A. Related work

A.1. Diffusion acceleration models

ODE-based acceleration A line of studies suggests methods to efficiently obtain the solution of a given PF-ODE by developing numerical solvers [14, 26, 44, 85] to reduce discretization error under few-step regimes. Another line of research proposes learning a continuous-flow ODE model by rectifying the curvature of the ODE trajectory [34, 35, 41–43], which reduces the number of steps required for solving the ODE. Another approach is to distill a solution of the ODE into few-step student models by utilizing the deterministic nature of the given ODE trajectory, which is most aligned with our work. Some papers suggest learning a coupling of initial noise and samples by utilizing a regression loss between the student's prediction and the teacher's prediction [28, 30, 45, 51, 62, 86]. While others suggest utilizing a regression loss based on the student's self-consistency [2, 17, 27, 46, 72].

Distribution matching distillation This category of studies focuses on distribution matching between the few-step student and teacher distributions without relying on the ODE trajectory. Some methods estimate the score function of the student model during training and ensure that it becomes similar to that of the teacher [48, 49, 81–83]. Other methods utilize a discriminator and an adversarial loss [25, 39, 67, 68], but this usually requires an extensive hyperparameter search for stable training. Note that for practical performance gains, a distribution matching loss is often combined with a regression loss in ODE-based distillation methods as well [27, 28, 30, 83]. These methods require the training of auxiliary neural networks (e.g., student score network, discriminator) for distillation, which demands extra computational resources and memory.

A.2. Transformer-based visual generative models

Autoregressive model The visual autoregressive model was initially applied to pixel space using CNN or RNN architectures [77, 78]. Subsequently, a series of studies proposed autoregressive modeling in the vector-quantized embedding space with an autoencoding module [15, 33, 59, 84]. These models conceptualize an image as a sequence of 1D discrete-value tokens and adopt transformer architectures similar to those used in language models. Further research has expanded autoregressive models to multimodal generation [73, 74, 80] by jointly estimating language and visual tokens – one token at a time. Recently, [76] proposed to predict a sequence of 2D token maps instead of the sequence of 1D tokens. This change significantly reduced the length of the predicted sequence, resulting in fewer calls to the model during inference. Our work also works on 2D token maps. However, unlike the approach in [76], which requires an ad-hoc autoencoder, our sequence of 2D token maps is obtained directly from any pre-trained diffusion transformer.

DART DART [18] is a concurrent work that proposes an autoregressive model with a sequence of 2D token maps derived from the diffusion process. However, there are fundamental differences with our work. We propose a distillation method (ARD) that allows reducing the number of inference steps down to 3 or 4 following the backward ODE trajectories of a pre-trained diffusion model, in contrast, DART is a vanilla diffusion model that is very inefficient and requires a lot of sampling steps. This results in a significant performance gap in Table 5, as our model is specifically designed to work in low-step regime. There is also a difference in how a sequence of 2D token maps is formed. In ARD, we directly use the teacher's backward ODE trajectory. While in DART, the trajectory is formed by applying the forward noising process to the ground truth data.

Table 5. Comparison of ARD and DART on ImageNet 256p.

Model	Steps ↓	GFLOPs ↓	FID ↓
DART [18]	16	2157	3.98
ARD (Ours)	**4**	**479.9**	**1.84**

Masked generative models Using the vector-quantized embedding space [59], masked prediction is another line of work for generation [4, 36, 52] similar to BERT [11] in language models. Unlike the autoregressive models, which predict the sequence of tokens in order, this approach starts with a full mask and predicts all masked tokens simultaneously. By repeating the process of re-masking and predicting, it eventually obtains the generated sample.

B. Experimental details

B.1. Class conditional image generation

We follow the teacher configuration [54] for student training, except for gradient clipping and batch size. Table 6 presents the training configuration for the 4-step student model with regression loss. We use a batch size of 128 for the 2-step student model and 256 for the 1-step student model. The student model is initialized from the teacher's weights. By default, we set the prediction target at s as $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_{\tau_s}]$. We use 8 NVIDIA A100 GPUs for training, which takes approximately 2 days. As shown in the blue line of Fig. 8a, the FID almost converges within 16 hours (100k iterations). We applied the same settings to the baseline (step distillation [62]) as well. Note that step distillation was not trained using the progressive algorithm proposed in the original paper [62]; instead, it was learned directly from the teacher. Please refer to Section 2.2 for the objective function.

Configuration	Setting
Learning Rate	10^{-4}
Weight Decay	0.0
Gradient Clipping	1.0
Batch Size	64
Iterations	$300k$
EMA Decay Rate	0.9999

Table 6. Details on class conditional generation.

When using an additional discriminator loss, we utilize the teacher network as a feature extractor similar to [67], and train only the discriminator heads on top of the extracted features from each transformer block [64]. Discriminator heads predict logits token-wise. We use hinge loss, as described in [66], and follow the discriminator head architecture proposed in the same work. The discriminator is trained on the final prediction of the student model $\hat{\mathbf{x}}_{\tau_0} = G_{\boldsymbol{\theta}}(\mathbf{x}_{\tau_S:\tau_1}, 1)$, and real data. We train it with a learning rate of $1e-3$ and no weight decay. We set adaptive balancing between the regression loss and the discriminator loss following [15]. A batch size of 48 is used for both the student model and the discriminator. By adding a discriminator loss and further finetuning a student model pre-trained with regression loss, we achieve an improvement in FID from 4.32 to 1.84 within just 40k iterations.

B.2. Text conditional image generation

The Emu teacher model[8] has 1.7B parameters, consists of 24 DiT layers, and uses cross-attention layers for text conditioning. Emu is a latent diffusion model, which encodes 1024×1024×3 images in a 128×128×8 latent space. The distillation setup is similar to the procedure in the class-conditional case. We follow the same training configuration as the teacher, except that we adjust the batch size. The student model uses a block-wise causal mask (M4), and $N = 1$ as it showed the best performance-quality trade-off. The prediction target is set to $\mathbf{x}_{\tau_s}$ for fast training. We use 32 H100 GPUs for the student model training. Table 7 shows the prompts (from left to right, top to bottom) that we used for Fig. 1.

