

RecTok: Reconstruction Distillation along Rectified Flow

Supplementary Material

Overview. In this supplementary file, we present more details in addition to the main paper. Here are the details:

- **Sec. 1:** Introduction video, we strongly recommend that reviewers take a look.
- **Sec. 2:** Implementation details.
- **Sec. 3:** More ablation studies.
- **Sec. 4:** More qualitative results.
- **Sec. 5:** More analysis on RecTok.
- **Sec. 6:** Limitations and future works.
- **Sec. 7:** Broader impacts.

1. Introduction Video

To help readers quickly grasp the primary idea of our work, we provide a 6-minute introduction video. Please refer to “[introduction_video.mp4](#)” in the supplementary file.

2. Implementation Details

In the Tab. 1, we provide detailed training configurations for both RecTok and DiT, including the hyperparameters for different DiT sizes, training epochs, learning-rate schedules, sampling schedules, and other related settings.

3. Additional Ablation Studies

Ablations on Encoder Initialization. We attempt to initialize the encoder of RecTok with VFMs, as shown in Tab. 2. We employ FSD and RAD as self-distillation [13] like methods to train the VFM initialized RecTok. However, the overall performance lags behind that of the randomly initialized encoder.

Ablations on Mask Ratio in RAD. We ablate the mask ratio used in reconstruction and alignment distillation (RAD). As shown in Tab. 3, we evaluate four mask ratio settings and report their effects on reconstruction, generation, and semantics. The results show that increasing the mask ratio improves generation quality while degrading reconstruction performance. Since reconstruction can be further enhanced through decoder finetuning, we set the upper bound of the mask ratio to 0.4.

Ablations on Noise Intensity in FSD. It is worth noting that the end point state of the forward flow in FSD, x_1 , does not need to follow a normal Gaussian distribution. As an alternative, we can instead place x_1 in a more expressive high-intensity noise space:

$$x_1 = \gamma \times \epsilon, \quad \epsilon \in \mathcal{N}(0, 1), \quad (1)$$

where ϵ denotes standard Gaussian noise, and γ controls the intensity of the noise, equivalently, the variance of x_1 .

A larger γ corresponds to stronger collisions at time t . We view this mechanism as analogous to the timestep shift introduced in our main paper, as both aim to mitigate information redundancy in high-dimensional settings. Therefore, we focus our analysis on the effect of γ and leave alternative formulations of x_1 to future work.

Ablations on KL Loss. We conduct an ablation study on the use of the KL loss as shown in Tab. 5. Although recent works [3, 4] remove the KL term and adopt an autoencoder (AE), our ablations show that incorporating KL regularization improves the generation performance. A detailed analysis of Tab. 5 reveals a distinct trade-off between reconstruction and generation performance. Specifically, the deterministic AE setting excels in reconstruction, achieving a lower rFID (0.35) and higher PSNR (29.89). This indicates that without the regularization constraint, the model can more freely encode high-frequency details into the latent space. However, this unconstrained latent space falls short in the generation stage, as evidenced by the degraded gFID (5.19). In contrast, enforcing the KL loss promotes a smooth and compact latent manifold. Although this results in a slight drop in reconstruction metrics, it significantly facilitates the learning process for the subsequent generative model, improving the gFID by over 50% ($5.19 \rightarrow 2.27$) and boosting the Inception Score by a large margin (+44.2). Since our work prioritizes the generation task, we retain the KL loss and use a variational autoencoder (VAE).

4. Additional Qualitative Results

Reconstruction Results. In Fig. 8, we present additional reconstruction results. RecTok accurately preserves the structure, color, and fine details of the input images.

Generation Results. In Fig. 5, Fig. 6, and Fig. 7. We provide additional generation results produced by a DiT^{DH} – XL model trained for 600 epochs. We show outputs both with and without classifier-free guidance.

5. Additional Analysis on RecTok

The Discriminative Ability along the Flow. We visualize the features along the forward flow x_t using t-SNE, uniformly sampling timesteps $t \in [0, 1]$. As shown in Fig. 2, our RecTok exhibits strong semantic consistency throughout the forward flow. The t-SNE visualization and linear probing accuracy demonstrate the stable discriminative ability of x_t . A more discriminative x_t also encourages forward trajectories to avoid intersections, which aligns with the objective of the original rectified flow [7].

