A. Implementation details

Table 1 shows the implementation details for DiT experiments in JAX. Typical DiT-B training run with 200,000 iterations takes 2 days using A6000 GPUs, while DiT-XL training takes 1 day using H100 GPUs. Hyperparameters and checkpoints for experiments with SiT follow the unofficial MeanFlow reimplementation in PyTorch[1]. All SiT models have been trained with 200,000 iterations from the provided checkpoints using 1e-4 learning rate.

Table 1. **Implementation details and hyperparameters.**

Dataset	CelebA-HQ-256	ImageNet-256	
Architecture	DiT-B	DiT-B	DiT-XL
Patch size	2×2	2×2	2×2
Hidden size	768	768	1152
Attention heads	12	12	16
MLP hidden ratio	4	4	4
Blocks $\{L_{\text{min}}, L\}$	$\{4, 12\}$	$\{4, 12\}$	$\{12, 28\}$
Default G, K	$4, 1/8$	$4, 1/8$	$8, 1/8$
Training iterations	$100K \rightarrow 100K$		
Learning rate	1e-4 $\rightarrow$ 1e-5		
Batch size	128	256	256
Initial weights	from scratch	from [1] checkpoint	
Schedule	90% const. with 5,000 iter. warmup $\rightarrow$ 10% cosine decay		
Optimizer	AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay $= 0.1$		
EMA ratio	0.999		
Class. free guid.	0	1.5	1.5
GPUs	4×A6000	8×A6000	8×H100
Generation	uncond.	class-cond.	class-cond.
Classes	1	1000	1000
ODE$_l$ solver	Euler by the neural network		
ODE$_t$ solver	Euler/Dopri5	Euler	Euler
EMA weights	✓	✓	✓

B. Qualitative results

Our additional qualitative sampling results for CelebA-HQ and ImageNet-1K with DiT-B and SiT-XL architectures are shown in Figures 1-2 using 3×3 image grids. The generation time steps $T = 128, 4, 1$ decrease horizontally from left to right. The neural network depth decreases vertically from top to bottom with $l = 12, 8, 4$ for DiT-B and $l = 28, 20$ for SiT-XL.

