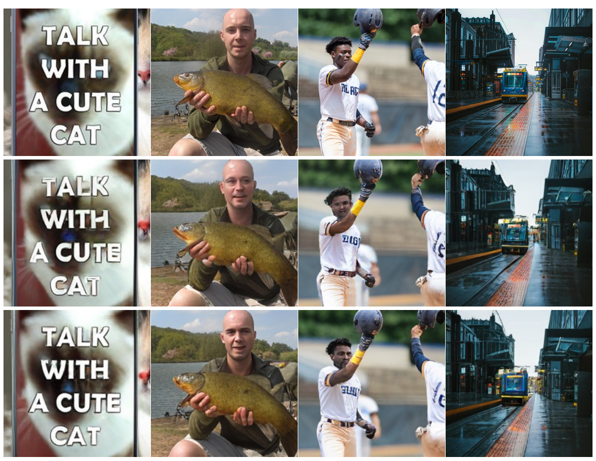


Figure 10. Comparison of image reconstruction quality. (a) Original images. (b) Reconstructions using the base pixel decoder. (c) Reconstructions using the enhanced (2× capacity) decoder. The enhanced decoder demonstrates superior preservation of fine-grained details, particularly in facial details and textual elements.

components frozen, thereby preserving the learned visual token mappings. This enables us to improve reconstruction fidelity without compromising perception ability of TokenFlow. As shown in Fig. 10, the enhanced decoder yields notable improvements in reconstruction quality. It demonstrates superior preservation of high-frequency details, particularly in facial details and text elements.

B.2. More Analysis of TokenFlow

Analysis of Joint Distribution Learning. To evaluate the effectiveness of our shared mapping mechanism, we conduct comparative experiments against VQKD [35] and VQGAN [13]. All models are configured with identical codebook sizes of 8,192 tokens for fair comparison. For baseline models, we utilize the official pretrained checkpoints from [35] and [48], respectively. Our TokenFlow model is trained on ImageNet-1K for 50 epochs. We deliberately excludes the multi-scale VQ design [51] to isolate the effects of the shared mapping in this experiment.

For evaluation, we process 50,000 images from the ImageNet-1K validation set through each model's encoder. We apply average pooling to the extracted features to obtain a 1×1 representation, and then identify the closest index in their respective codebooks using l_2 distance. As shown in Fig. 7, TokenFlow exhibits significantly smoother distribution against compared to others. The total non-empty clusters of TokenFlow are 7161/8192 (87.4%), which is significantly larger than that of VQGAN (2.5%) and VQKD (27.1%). These results demonstrate that our shared mapping design enables effective learning of joint distributions across high-level semantic and low-level pixel representations. By simultaneously encoding multiple levels of visual information, we induces a joint representation space compared to single-representation architectures. This directly contributes to the superior codebook utilization observed in our experiments. Even when expanding the codebook to over 131K entries, TokenFlow maintains an exceptional utilization ratio exceeding 95%. The clustered results is shown in Fig. 11.

Automatic Balancing between Semantic Distance and Pixel Distance. In our structure, the optimal quantize index is determined by $\arg\min_i(d_{\text{sem},i} + w_{\text{dis}} \cdot d_{\text{pix},i})$. There exists an automatic balancing mechanism between semantic distance and pixel distance. For instance, when encountering a case where $d_{\text{sem},i}$ is relatively small while $d_{\text{pix},i}$ is large, during backpropagation, both commit loss and perceptual loss will contribute to reducing the distance between the encoded features and their quantized counterparts. This mechanism naturally narrows the gap between these two distance metrics. Therefore, we set w_{dis} to 1.0 across all experiments.

Comparison between TokenFlow and their corresponding semantic teachers. Table 9 presents a fair

Table 9. Quantitative comparison of multimodal understanding capabilities between our discrete TokenFlow and their corresponding continuous semantic teachers. All experiments are trained with LLaVA-1.5 data for fair comparison. When calculating average, we use MME-P and divide it by 20 to have the same scale with other benchmarks.

