Appendix

We include additional derivations, experimental details, and results in the appendix.

- In Sec. A, we provide a detailed formula derivation of the State Transition Identity, and summarize the training and sampling algorithms.
- In Sec. B, we discuss TiM's relationships with existing methods, including diffusion models, consistency models, and other training approaches.
- In Sec. C, we provide the implementation details of text-to-image generation, including native-resolution training, resolution-dependent timestep shifting, model-guidance training, and from-scratch training.
- In Sec. D, we provide additional ablation results on class-guided image generation.
- In Sec. E, we provide more qualitative results of TiM.

A. Transition Model Framework

In this section, we first provide the derivation of the TiM identity equation Eq. (8). Then we provide the training and sampling algorithms. Finally, we provide a systematic analysis of the connections with other existing methods.

A.1. TiM Identity Equation Derivation

We demonstrate the derivation from Eq. (7) to the TiM identity equation Eq. (8). We start from the detailed expansion of each term of Eq. (7). Firstly, we have:

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\varepsilon}, \tag{14}$$

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t}\boldsymbol{x} + \frac{\mathrm{d}\sigma_t}{\mathrm{d}t}\boldsymbol{\varepsilon}, \tag{15}$$

where $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}$ is the PF-ODE of diffusion and Eq. (15) has already been proved in previous works [45, 47, 71, 74]. For $A_{t,r} = \frac{\alpha_r\hat{\sigma}_t - \sigma_r\hat{\alpha}_t}{\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t}$, $B_{t,r} = \frac{\sigma_r\alpha_t - \alpha_r\sigma_t}{\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t}$, we have:

$$\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t} = \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\alpha_t}\cdot\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} + \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\sigma_t}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t} + \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\alpha}_t}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} + \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\sigma}_t}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \tag{16}$$

$$\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t} = \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\alpha_t}\cdot\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} + \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\sigma_t}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t} + \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\hat{\alpha}_t}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} + \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\hat{\sigma}_t}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \tag{17}$$

We use $C_{t,r} = \hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t$ for simplicity, which is the denominator of $A_{t,r}$ and $B_{t,r}$. For Eq. (16) and Eq. (17), each term is calculated as:

$$\begin{cases} \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\alpha_t} = \frac{(\sigma_r\hat{\alpha}_t - \alpha_r\hat{\sigma}_t)\hat{\sigma}_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = -\hat{\sigma}_t\frac{A_{t,r}}{C_{t,r}} \\ \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\sigma_t} = \frac{(\alpha_r\hat{\sigma}_t - \sigma_r\hat{\alpha}_t)\hat{\alpha}_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = \hat{\alpha}_t\frac{A_{t,r}}{C_{t,r}} \\ \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\alpha}_t} = \frac{(\alpha_r\sigma_t - \sigma_r\alpha_t)\sigma_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = -\hat{\sigma}_t\frac{B_{t,r}}{C_{t,r}} \\ \frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\sigma}_t} = \frac{(\alpha_t\sigma_r - \alpha_r\sigma_t)\alpha_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = \hat{\alpha}_t\frac{B_{t,r}}{C_{t,r}} \end{cases}; \quad \begin{cases} \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\alpha_t} = \frac{(\alpha_r\hat{\sigma}_t - \sigma_r\hat{\alpha}_t)\sigma_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = \sigma_t\frac{A_{t,r}}{C_{t,r}} \\ \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\sigma_t} = \frac{(\sigma_r\hat{\alpha}_t - \alpha_r\hat{\sigma}_t)\alpha_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = -\alpha_t\frac{A_{t,r}}{C_{t,r}} \\ \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\hat{\alpha}_t} = \frac{(\alpha_r\sigma_t - \sigma_r\alpha_t)\sigma_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = \sigma_t\frac{B_{t,r}}{C_{t,r}} \\ \frac{\mathrm{d}B_{t,r}}{\mathrm{d}\hat{\sigma}_t} = \frac{(\alpha_r\sigma_t - \sigma_r\alpha_t)\alpha_t}{(\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t)^2} = -\alpha_t\frac{B_{t,r}}{C_{t,r}} \end{cases}. \tag{18}$$

Substituting Eq. (18) into Eq. (16) and Eq. (17), we have:

$$\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t} = -\hat{\sigma}_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} + \hat{\alpha}_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t} - \hat{\sigma}_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} + \hat{\alpha}_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}, \tag{19}$$

$$\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t} = \sigma_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \alpha_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t} + \sigma_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} - \alpha_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}. \tag{20}$$