[1] MeanFlow reimplementation at github.com/zhuyu-cs/MeanFlow

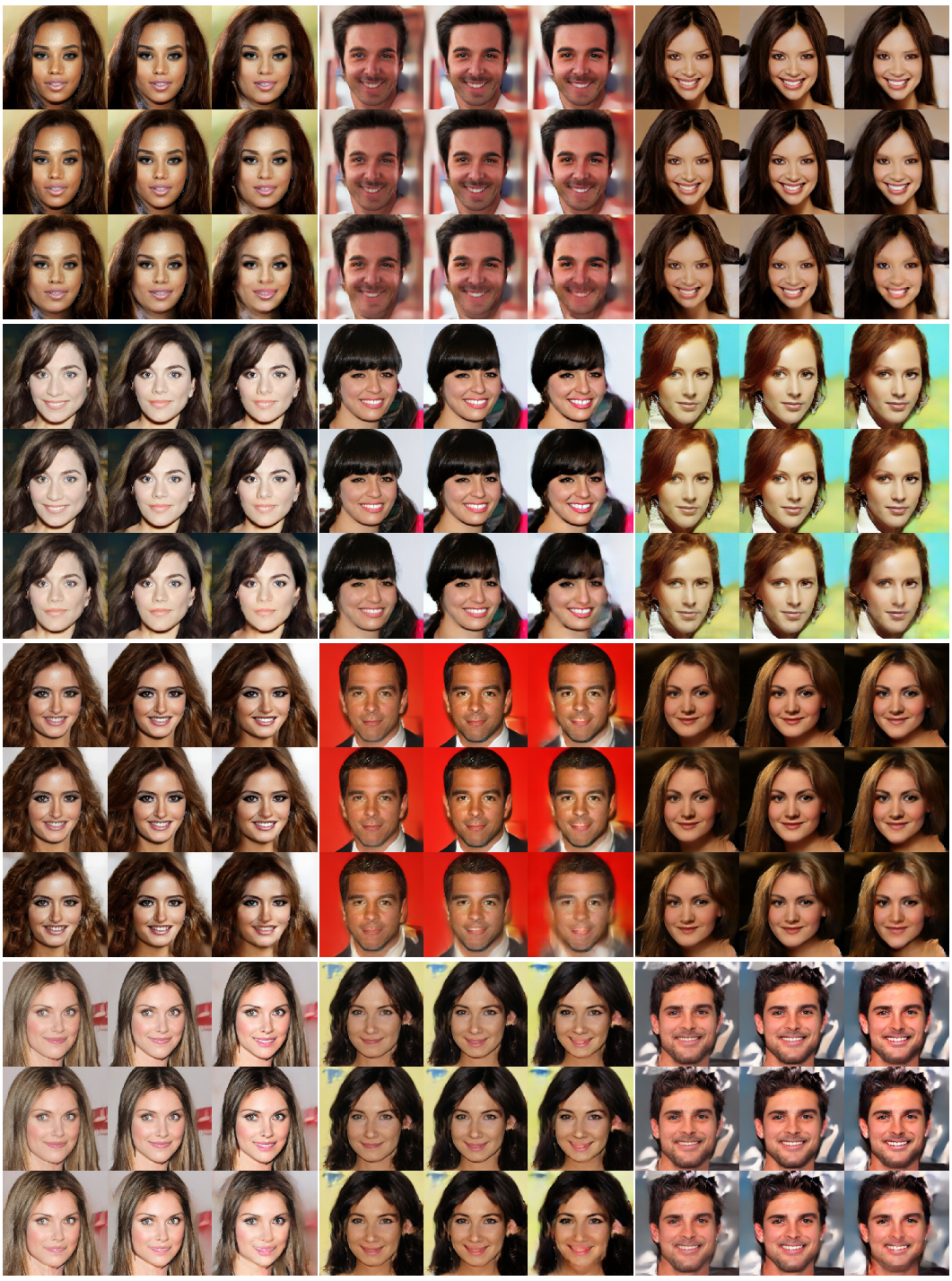

Figure 1. CelebA-HQ-256 images generated by $\text{ODE}_t(\text{ODE}_l)$ with DiT-B. Time shortcuts [1] ($T = 128, 4, 1$ from left to right) alter the image style, our length shortcuts ($l = 12, 8, 4$ from top to bottom) preserve the style and iteratively compress the image details. The bottom right corner combines both approaches with artifacts of two types and the benefit of extremely fast sampling with a factor of $400\times$ lower latency w.r.t. the top left corner.

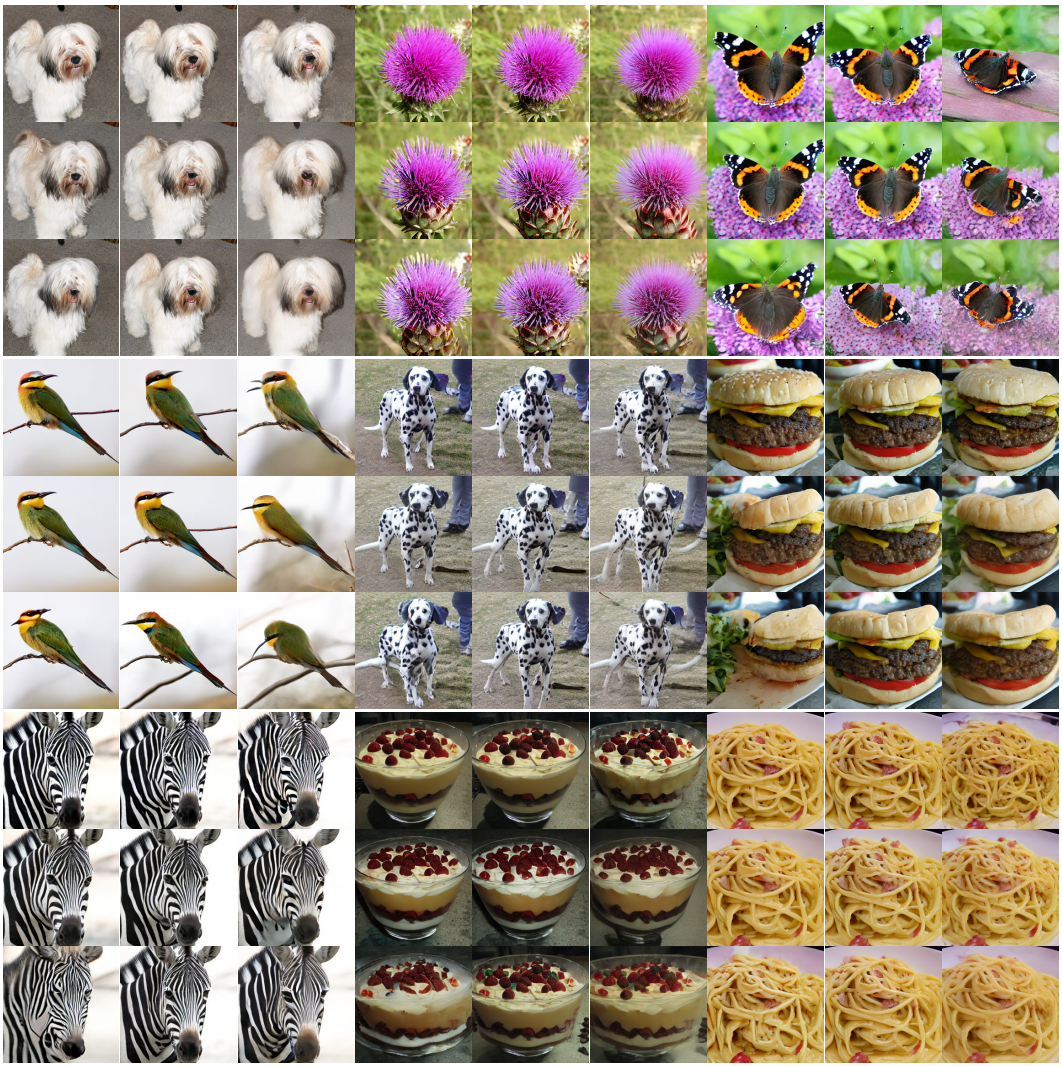

Figure 2. ImageNet-256 images generated by the MF [2] and $\text{ODE}_t(\text{ODE}_l)$ with SiT-XL. Euler steps $T = 128, 4, 1$ decrease from left to right. Top row is the MF [2] baseline, while our $\text{ODE}_t(\text{ODE}_l)$ is showed in the second row with $l = 28$ and the last row with $l = 20$. Our $\text{ODE}_t(\text{ODE}_{l=20})$ provides competitive quality when $T > 1$ (NFE) at 30% lower computational complexity.

References

[1] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *ICLR*, 2025. 1, 2

[2] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *NeurIPS*, 2025. 3