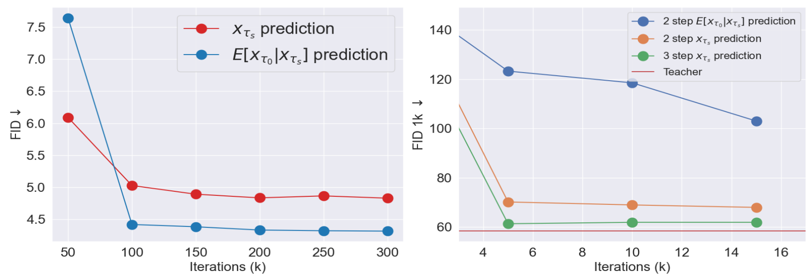
A majestic unicorn prancing through a vibrant field of rainbows and flowers, with its face visible and distinguishable. The unicorn's coat is a shimmering white, with a spiral horn protruding from its forehead and a flowing mane that shimmers in the sunlight. Its legs are slender and powerful, with hooves that barely touch the lush green grass. The rainbows in the field are multicolored, arching up from the ground and swirling around the unicorn's body, while the flowers are a variety of bright colors, blooming in every direction. The atmosphere is one of joy and wonder, as if the unicorn is dancing through a magical paradise.
A scale-up 4k photo-realistic image of a cat wearing a golden crown, with its face visible and distinguishable. The cat is sitting on a chair made by avocado wood, with a soft and luxurious texture. The cat's fur is a vibrant shade of orange, with a few white stripes on its face. The crown is made of pure gold, with intricate carvings and a sparkling gemstone in the center. The cat's eyes are fixed intently on the camera, exuding a sense of confidence and royalty.
A serene scene of a couple of sheep relaxing in a lush green field, with their face visible and distinguishable. The sheep are lying down, with one sheep resting its head on the other's back, conveying a sense of comfort and companionship. The sheep's wool is a soft, fluffy brown color, and their ears are slightly perked up, as if they are enjoying the peaceful atmosphere. The field is filled with wildflowers of various colors, adding a pop of vibrancy to the scene. The sun casts a warm glow over the entire scene, creating long shadows that stretch across the field.
The image shows a corgi sitting on a white background, wearing a bright red bowtie around its neck and a vibrant purple party hat on its head. The bowtie is made of silk and has a subtle sheen to it, while the party hat is adorned with glittery decorations. The corgi's fur is a sandy brown color, and its ears are perked up, as if listening intently. Its big brown eyes are looking directly at the camera, and its tongue is slightly out, as if it's panting happily. The overall atmosphere is festive and playful, suggesting a celebratory occasion.
A very bright and clear 4k close-up image of a tiger visible face and distinguishable features. The background is a mountain with a river flowing through it. The tiger is sitting on a rock, looking directly at the camera, with its arms relaxed and its paws clasped together. The tiger's fur is a soft, golden-brown color, and its eyes are a deep, dark brown. The mountain is a vibrant green, with a few snow-capped peaks visible in the distance. The river is a clear blue, with a few small fish swimming in the water. The overall atmosphere is serene and peaceful, as if the tiger is enjoying a quiet moment in nature.
The image depicts a photo-realistic scene of a dog recklessly driving a go-kart on a winding road. The dog, a sleek black feline with bright eyes, is hunched over the steering wheel, its paws gripping the wheel tightly as it takes a sharp turn. The go-kart is a miniature version of a real race car, with shiny metal bodywork and large wheels. The road is lined with lush greenery and rocks, and the background shows a cloudy sky with a hint of sunlight peeking through. The dog's fur is ruffled by the wind, and its ears are flapping wildly as it speeds along.
A serene landscape with a majestic mountain range in the background, its peaks covered in a blanket of snow. In the foreground, a vibrant meadow is filled with an array of colorful wildflowers, swaying gently in the breeze. A tranquil lake reflects the beauty of the surroundings, its crystal-clear waters glistening in the sunlight. The atmosphere is peaceful, with a sense of harmony between the natural elements, as if nature is offering a sacrifice of its own beauty.
The image shows a vibrant and colorful spread of traditional Chinese cuisine. A steaming plate of dumplings sits alongside a plate of savory Peking duck, complete with crispy skin and tender meat. A bowl of steaming hot noodles is placed nearby, garnished with sliced green onions and a sprinkle of sesame seeds. A small dish of egg rolls and a plate of fresh fruit complete the spread. The food is arranged on a red tablecloth, and the background is a warm and inviting Chinese-style restaurant setting with intricately carved wooden panels and ornate lanterns.
A burger on a wooden board with fries and ketchup. The burger has a sesame seed bun, lettuce, tomato slices, and a beef patty. There are two skewers sticking out of the top of the burger. On the left side of the burger, there is a small white bowl filled with ketchup. To the right of the burger, there are french fries on a wooden board. In front of the burger, there is a knife with a wooden handle laying on its side. Behind the burger, there is a small white bowl filled with mayonnaise. In the background, there is a wooden cutting board with a bowl of cherry tomatoes, a green chili pepper, and a brown cloth.

Table 7. The prompts used to generate the images for Fig. 1.

C. Additional experimental results

C.1. More ablations on the ImageNet 256p

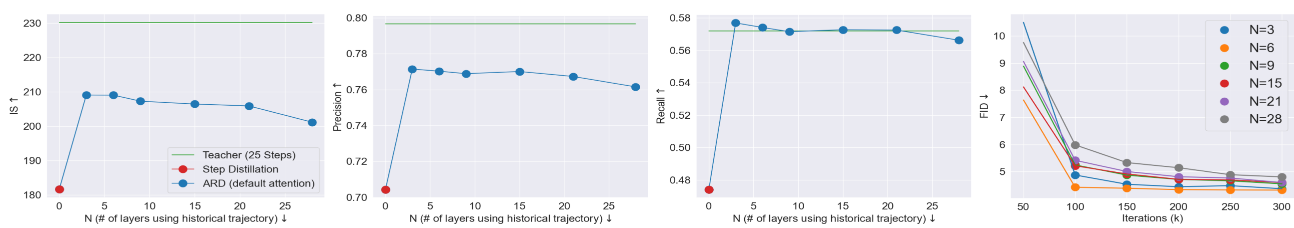
Prediction targets The teacher model provides the ODE path $\mathbf{x}_{\tau_s} \in [\mathbf{x}_{\tau_S}, ..., \mathbf{x}_{\tau_0}]$. When we solve the teacher ODE, the teacher also provides the equivalent target $\mathbb{E}[\mathbf{x}_{\tau_0}|\mathbf{x}_{\tau_s}]$, which is known as "predicted $\mathbf{x}_{\tau_0}$" [69] at each step s. Figure 8a shows the ablation of two prediction targets in class-conditional experiments. The convergence of the $\mathbf{x}_{\tau_s}$ prediction is faster during the first 50k iterations, but with the target $\mathbb{E}[\mathbf{x}_{\tau_0}|\mathbf{x}_{\tau_s}]$ the model finally converges to a better local optimum. We suspect that $\mathbb{E}[\mathbf{x}_{\tau_0}|\mathbf{x}_{\tau_s}]$ has an advantage in providing fine-grained information as an input because unnecessary noise is removed. The difference in learning speed between the two estimation targets is more pronounced in the text-conditional experiments, as shown in Fig. 8b. In our early work, we observed that the $\mathbf{x}_{\tau_s}$ prediction quickly converged with satisfactory performance in the 2-step experiment, so we chose the $\mathbf{x}_{\tau_s}$ prediction for the remaining experiments.