There exists some symmetry between the above two equations, which is the key to our TiM identity. Combining Eqs. (14), (15) and (19), we have:

$$\boldsymbol{x}_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t} + A_{t,r}\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = (A_{t,r}\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} + \alpha_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t})\boldsymbol{x} + (A_{t,r}\frac{\mathrm{d}\sigma_t}{\mathrm{d}t} + \sigma_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t})\boldsymbol{\varepsilon}. \tag{21}$$

The coefficient of $\boldsymbol{x}$ in the above equation can be decomposed as:

$$
\begin{aligned}
&A_{t,r}\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}+\alpha_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t}\\
=&(A_{t,r}+\alpha_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\alpha_t})\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}+\alpha_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\sigma_t}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}+\alpha_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\alpha}_t}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}+\alpha_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\sigma}_t}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\\
=&-\hat{\alpha}_t\sigma_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}+\hat{\alpha}_t\alpha_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}-\hat{\sigma}_t\alpha_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}+\hat{\alpha}_t\alpha_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\\
=&-\hat{\alpha}_t\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}+(\hat{\alpha}_t\sigma_t-\hat{\sigma}_t\alpha_t)\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\hat{\alpha}_t}{\mathrm{d}t}\\
=&-\hat{\alpha}_t\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}-B_{t,r}\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}.
\end{aligned}
\tag{22}
$$

Similarly, the coefficient of $\boldsymbol{\varepsilon}$ in the Eq. (21) can be decomposed as:

$$
\begin{aligned}
&A_{t,r}\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}+\sigma_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t}\\
=&\sigma_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\alpha_t}\cdot\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}+(A_{t,r}+\sigma_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\sigma_t})\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\sigma_t}\cdot\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}+\sigma_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\alpha}_t}\cdot\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}+\sigma_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}\hat{\sigma}_t}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\\
=&-\hat{\sigma}_t\sigma_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}+\hat{\sigma}_t\alpha_t\frac{A_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}-\hat{\sigma}_t\sigma_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}+\hat{\alpha}_t\sigma_t\frac{B_{t,r}}{C_{t,r}}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\\
=&-\hat{\sigma}_t\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}+(\hat{\alpha}_t\sigma_t-\hat{\sigma}_t\alpha_t)\frac{B_{t,r}}{\mathrm{d}t}\cdot\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\\
=&-\hat{\sigma}_t\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}-B_{t,r}\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}.
\end{aligned}
\tag{23}
$$

Substituting Eqs. (21) to (23) into Eq. (7), we have:

$$
\begin{aligned}
&\boldsymbol{x}_t\frac{\mathrm{d}A_{t,r}}{\mathrm{d}t}+A_{t,r}\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}+\boldsymbol{f}_{\theta,t,r}\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}+B_{t,r}\frac{\mathrm{d}\boldsymbol{f}_{\theta,t,r}}{\mathrm{d}t}=0.\\
\Rightarrow&(-\hat{\alpha}_t\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}-B_{t,r}\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t})\boldsymbol{x}+(-\hat{\sigma}_t\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}-B_{t,r}\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t})\boldsymbol{\varepsilon}+\boldsymbol{f}_{\theta,t,r}\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}+B_{t,r}\frac{\mathrm{d}\boldsymbol{f}_{\theta,t,r}}{\mathrm{d}t}=0.\\
\Rightarrow&(\hat{\alpha}_t\boldsymbol{x}+\hat{\sigma}_t\boldsymbol{\varepsilon}-\boldsymbol{f}_{\theta,t,r})\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}+B_{t,r}(\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}+\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}-\frac{\mathrm{d}\boldsymbol{f}_{\theta,t,r}}{\mathrm{d}t})=0\\
\Rightarrow&(\hat{\alpha}_t\boldsymbol{x}+\hat{\sigma}_t\boldsymbol{\varepsilon}-\boldsymbol{f}_{\theta,t,r})\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}+B_{t,r}\frac{\mathrm{d}(\hat{\alpha}_t\boldsymbol{x}+\hat{\sigma}_t\boldsymbol{\varepsilon}-\boldsymbol{f}_{\theta,t,r})}{\mathrm{d}t}=0\\
\Rightarrow&\frac{\mathrm{d}(B_{t,r}\cdot(\hat{\alpha}_t\boldsymbol{x}+\hat{\sigma}_t\boldsymbol{\varepsilon}-\boldsymbol{f}_{\theta,t,r}))}{\mathrm{d}t}=0.
\end{aligned}
\tag{24}
$$

This is the TiM identity equation in Eq. (8), the proof is completed.

A.2. TiM Training Algorithm

We provide the detailed training algorithm of TiM in Algorithm 1. It is noteworthy that the TiM models are entirely trained from scratch.