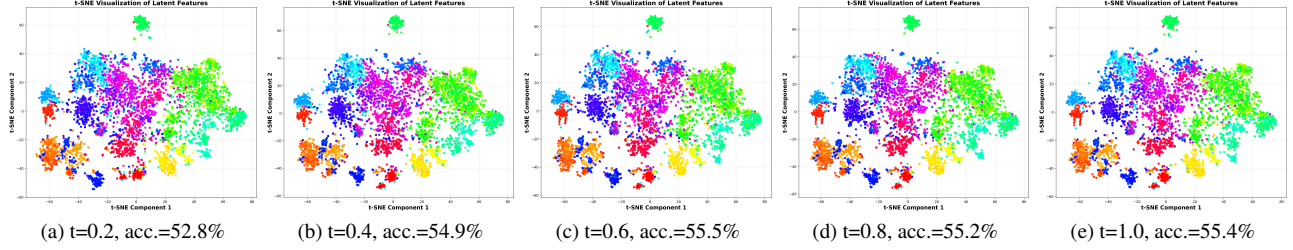


Figure 2. **t-SNE visualizations under different timesteps (t).** Our RecTok shows a clear advantage in semantic consistency on the forward flow, even with a high level of noise and disturbance.

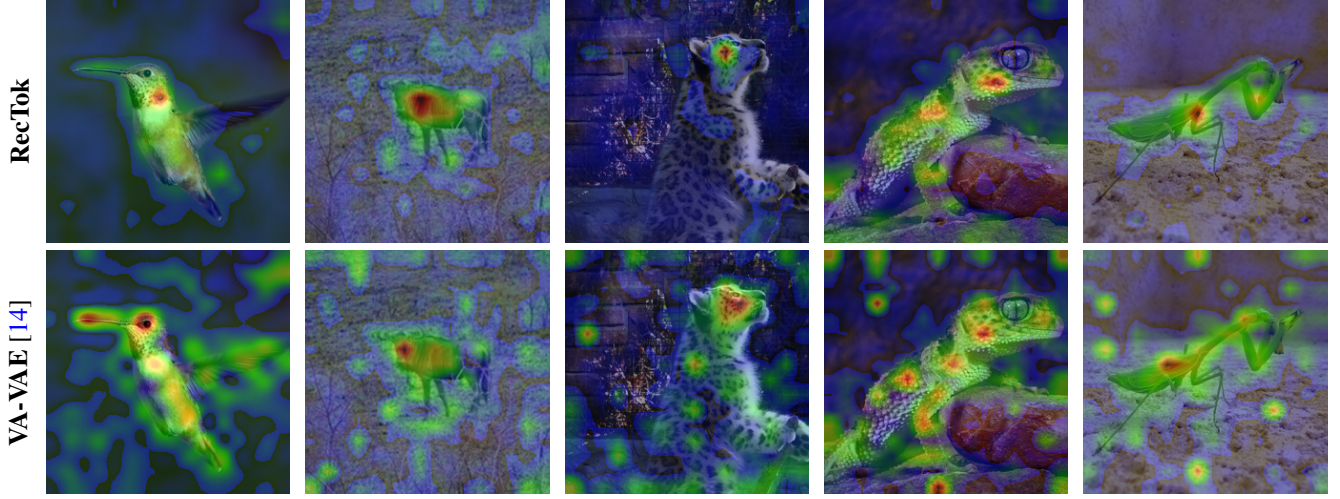


Figure 3. **Visualizations of latent feature through cosine similarity.** We compare the latent features of RecTok (top row) and VA-VAE (bottom row). The RecTok features exhibit stronger semantic localization on foreground objects compared to VA-VAE.

Latent Feature Visualization. We visualize the latent features of RecTok using cosine similarity. Specifically, we extract the latent features using RecTok and obtain a global feature via spatial pooling. We then compute the cosine similarity between the global and latent features. We also show the PCA results, as shown in Fig. 3 and Fig. 4, the heatmaps reveal that the learned latent features possess distinct semantic localization capabilities. Even without explicit segmentation supervision, the high similarity regions (indicated in warm colors) consistently align with the parts of the foreground objects, such as the head or the body structure. Conversely, background clutter and irrelevant textures are effectively suppressed. This suggests that RecTok’s tokenization process preserves spatial semantic integrity, ensuring that the aggregated global feature is highly representative of the core visual content and discriminative for downstream tasks.