(a) Class-conditional experiments (b) Text-conditional experiments

Figure 8. Ablation on the prediction targets.

Ablation of the number of layers using historical trajectory (N) We provide more evaluation results by ablating N for our class-conditional ARD model, where N is the number of layers that use the historical trajectory. Figures 9a to 9c show the performance of our method for different values of N. Each metric shows a similar trend as the FID, as shown in Fig. 6c. Figure 9d shows the training curves for various values of N. We find that $N = 6$ is the optimal value. The performance advantage is maintained after 100k iterations and $N = 6$ not only has a better convergence point but also learns faster.



(a) Inception Score (b) Precision (c) Recall (d) FID per itertions

Figure 9. More evaluation results on N ablations.

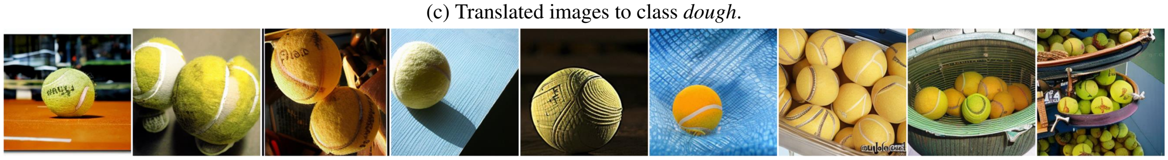
C.2. Image manipulation

ARD provides image manipulation capability similar to [50]. Starting sampling from initial noise $\mathbf{x}_{\tau_S}$, ARD denoises it to the target class by using the source image $\mathbf{x}^{\text{src}}$ as input instead of the prediction $\hat{\mathbf{x}}_{\tau_s}$ at a certain time step s (i.e., $\hat{\mathbf{x}}_{\tau_{s-1}} = G_{\boldsymbol{\theta}}([\hat{\mathbf{x}}_{\tau_S:\tau_{s+1}}, \mathbf{x}^{\text{src}}], s)$). The subsequent sampling process remains the same as before. Figure 10 shows the image translation results using 4-step ARD. The prediction of the first step is replaced with the source images and is fed into the ARD model.

(a) Source images from class *orange*.

(b) Translated images to class *lemon*.

(c) Translated images to class *dough*.



(d) Translated images to class *tennis*.

Figure 10. Image translation with ARD (R+D) 4-step model.

C.3. Reference teacher performance for T2I Image-Text alignment

Table 8 shows the performance of the corresponding teacher of the 3-step distillation models listed in Table 3. ARD (distilled from Emu (3.0 CFG)) has a gap of 2.3 ($= 58.5 - 56.2$) in average score, which is the smallest among all the competitors. Pixart-delta [6] (distilled from Pixart-alpha) shows a gap of 4.2, and LCM-LoRA [47] (distilled from SSD and LDM-XL) shows gaps of 8.0 and 3.3, respectively. All the 768-resolution student models are distilled from Emu (2.7B), and the best model Imagine Flash [30] has a gap of 4.9.

Table 8. The performance on CompBench for the target teachers for each 3-step T2I distillation model.

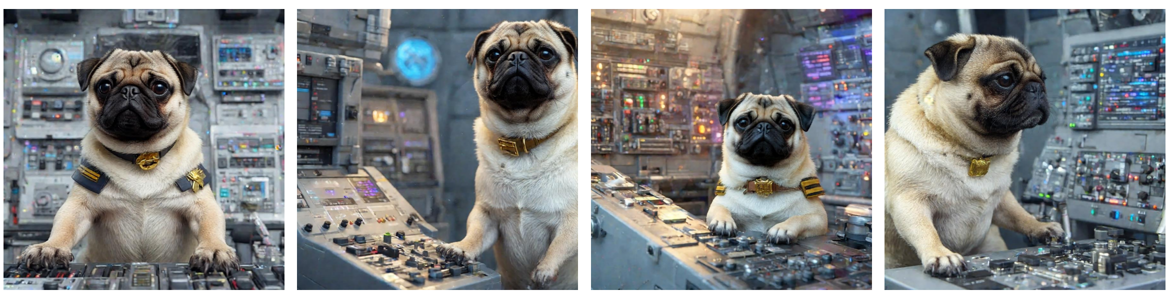
				CompBench ↑ (%)						
Teacher	Params ↓	Res.	CFG	Color	Shape	Texture	Spatial	Non-spatial	Complex	AVG ↑
Emu [8]	2.7B	768	6	51.8	39.8	53.5	60.3	67.9	46.0	53.2
Pixart-alpha [7]	0.6B	1024	4.5	41.7	39.1	45.9	60.7	62.7	43.0	48.9
SSD [58]	1.3B	1024	7.5	60.3	44.5	51.4	61.7	65.4	45.3	54.8
LDM-XL [56]	2.6B	1024	7.5	62.8	51.2	54.9	61.3	61.9	43.3	55.9
Emu [8] (Our teacher)	1.7B	1024	3.0	67.9	50.7	64.5	63.0	59.6	45.4	58.5
Emu [8] (Our teacher)	1.7B	1024	7.5	72.6	55.4	70.4	69.6	64.4	49.9	63.7

C.4. Additional samples

This section presents additional generated samples. Figure 11 illustrates the text-conditional samples with various initial noise $\mathbf{x}_{\tau_S}$. Figure 12 shows the comparison between ARD and the Emu teacher with the same number of sampling steps. Figures 13 to 20 show the class-conditioned generations across various classes.

A 4k close-up clear image of majestic tiger prancing through a vibrant field of rainbows and flowers, with its face visible and distinguishable. The tiger's coat is a shimmering, with sharp teeth protruding from its mouth and a flowing mane that shimmers in the sunlight. Its legs are slender and powerful, with hooves that barely touch the lush green grass. The rainbows in the field are multicolored, arching up from the ground and swirling around the tiger's body, while the flowers are a variety of bright colors, blooming in every direction. The atmosphere is one of joy and wonder, as if the tiger is dancing through a magical paradise.