Method	# Params	Visual Encoder	Res.	SEEDB	MMV	POPE	VQAv2	GQA	TQA	AI2D	RWQA	MMMU	MMB	MME	MME-P	Avg.
Continuous Visual Input																
LLaVA-1.5	Vicuna-13B	CLIP ViT-B/14 [37]	224	64.1	30.8	85.1	73.8	61.3	53.4	57.8	50.9	35.1	62.0	1737.0	1460.9	58.9
		ViTamin-XL [8]	256	65.7	34.6	85.8	76.8	62.6	57.4	59.4	54.4	35.0	66.4	1839.1	1514.5	61.3
		SigLIP-SO400M [69]	384	67.5	38.1	86.5	78.6	63.8	62.2	59.5	57.4	35.4	68.3	1802.1	1488.2	62.9
Discrete Visual Input																
Ours	Vicuna-13B	TokenFlow-B	224	60.4	22.4	84.0	70.2	59.3	49.8	54.2	49.4	34.2	55.3	1660.4	1353.6	55.2 (93.7%)
		TokenFlow-L	256	62.6	27.7	85.0	73.9	60.3	54.1	56.6	49.2	34.4	60.3	1622.9	1365.4	57.5 (93.8%)
		TokenFlow-XL	384	65.3	41.2	86.2	76.6	63.0	57.5	56.8	53.3	34.7	62.7	1794.4	1502.3	61.1 (97.1%)

Table 10. Comparison of generation quality on GenEval and DPG-Bench. Obj.: Object. Attri.: Attribute. † result is with rewriting.

	GenEval							DPG-Bench					
Method	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall	Global	Entity	Attribute	Relation	Other
Diffusion-based													
SDv1.5 [41]	0.43	0.97	0.38	0.35	0.76	0.04	0.06	63.18	74.63	74.23	75.39	73.49	67.81
DALL-E 2 [39]	0.52	0.94	0.66	0.49	0.77	0.10	0.19	–	–	–	–	–	–
SDv2.1 [41]	0.50	0.98	0.51	0.44	0.85	0.07	0.17	–	–	–	–	–	–
SDXL [36]	0.55	0.98	0.74	0.39	0.85	0.15	0.23	74.65	83.27	82.43	80.91	86.76	80.41
PixArt-alpha [7]	0.48	0.98	0.50	0.44	0.80	0.08	0.07	71.11	74.97	79.32	78.60	82.57	76.96
DALL-E 3 [4]	0.67†	0.96†	0.87†	0.47†	0.83†	0.43†	0.45†	83.50	90.97	89.61	88.39	90.58	89.83
Autoregressive meets diffusion													
Show-o [62]	0.53	0.95	0.52	0.49	0.82	0.11	0.28	67.27	79.33	75.44	78.02	84.45	60.80
Transfusion [74]	0.63	–	–	–	–	–	–	–	–	–	–	–	–
Autoregressive-based													
Chameleon [48]	0.39	–	–	–	–	–	–	–	–	–	–	–	–
LlamaGen [44]	0.32	0.71	0.34	0.21	0.58	0.07	0.04	64.84	81.76	75.43	76.17	84.76	58.40
EMU3 [55]	0.54	0.98	0.71	0.34	0.81	0.17	0.21	80.60	85.21	86.68	86.84	90.22	83.15
VAR [51]	0.53	0.95	0.60	0.41	0.81	0.16	0.24	71.08	77.51	78.17	77.80	85.80	62.00
Ours	0.55 0.63†	0.97 0.93†	0.66 0.72†	0.40 0.45†	0.84 0.82†	0.17 0.45†	0.26 0.42†	73.38	78.72	79.22	81.29	85.22	71.20

comparison between our discrete TokenFlow variants and their corresponding semantic teachers under the LLaVA-1.5 training paradigm. TokenFlow exhibits a relative performance gap compared to its semantic teachers due to vector quantized distillation. However, this gap diminishes as resolution increases: from 6.3% at 224×224 to 6.2% at 256×256, and finally to 2.9% at 384×384. This improvement can be attributed to the increased number of discrete tokens and additional scales supplementing the residual features at higher resolutions.

B.3. More Visual Generation Results

Quantitative Results. In Tab. 10, we present the complete scores for both GenEval [15] and DPG-Bench [18]. Following DALL-E 3 [4], we report our GenEval results using GPT-4V as a rewriter. For DPG-Bench, we tested the results of LlamaGen and Show-o using their released checkpoints. We compare against VAR [51] by using their released tokenizer and training the visual generation model under identical settings to ensure fair comparison.