6. Limitations

In terms of semantics, although RecTok enhances semantic structure by increasing the latent dimensionality, its dis-

criminative capability still lags behind that of VFMs. For example, DINOv3 [9], SigLIP 2 [11], and SAM [6]. Regarding reconstruction, while the KL loss smooths the latent space and improves generation quality, it inevitably weakens reconstruction ability, resulting in RecTok performing worse than an AE model with the same architecture. We leave these challenges as open questions for future work. We believe they can be addressed by further increasing the latent dimensionality and refining the KL regularization.

7. Broader Impacts

RecTok further expands the dimensionality of visual tokenizers while delivering consistent improvements in reconstruction, generation, and semantic representation. Its effectiveness suggests that the community may benefit from exploring higher-dimensional latent spaces that excel at both generative and understanding tasks [1, 2, 8, 10, 15]. Such a latent space serves as a *real unified representation*, removing the need for two separate image tokenizers as in prior unified models [12]. A shared feature space for both generation and understanding can promote mutual benefits across

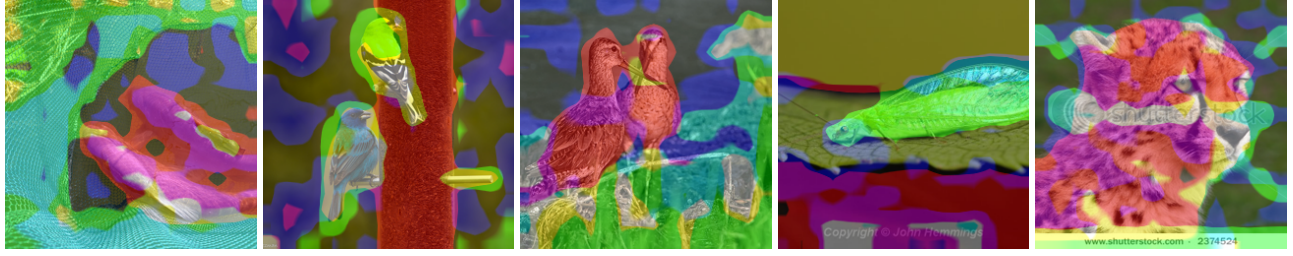


Figure 4. Visualizations of latent feature through PCA.

Table 1. **Implementation details.** We report the detailed architectural specifications, training hyperparameters, and sampling settings for different model variants.

architecture	DiT ^{DH} -S	DiT ^{DH} -B	DiT ^{DH} -L	DiT ^{DH} -XL
depth	12	12	24	28
hidden dim	384	768	1024	1152
heads	6	12	16	16
DDT depth			2	
DDT hidden dim			2048	
DDT heads			16	
training	RecTok	DiT ^{DH}		
epochs	200	80 (ablation), 600		
warmup epochs	50 (linear)	0		
decay epochs	50 - 200	40 - 800		
optimizer	Adam [5], $\beta_1, \beta_2 = 0.9, 0.95$			
batch size	1024			
learning rate	4e-4	2e-4		
learning rate schedule	cosine decay	linear decay		
weight decay	1e-4	0.0		
ema rate	0.999	0.9995		
noise schedule	$t = \frac{st'}{1+(s-1)t'}$, None	$t' \sim \mathcal{U}(0, 1)$,	$s = \sqrt{\frac{4096}{r^2d}}$	
class token drop (for CFG)			0.1	
sampling				
ODE solver	Euler			
ODE steps	50 (ablation), 150			
time steps	shift in [0.0, 1.0] according to noise schedule			
CFG scale	1.29			
CFG interval	[0, 1] (not used)			

Table 2. **Ablations on encoder initialization methods.** We notice that using the randomly initialized encoder yields better generation performance.

Encoder Initialization	L.P. Acc.	rFID	gFID	IS
VFM	54.5	0.57	2.37	189.2
Random	55.4	0.65	2.27	196.4

tasks, making unified models more coherent and meaningful.

Table 3. **Ablation on mask ratio.** We observe that a higher mask ratio of 0.4 yields the best overall generative performance (gFID and IS), suggesting that a more challenging task benefits RAD.

Mask Ratio	rFID	PSNR	gFID	IS
0.3	0.60	25.45	2.35	192.8
0.4	0.65	25.28	2.27	196.4
0.5	0.66	25.24	2.28	197.1
0.6	0.67	25.22	2.27	192.7

Table 4. **Effect of γ on reconstruction and generation quality.** We observe that $\gamma = 1.0$ achieves the optimal trade-off. While larger γ values marginally improve IS, they lead to a degradation in both reconstruction fidelity (rFID/PSNR) and generative distribution alignment (gFID).