The image shows a small chihuahua running through the waves on a sunny beach. The dog is tiny, with a brown coat and a fluffy tail, and it's running towards the left side of the image. The surf is gentle, with small waves rolling in from the ocean. The beach is lined with palm trees, and there are a few beach towels and sun loungers scattered in the distance. The chihuahua's legs are moving quickly as it runs, and its ears are flapping in the wind. The atmosphere is lively and joyful, capturing the carefree spirit of a day at the beach.



Captain Pug Dog, a small but determined-looking dog, standing at the command center of the Starship Enterprise. With his paws on the control panel, he is issuing commands to the crew, his face visible and distinguishable as he surveys the vast galaxy. He is wearing a miniature version of Captain Kirk's iconic uniform, complete with a gold pin on his collar and a confident expression. The Enterprise's sleek, silver hull looms behind him, with the ship's iconic logo emblazoned on the side. The scene conveys a sense of excitement and adventure, with Captain Pug Dog at the helm, ready to explore the final frontier.

A plate of juicy chicken and a variety of colorful vegetables arranged artfully next to a steaming bowl of white rice. The plate is made of ceramic and has a simple, elegant design, with the chicken pieces arranged in a neat row. The vegetables include carrots, broccoli, and bell peppers, all cut into bite-sized pieces. The bowl of rice is placed directly next to the plate, with a small amount of soy sauce or sesame oil drizzled on top. The overall presentation is appetizing and inviting, with the warm golden tones of the chicken and rice complementing the vibrant colors of the vegetables.

Figure 11. Samples generated by our 3-step ARD model, distilled from a 1.7B Emu.

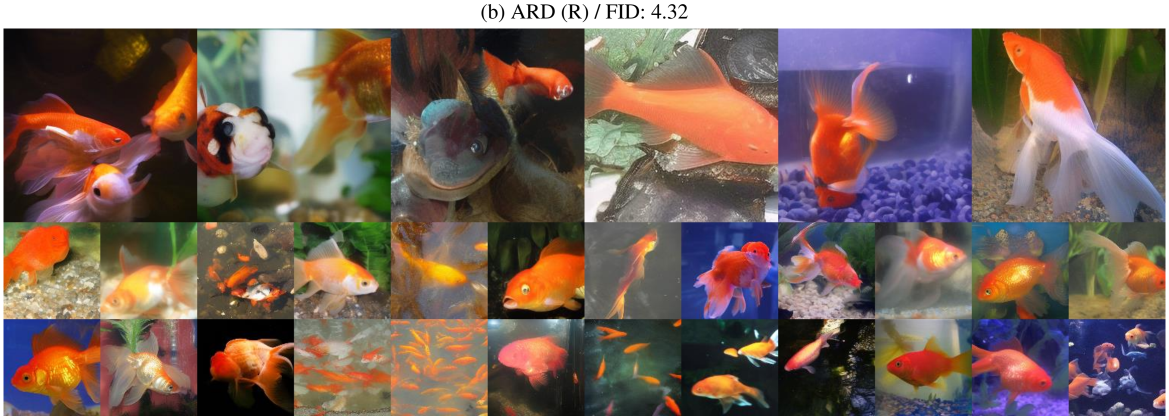
(a) Samples from 3-step ARD

(b) Samples from the Emu teacher with 3-step

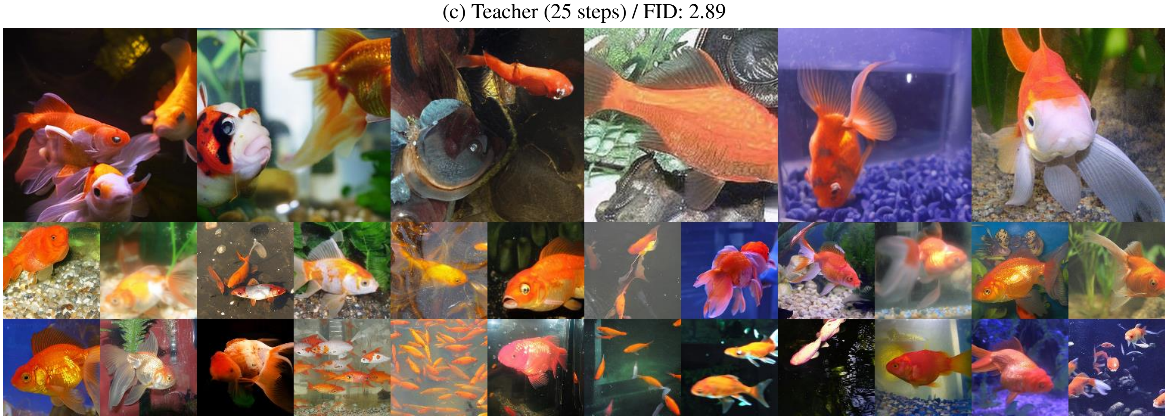
Figure 12. Comparison of generations between ARD and the Emu teacher.