A.3. TiM Sampling Algorithms

We provide the TiM sampling algorithm in Algorithm 2. For few-step sampling, we use deterministic sampling process to ensure fidelity. For multi-step sampling, we can further incorporate stochasticity into the sampling process for improved diversity. In multi-step scenarios, the TiM sampling is similar to the diffusion sampling process, but with a new condition for the next step. Therefore, we can construct a stochastic sampling from the SDE (stochastic differential equation) diffusion process. Given $\boldsymbol{x}_t=\alpha_t+\sigma_t\boldsymbol{\varepsilon}$, Song et al. [71] has shown that the SDE forward and reverse are:

$$
\begin{aligned}
&\text{forward}:\ \mathrm{d}\boldsymbol{x}_t=\mathrm{f}(\boldsymbol{x}_t,t)+\mathrm{g}(t)\mathrm{d}\mathbf{w},\\
&\text{reverse}:\ \mathrm{d}\boldsymbol{x}_t=[\mathrm{f}(\boldsymbol{x}_t,t)-\frac{1}{2}\mathrm{g}(t)^2\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)]\mathrm{d}t+\mathrm{g}(t)\mathrm{d}\mathbf{w}.
\end{aligned}
\tag{25}
$$

Algorithm 1 Training Algorithm of Transition Models (TiM).

Input: dataset $\mathcal{D}$ with standard deviation σ_{d}, model $\boldsymbol{f}_\theta$, diffusion parameterization $\{\alpha_t, \sigma_t, \hat{\alpha}_t, \hat{\sigma}_t\}$, weighting w, learning rate η, time distribution $\mathcal{T}$, constant ϵ, constant c.
Init: Iters $\leftarrow 0$
repeat
  $\boldsymbol{x}_{\text{d}} \sim \mathcal{D},\ \boldsymbol{x} = c_{\text{data}}(\boldsymbol{x}_{\text{d}}),\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ r < t \sim \mathcal{T},\ \boldsymbol{x}_t \leftarrow \alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\varepsilon}$
  $B_{t,r} \leftarrow (\sigma_r \alpha_t - \alpha_r \sigma_t)/(\hat{\sigma}_t \alpha_t - \hat{\alpha}_t \sigma_t)$
  $\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}}{\mathrm{d}t} = \frac{1}{2\epsilon}(\boldsymbol{f}_{\theta^-}(\boldsymbol{x}_{t+\epsilon}, t+\epsilon, r) - \boldsymbol{f}_{\theta^-}(\boldsymbol{x}_{t-\epsilon}, t-\epsilon, r))$ ▷ DDE Calculation
  $\hat{\boldsymbol{f}} \leftarrow \hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon} + (\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} \cdot \boldsymbol{x} + \frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \cdot \boldsymbol{\varepsilon} - \frac{\mathrm{d}\boldsymbol{f}_{\theta^-}}{\mathrm{d}t}) \cdot B_{t,r} / \frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}$ ▷ TiM Target
  $\mathcal{L}(\theta) \leftarrow \|\boldsymbol{f}_\theta - \hat{\boldsymbol{f}}\|_2^2 + L_{\cos}(\boldsymbol{f}_\theta, \hat{\boldsymbol{f}})$
  $\mathcal{L}(\theta) \leftarrow w(t,r) \cdot \mathcal{L}(\theta)/(\|\mathcal{L}(\theta)\| + c)$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$
  Iters $\leftarrow$ Iters $+ 1$
until convergence

Algorithm 2 Sampling Algorithm of Transition Models (TiM).