γ	Noise Schedule	rFID	PSNR	gFID	IS
1.0	Shift	0.65	25.28	2.27	196.4
2.0	Shift	0.69	24.79	2.34	198.1
3.0	Shift	0.72	24.45	2.39	200.2

Table 5. **Ablation on the KL loss.** We compare the deterministic AE (w/o KL) and the VAE (w KL) settings. While removing the KL term improves reconstruction fidelity (rFID 0.35), it results in a disjointed latent space that hinders generation (gFID 5.19). Incorporating KL regularization significantly boosts generative performance (gFID 2.27), validating its necessity for our framework.

Setting	rFID	PSNR	gFID	IS
w/o KL	0.35	29.89	5.19	152.2
w KL	0.65	25.28	2.27	196.4

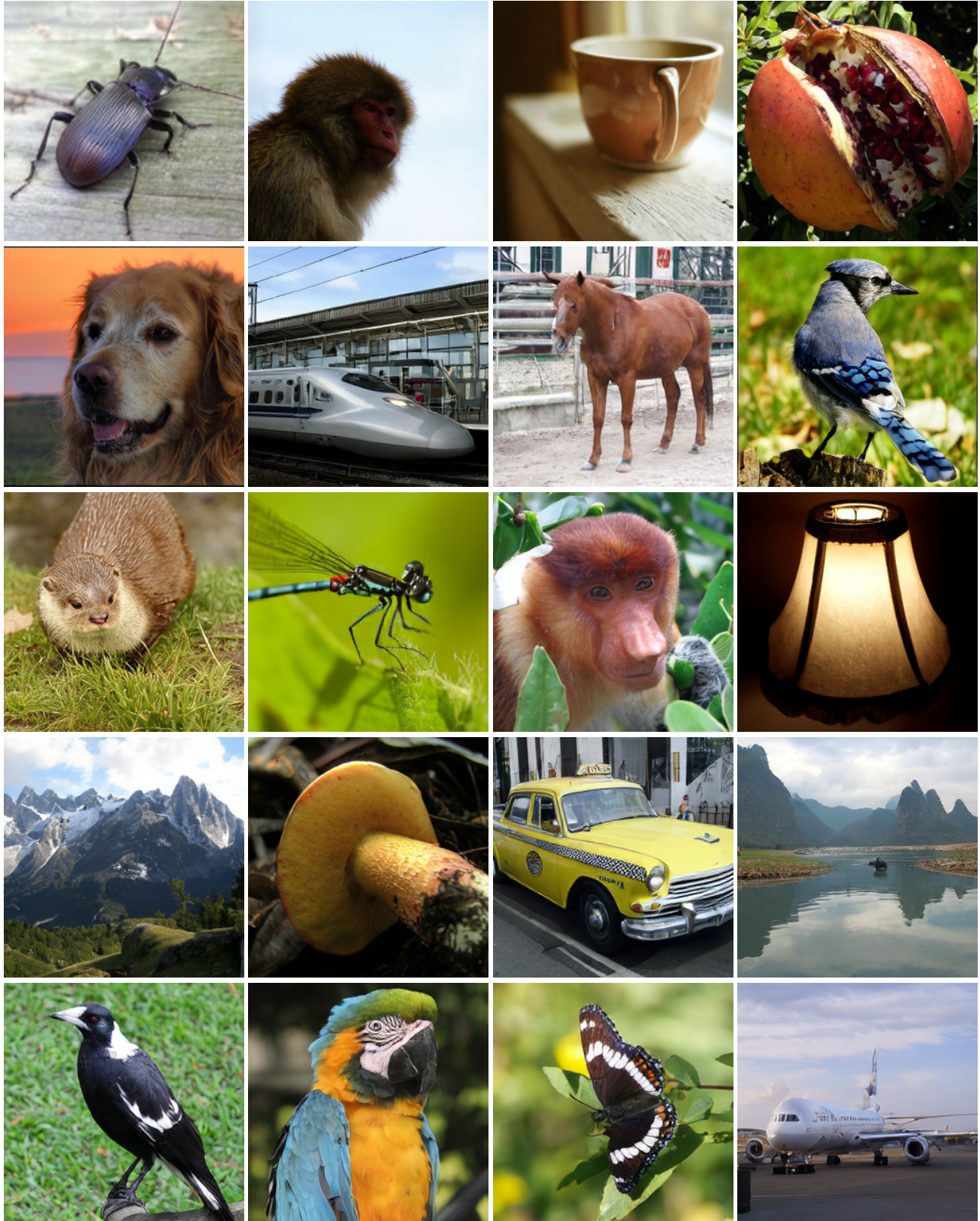


Figure 5. Supplementary generations (1/3).