(a) Step Distillation (R) / FID: 10.25

(b) ARD (R) / FID: 4.32



(c) Teacher (25 steps) / FID: 2.89

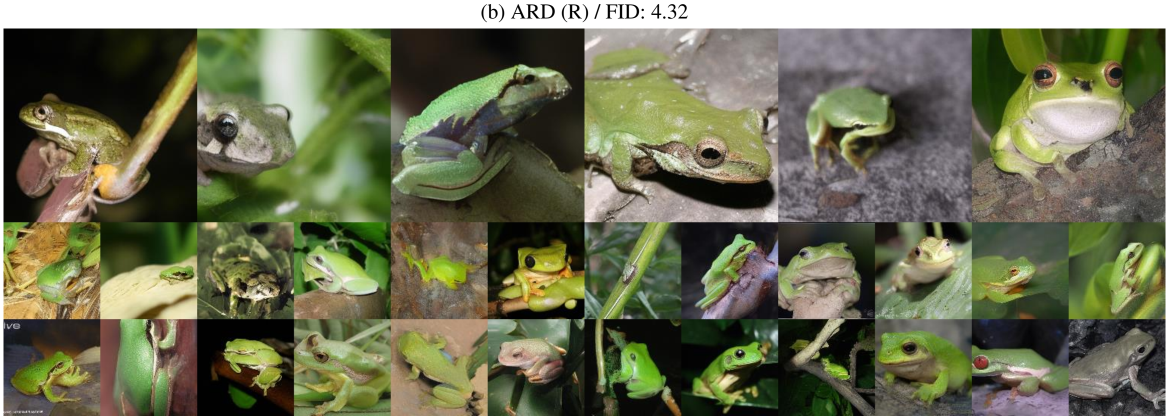


(d) ARD (R+D) / FID: 1.84

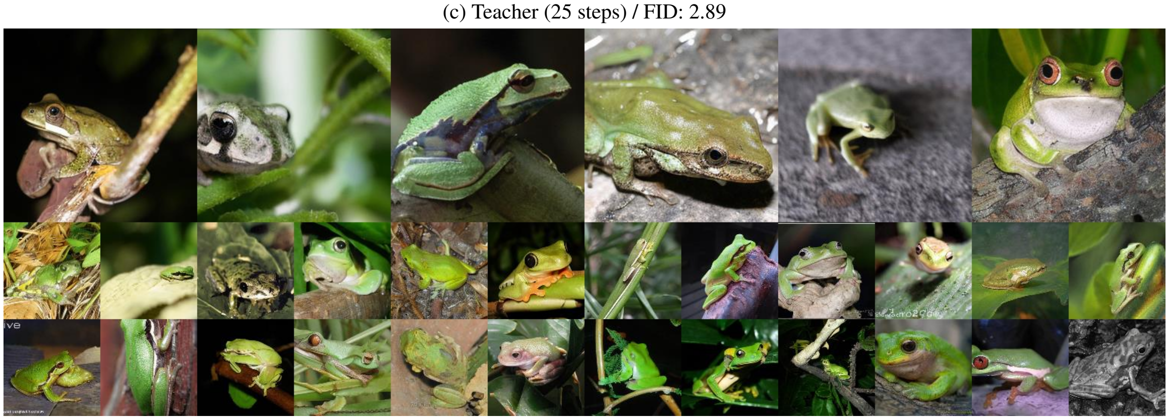
Figure 13. Randomly generated ImageNet 256p samples for class *goldfish*. All distilled models are 4-step models.

(a) Step Distillation (R) / FID: 10.25

(b) ARD (R) / FID: 4.32



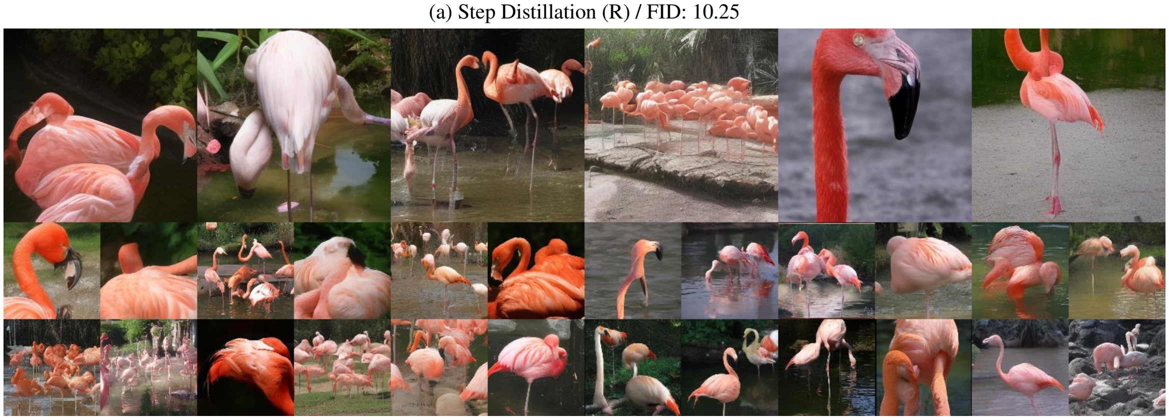
(c) Teacher (25 steps) / FID: 2.89



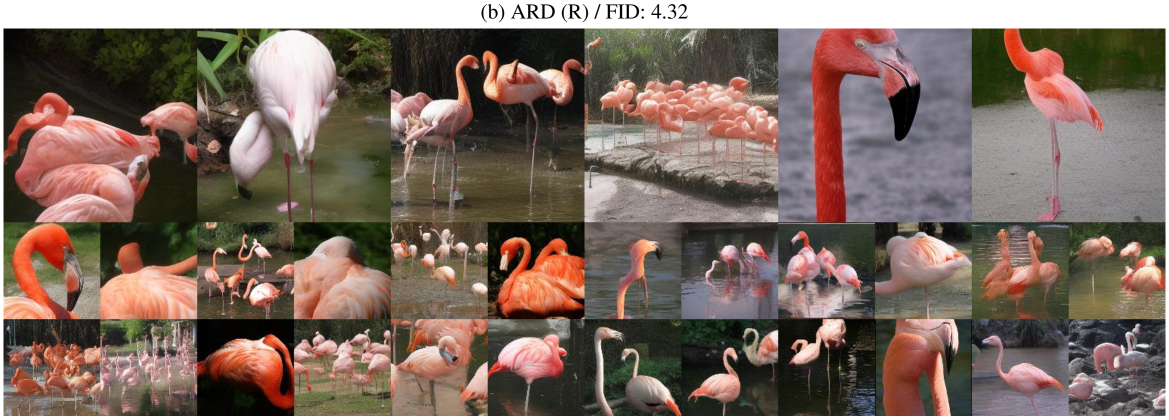
(d) ARD (R+D) / FID: 1.84

Figure 14. Randomly generated ImageNet 256p samples for class *tree frog*. All distilled models are 4-step models.