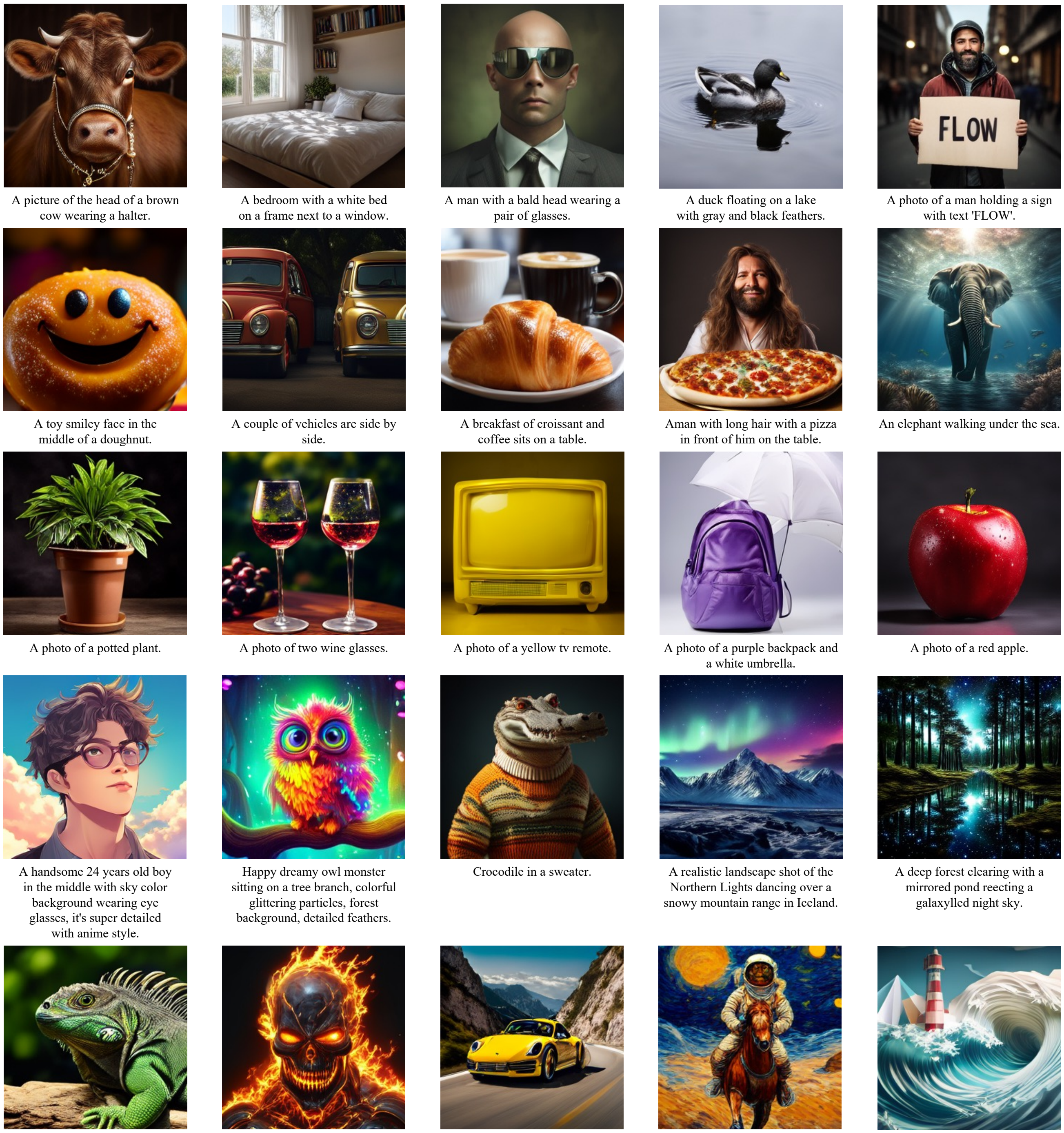
Qualitative Results. We present additional visual generation results in Fig. 12. Our method can generate images with various styles, subjects, and scenarios.

C. Limitation and Future Work

A primary limitation of TokenFlow lies in the performance gap in multimodal understanding between our discrete tokenizer and its continuous semantic teacher, which stems from the vector quantization distillation process. While this gap narrows to 2.9% at 384×384 resolution, several methods remain for further improvement, such as incorporating text alignment loss during tokenizer training.

In this work, we primarily focused on designing TokenFlow and validating its effectiveness separately in multimodal understanding and visual generation tasks. A natural extension of this work is the development of a fully unified model for both multimodal understanding and generation. This unification can be achieved through joint training on interleaved vision-language data. This is currently in our high priority for exploration.

Figure 11. Qualitative comparison of images clustered by VQKD [35], VQGAN [13] and our TokenFlow. VQKD clusters exhibit semantic similarity, while VQGAN clusters exhibit low-level similarity (*i.e.* color and texture). Our TokenFlow can successfully combine both semantic and low-level similarity (*e.g.* birds with different background can be mapped into two different index).



A picture of the head of a brown cow wearing a halter.

A bedroom with a white bed on a frame next to a window.

A man with a bald head wearing a pair of glasses.

A duck floating on a lake with gray and black feathers.

A photo of a man holding a sign with text 'FLOW'.

A toy smiley face in the middle of a doughnut.