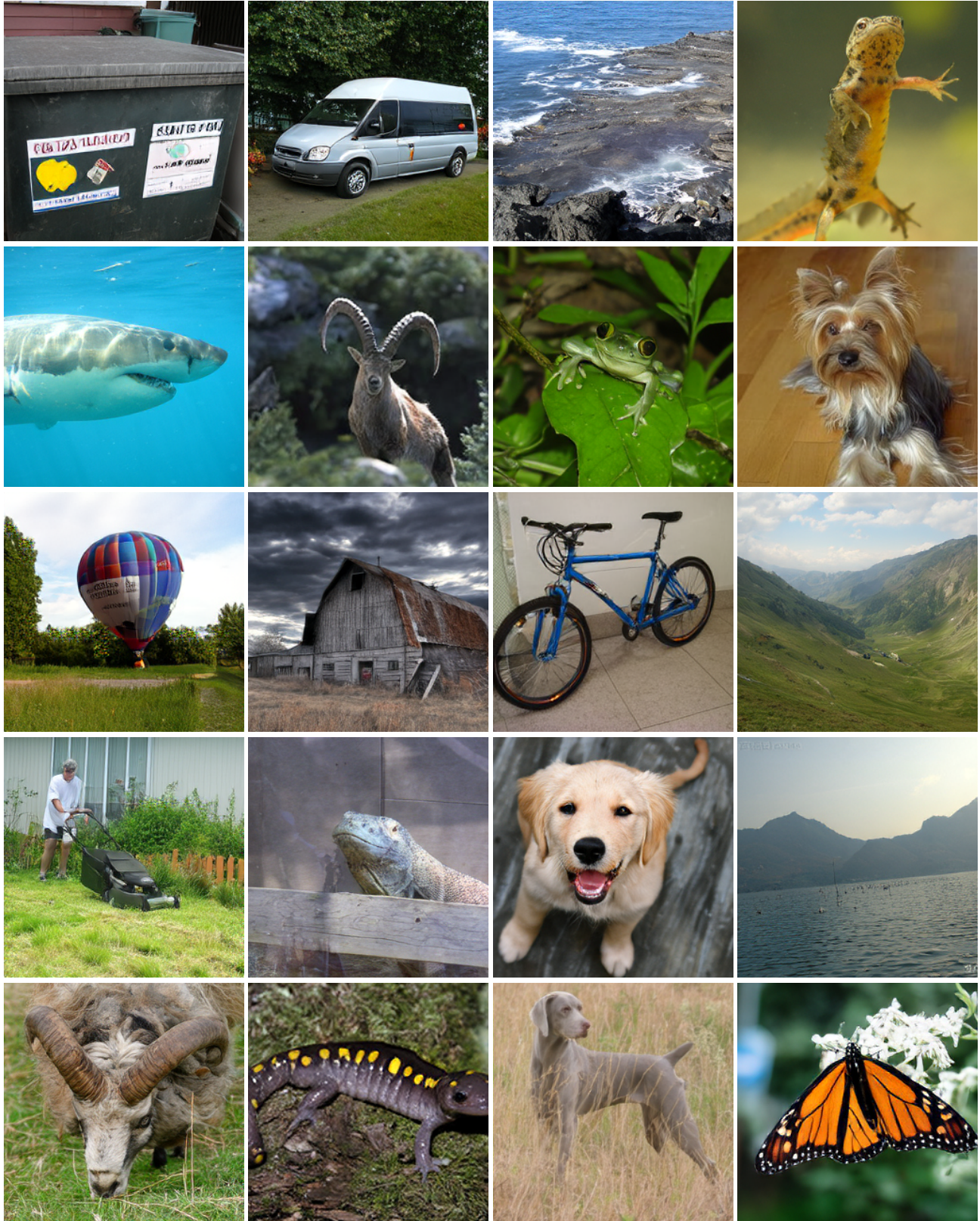


Figure 6. Supplementary generations (2/3).