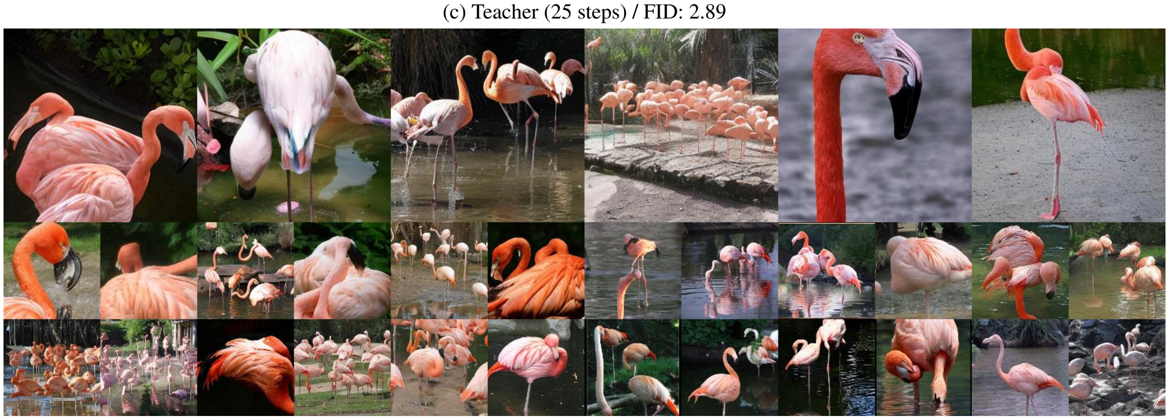
(a) Step Distillation (R) / FID: 10.25



(b) ARD (R) / FID: 4.32



(c) Teacher (25 steps) / FID: 2.89

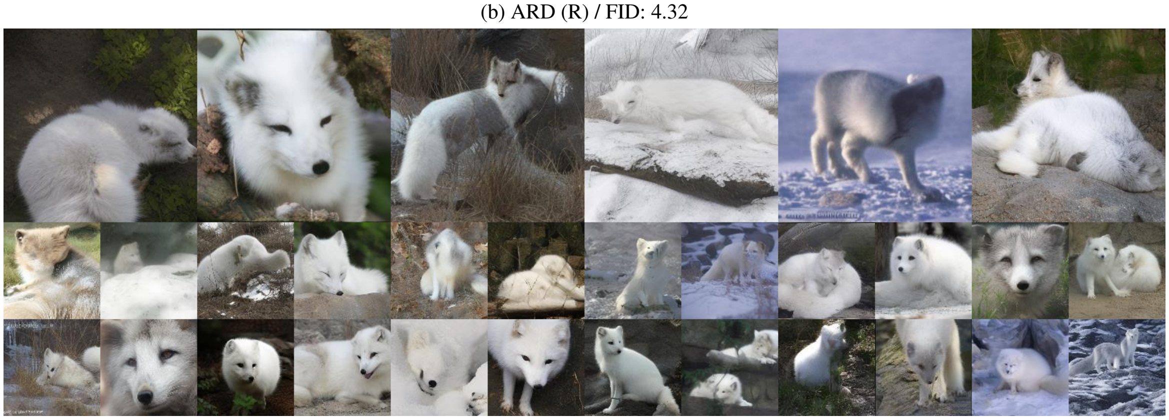


(d) ARD (R+D) / FID: 1.84

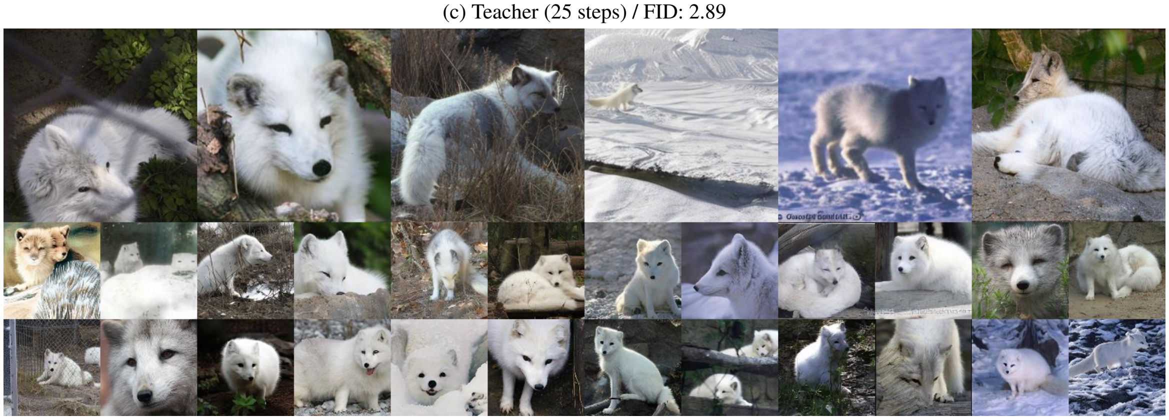
Figure 15. Randomly generated ImageNet 256p samples for class *flamingo*. All distilled models are 4-step models.

(a) Step Distillation (R) / FID: 10.25

(b) ARD (R) / FID: 4.32



(c) Teacher (25 steps) / FID: 2.89

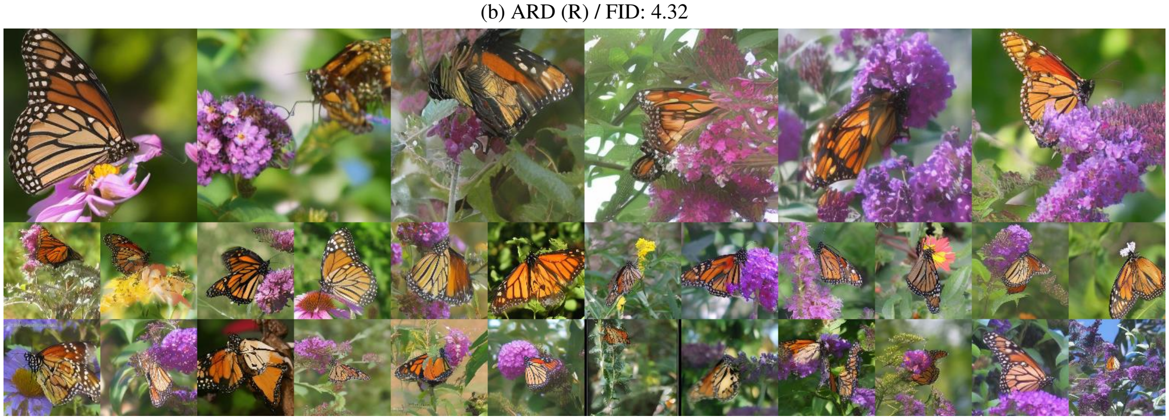


(d) ARD (R+D) / FID: 1.84

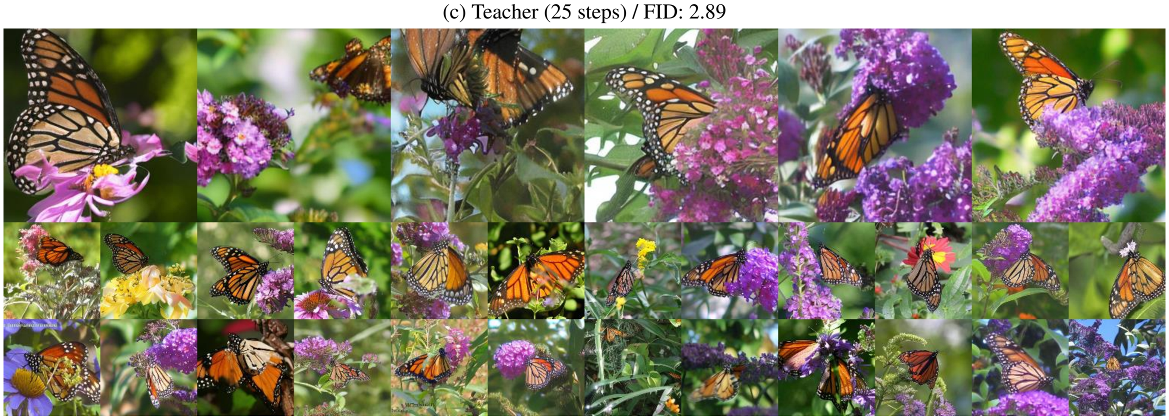
Figure 16. Randomly generated ImageNet 256p samples for class *Arctic fox*. All distilled models are 4-step models.