A couple of vehicles are side by side.

A breakfast of croissant and coffee sits on a table.

Aman with long hair with a pizza in front of him on the table.

An elephant walking under the sea.

A photo of a potted plant.

A photo of two wine glasses.

A photo of a yellow tv remote.

A photo of a purple backpack and a white umbrella.

A photo of a red apple.

A handsome 24 years old boy in the middle with sky color background wearing eye glasses, it's super detailed with anime style.

Happy dreamy owl monster sitting on a tree branch, colorful glittering particles, forest background, detailed feathers.

Crocodile in a sweater.

A realistic landscape shot of the Northern Lights dancing over a snowy mountain range in Iceland.

A deep forest clearing with a mirrored pond reecting a galaxylled night sky.

A vivid green iguana is perched motionlessly atop a worn wooden log, its intricate scales exhibiting various shades of green and black.

An intricately detailed representation of the Marvel character Ghost Rider featuring a human skull, with flames licking around the contours of the skull and rising above it in a fierce expression of fiery vengeance.

A vibrant yellow 2017 Porsche 911 is captured in motion, navigating a winding mountain road with its sleek body hugging the curve.

An astronaut riding a horse on the moon, oil painting by Van Gogh.

A lighthouse in a giant wave, origami style.

Figure 12. More Visual Generation Results with TokenFlow. We present diverse 256×256 results across various styles, subjects, and scenarios.

Table 11. Detail settings of TokenFlow-B, TokenFlow-L and TokenFlow-XL.

Tokenizer	TokenFlow-B	TokenFlow-L	TokenFlow-XL
Tokenizer settings:			
Input resolution	224	256	384
Codebook size	32,768	32,768	32,768
Semantic teacher	CLIP ViT-B/14-224 [37]	ViTamin-XL-256 [8]	SigLIP-SO400M-patch14-384 [69]
Multi-scale settings	[1, 2, 4, 6, 8, 10, 12, 14]	[1, 2, 3, 4, 6, 8, 10, 12, 14, 16]	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 17, 22, 27]
Semantic codebook embedding dimension	32	32	32
Pixel codebook embedding dimension	8	8	8
Training settings:			
Learning rate	1e-4	1e-4	1e-4
Batch size	256	256	256
Training steps	1,000,000	500,000	500,000
Distance balance weight w_{dis}	1.0	1.0	1.0
Commitment loss factor β	0.25	0.25	0.25
Adversarial loss factor λ_{G}	0.5	0.5	0.5
Max gradient norm	1.0	1.0	1.0

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 6, 7

[4] 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. 8, 4

[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. 6

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 6, 8, 4

[8] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12954–12966, 2024. 3, 6, 1, 4, 7

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 7

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 6

[11] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3, 4, 6, 8, 1, 5

[14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3, 6, 7

[15] 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, 7, 8, 2, 4

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 6, 7

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5, 6

[18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 6, 7, 8, 4

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 6, 7

[20] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 6, 7

[21] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 3, 6

[22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6, 7

[23] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 3, 7

[24] 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. 7

[25] Xiang Li, Hao Chen, Kai Qiu, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024. 3

[26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6, 7

[27] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1

[28] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 3

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 3, 6, 7

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7

[31] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 7

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 6, 7

[33] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 1

[34] Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Huaian Chen, and Yi Jin. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*, 2024. 3

[35] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 3, 4, 5, 1

[36] 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, 8, 4

[37] 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. 3, 6, 1, 4, 7

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 8, 4

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 3, 8, 4

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[43] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3, 6, 7

[44] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1, 2, 3, 6, 8, 4

[45] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3

[46] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 1, 6

[47] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 8

[48] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3, 5, 6, 7, 8, 4

[49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[50] Qwen Team. Qwen2.5: A party of foundation models, 2024. 6, 7

[51] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 1, 3, 5, 6, 7, 8, 4

[52] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang,

Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 6

[53] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5, 6, 2

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[55] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2, 3, 7, 8, 4

[56] Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023. 5

[57] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 2, 3, 7

[58] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 7

[59] Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding. *arXiv preprint arXiv:2406.09781*, 2024. 1

[60] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 2, 3, 6, 7

[61] XAI. Realworldqa, 2024. 6, 7

[62] 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, 7, 8, 4

[63] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[64] 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, 4

[65] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1, 3

[66] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 3, 1

[67] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6, 7

[68] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 6, 7

[69] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3, 6, 1, 4, 7

[70] Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18467–18476, 2023. 3

[71] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 7

[72] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024. 2

[73] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022. 2, 3

[74] 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. 8, 4

[75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 7

[76] Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*, 2024. 3, 1