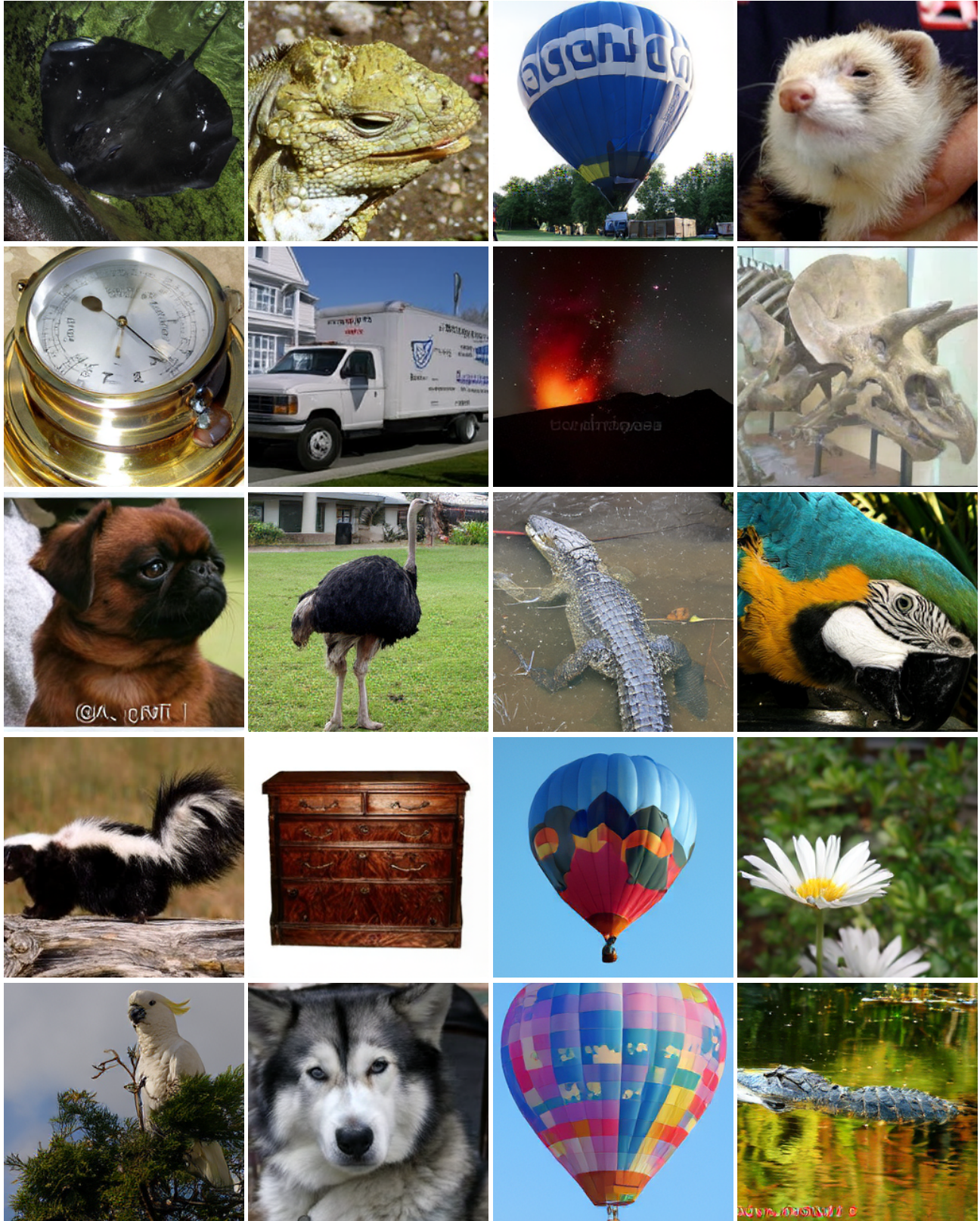


Figure 7. Supplementary generations (3/3).