(a) Step Distillation (R) / FID: 10.25

(b) ARD (R) / FID: 4.32



(c) Teacher (25 steps) / FID: 2.89



(d) ARD (R+D) / FID: 1.84

Figure 17. Randomly generated ImageNet 256p samples for class *monarch butterfly*. All distilled models are 4-step models.

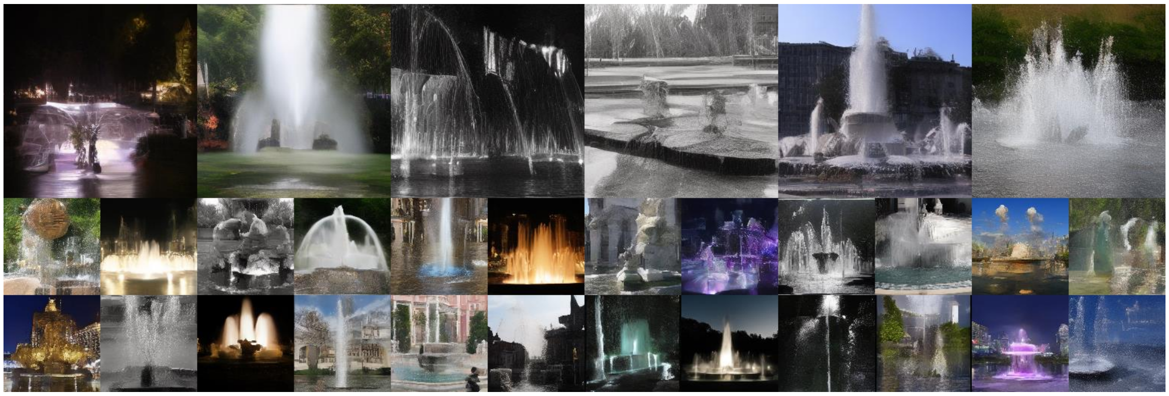
(a) Step Distillation (R) / FID: 10.25

(b) ARD (R) / FID: 4.32

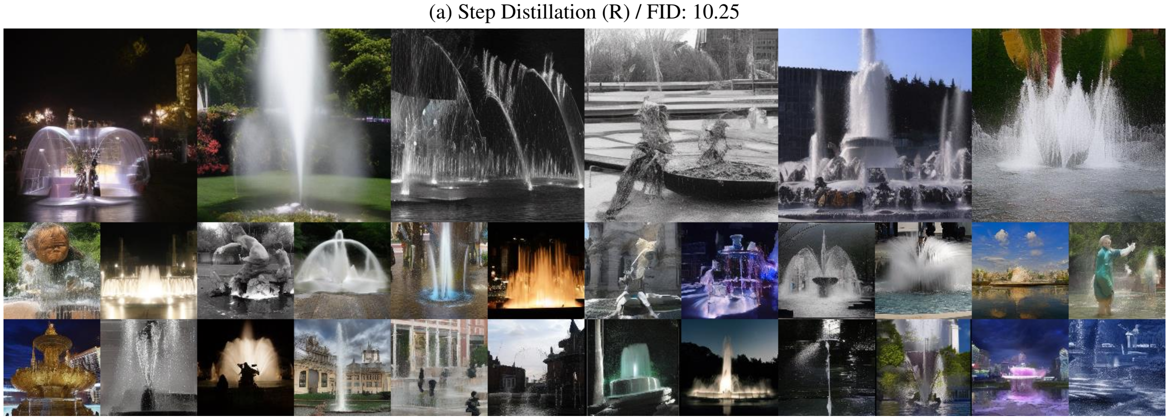
(c) Teacher (25 steps) / FID: 2.89

(d) ARD (R+D) / FID: 1.84

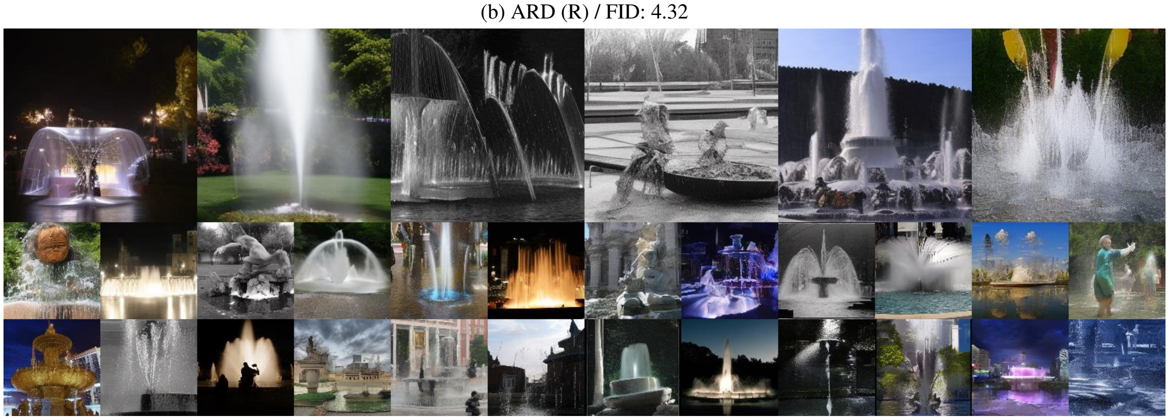
Figure 18. Randomly generated ImageNet 256p samples for class *balloon*. All distilled models are 4-step models.



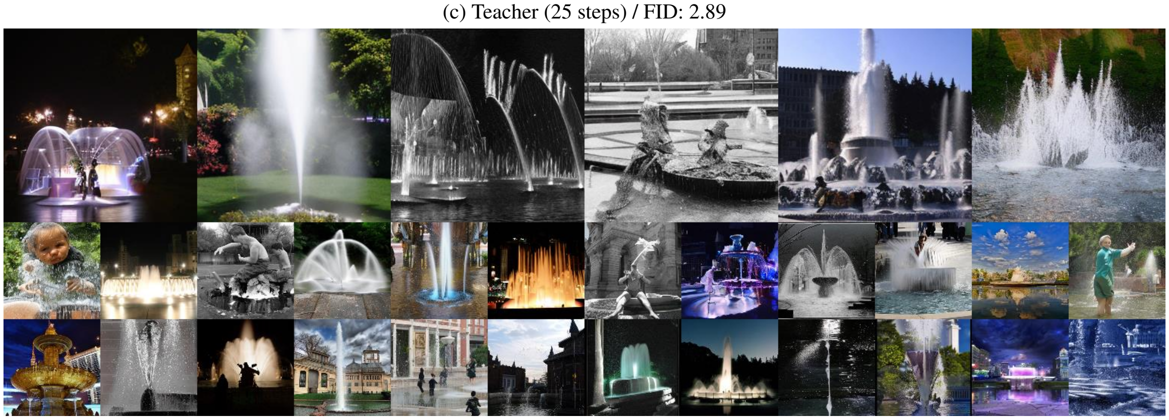
(a) Step Distillation (R) / FID: 10.25



(b) ARD (R) / FID: 4.32



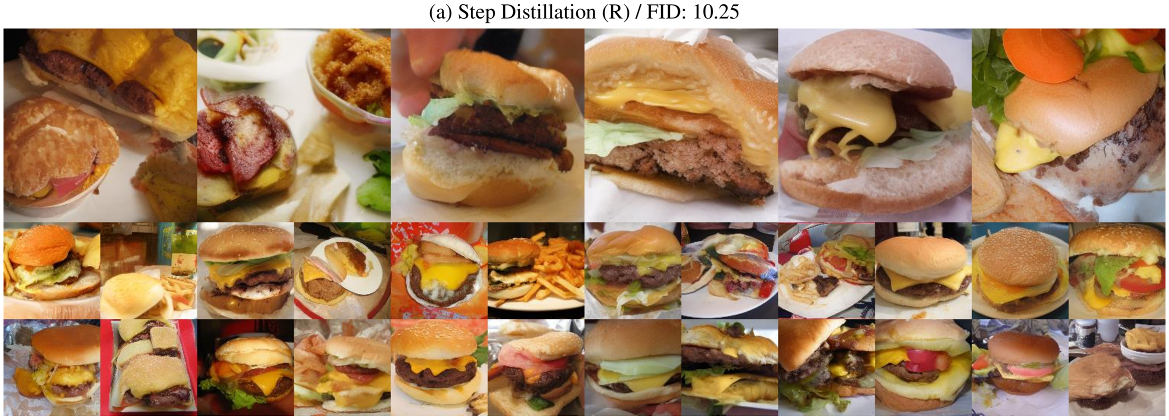
(c) Teacher (25 steps) / FID: 2.89



(d) ARD (R+D) / FID: 1.84

Figure 19. Randomly generated ImageNet 256p samples for class *fountain*. All distilled models are 4-step models.

(a) Step Distillation (R) / FID: 10.25



(b) ARD (R) / FID: 4.32

(c) Teacher (25 steps) / FID: 2.89

(d) ARD (R+D) / FID: 1.84

Figure 20. Randomly generated ImageNet 256p samples for class *cheeseburger*. All distilled models are 4-step models.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 2

[2] David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023. 1

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 6

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 6, 1

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 2

[6] Junsong Chen, YU Jincheng, GE Chongjian, 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 *The Twelfth International Conference on Learning Representations*, 2024. 2, 5

[7] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 7, 5

[8] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2, 8, 5

[9] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021. 2

[10] 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

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 1

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 6

[13] Sander Dieleman. Diffusion is spectral autoregression, 2024. 4

[14] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. GENIE: Higher-order denoising diffusion solvers. In *Advances in Neural Information Processing Systems*, 2022. 1

[15] 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. 6, 1, 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[17] Jiatao Gu, Chen Wang, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M. Susskind. Data-free distillation of diffusion models with bootstrapping. In *Forty-first International Conference on Machine Learning*, 2024. 2, 1

[18] Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, and Shuangfei Zhai. Dart: Denoising autoregressive transformer for scalable text-to-image generation, 2024. 1

[19] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2024. 4

[20] 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

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6, 8

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 8

[24] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 6

[25] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling Diffusion Models into Conditional GANs. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[26] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022. 1

[27] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 1

[28] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. PagoDA: Progressive growing of a one-step generator from a low-resolution diffusion teacher. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[29] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36, 2023. 6

[30] Jonas Kohler, Albert Pumarola, Edgar Schönfeld, Artsiom Sanakoyeu, Roshan Sumbaly, Peter Vajda, and Ali Thabet. Imagine flash: Accelerating emu diffusion models with backward distillation. *arXiv preprint arXiv:2405.05224*, 2024. 7, 8, 1, 5

[31] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 6

[32] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 29–37. JMLR Workshop and Conference Proceedings, 2011. 3

[33] 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. 6, 1

[34] Sangyun Lee, Beomsu Kim, and Jong Chul Ye. Minimizing trajectory curvature of ODE-based generative models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18957–18973. PMLR, 2023. 1

[35] Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[36] Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis, 2023. 1

[37] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 6

[38] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 6

[39] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 7, 8, 1

[40] 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. 8

[41] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[42] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[43] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and qiang liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[44] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 2, 1

[45] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 2, 5, 6, 1

[46] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1

[47] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 7, 8, 5

[48] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023. 1

[49] Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-Jun Qi. One-step diffusion distillation through score implicit matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[50] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 5

[51] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2, 3, 8, 1

[52] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition*, pages 7007–7016, 2024. 6, 1

[53] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[54] 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. 2, 6, 7

[55] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[56] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5

[57] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, 2016. 2, 3

[58] Rohit Rao. Announcing ssd-1b: A leap in efficient t2i generation, 2023. 5

[59] 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. 6, 1

[60] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[61] 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, 6

[62] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 2, 3, 5, 6, 7, 1

[63] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[64] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 2

[65] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 6

[66] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR, 2023. 2

[67] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation, 2024. 1, 2

[68] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 7, 1

[69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 4

[70] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[72] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023. 2, 1

[73] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 1

[74] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[75] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 6

[76] Keyu Tian, Yi Jiang, Zehuan Yuan, BINGYUE PENG, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 4, 6, 1

[77] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 1

[78] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 1

[79] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 2

[80] 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. 1

[81] Sirui Xie, Zhisheng Xiao, Diederik P Kingma, Tingbo Hou, Ying Nian Wu, Kevin Patrick Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. EM distillation for one-step diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[82] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7, 8

[83] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6613–6623, 2024. 1

[84] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification. 1

[85] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[86] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International conference on machine learning*, pages 42390–42402. PMLR, 2023. 2, 1