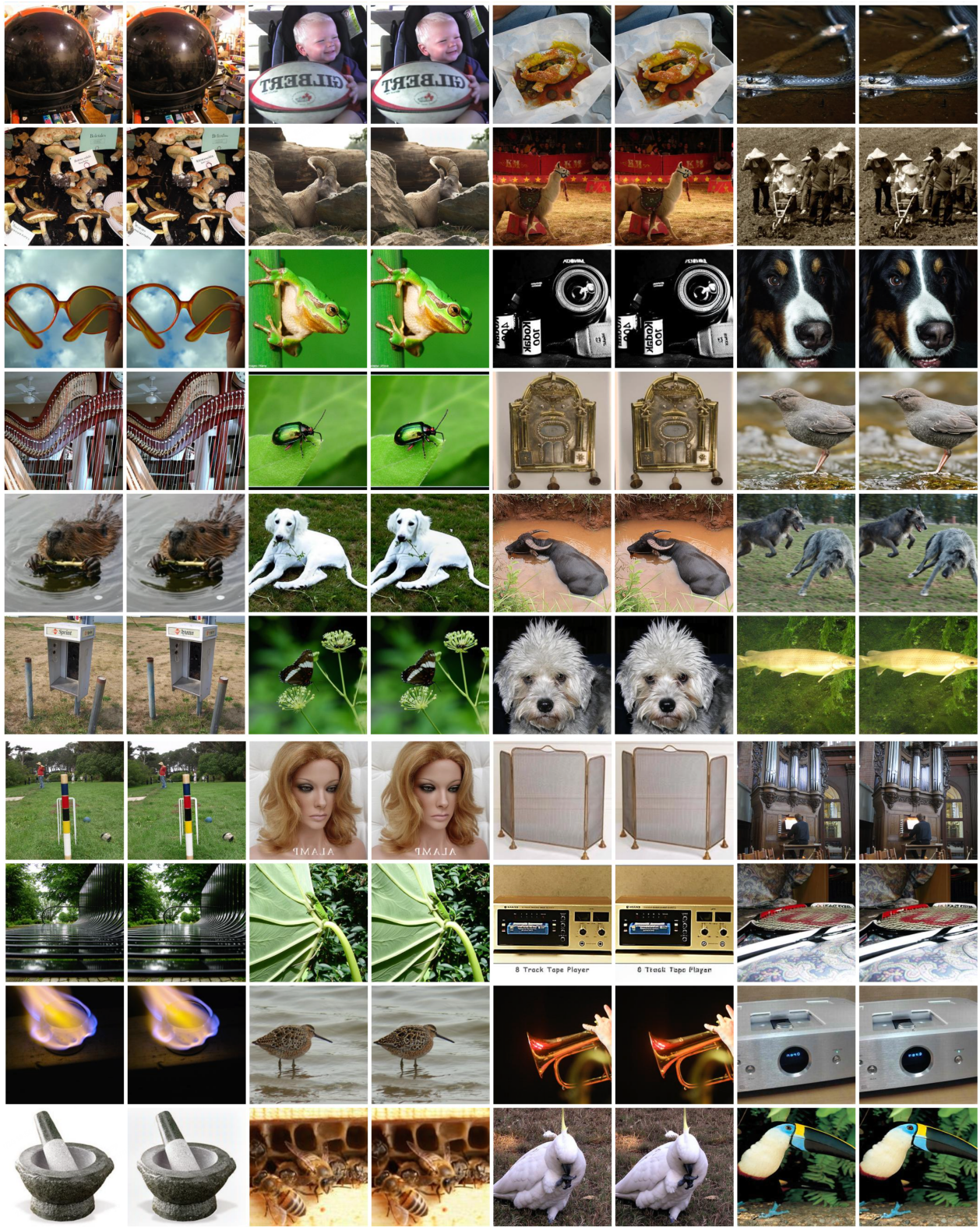


Figure 8. **Supplementary reconstruction (1/1).** We put the original images on the left and the reconstructed images on the right.

References

- [1] Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*, 2024. 2
- [2] Qifeng Cai, Hao Liang, Hejun Dong, Meiyi Qiang, Ruichuan An, Zhaoyang Han, Zhengzhou Zhu, Bin Cui, and Wentao Zhang. Lovr: A benchmark for long video retrieval in multimodal contexts. *arXiv preprint arXiv:2505.13928*, 2025. 2
- [3] Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. *arXiv preprint arXiv:2509.25162*, 2025. 1
- [4] Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *ICML*, 2025. 1
- [5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3
- [6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2
- [7] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 1
- [8] Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*, 2024. 2
- [9] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 2
- [10] Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*, 2025. 2
- [11] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2
- [12] 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
- [13] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024. 1
- [14] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025. 2
- [15] Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, et al. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*, 2024. 2