Input: sampling step N, mixiumum timestep T_{max}, model $\boldsymbol{f}_\theta$, diffusion parameterization $\{\alpha_t, \sigma_t, \hat{\alpha}_t, \hat{\sigma}_t\}$, stochasticity ratio ρ.
Init: data $\boldsymbol{x}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, timesteps $\mathcal{T} = \{t_i\}_{i=N}^0$ where $t_N = T_{\text{max}}, t_0 = 0$
for $i = N$ to 1 **do**
  $\boldsymbol{x}_{t_{i-1}} = \frac{\alpha_{t_{i-1}}\hat{\sigma}_{t_i} - \sigma_{t_{i-1}}\hat{\alpha}_{t_i}}{\hat{\sigma}_{t_i}\alpha_{t_i} - \hat{\alpha}_{t_i}\sigma_{t_i}} \boldsymbol{x}_{t_i} + \frac{\sigma_{t_{i-1}}\alpha_{t_i} - \alpha_{t_{i-1}}\sigma_{t_i}}{\hat{\sigma}_{t_i}\alpha_{t_i} - \hat{\alpha}_{t_i}\sigma_{t_i}} \boldsymbol{f}(\boldsymbol{x}_{t_i}, t_i, t_{i-1})$
  if $\rho > 0$ **then**:
    $\hat{\boldsymbol{\varepsilon}} \leftarrow \frac{\alpha_{t_i}}{\hat{\sigma}_{t_i}\alpha_{t_i}\hat{\alpha}_{t_i}\sigma_{t_i}} \boldsymbol{f}_\theta(\boldsymbol{x}_{t_i}, t_i, t_0) - \frac{\hat{\alpha}_{t_i}}{\hat{\sigma}_{t_i}\alpha_{t_i} - \hat{\alpha}_{t_i}\sigma_{t_i}} \boldsymbol{x}_{t_i}$
    $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathrm{d}t = t_i - t_{i-1}$
    $\boldsymbol{x}_{t_{i-1}} \leftarrow \boldsymbol{x}_{t_{i-1}} - \rho(\alpha_{t_i}\dot{\sigma}_{t_i} - \dot{\alpha_{t_i}}\sigma_{t_i})\hat{\boldsymbol{\varepsilon}}\mathrm{d}t - \sqrt{2\rho(\alpha_{t_i}\dot{\sigma}_{t_i} - \dot{\alpha}_{t_i}\sigma_{t_i})}\boldsymbol{\varepsilon}_i\sqrt{\mathrm{d}t}$
  end if
  $\boldsymbol{x}_i = \boldsymbol{x}_{t_{i-1}}$
end for

Previous works[37, 47, 71] has provided the explicit form of $\mathrm{f}(\boldsymbol{x}_t, t)$, $\mathrm{g}(t)$ and $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$:

$$\mathrm{f}(\boldsymbol{x}_t, t) = \frac{\dot{\alpha}_t}{\alpha_t}\boldsymbol{x}_t, \quad \mathrm{g}(t) = 2\sigma_t\dot{\sigma}_t - 2\frac{\dot{\alpha}_t}{\alpha_t}\sigma_t^2, \quad \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = -\frac{\boldsymbol{\varepsilon}}{\sigma_t}, \tag{26}$$

where $\dot{\alpha}_t$ and $\dot{\sigma}_t$ represent the derivation of α_t and σ_t respectively. For PF-ODE, it is defined as:

$$\mathbf{v}_t = \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \mathrm{f}(\boldsymbol{x}_t, t) - \frac{1}{2}\mathrm{g}(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \dot{\alpha}_t \boldsymbol{x} + \dot{\sigma}_t \boldsymbol{\varepsilon}. \tag{27}$$

For reverse-SDE, it is defined as:

$$\begin{aligned}
\mathrm{d}\boldsymbol{x}_t &= [\mathrm{f}(\boldsymbol{x}_t, t) - \mathrm{g}(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)]\mathrm{d}t + \mathrm{g}(t)\mathrm{d}\mathbf{w} \\
&= \underbrace{[\mathrm{f}(\boldsymbol{x}_t, t) - \frac{1}{2}\mathrm{g}(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)]\,\mathrm{d}t}_{\text{PF-ODE Term}} \underbrace{-\frac{1}{2}\mathrm{g}(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\mathrm{d}t + \mathrm{g}(t)\mathrm{d}\mathbf{w}}_{\text{Stochastic Term}} \\
&= \mathbf{v}_t \mathrm{d}t + [\dot{\sigma}_t - \frac{\dot{\alpha}_t}{\alpha_t}\sigma_t]\boldsymbol{\varepsilon}\mathrm{d}t + \sqrt{2\sigma_t\dot{\sigma}_t - 2\frac{\dot{\alpha}_t}{\alpha_t}\sigma_t^2}\mathrm{d}\mathbf{w}.
\end{aligned} \tag{28}$$

In the TiM sampling, we can take the stochastic term in the above equation to enhance diversity. To balance the stochasticity and stability, we incorporate a scaling factor $s(t) = \rho\alpha_t$, leading to a scaled $\tilde{\mathrm{g}} = \rho\alpha_t \mathrm{g}(t) = 2\rho(\alpha_t\sigma_t\dot{\sigma}_t - \dot{\alpha}_t\sigma_t^2)$. Therefore, the stochastic term is: $\rho[\alpha_t\dot{\sigma}_t - \dot{\alpha}_t\sigma_t]\boldsymbol{\varepsilon}\mathrm{d}t + \sqrt{2\rho(\alpha_t\sigma_t\dot{\sigma}_t - \dot{\alpha}_t\sigma_t^2)}\mathrm{d}\mathbf{w}$.

B. Connections with Existing Methods

In this section, we highlight the connection between TiM and other existing methods. We first demonstrate the properties of TiM compared with diffusion models. Then we demonstrate the connections of TiM with other training strategies.

Table 9. Time distribution of diffusion diffusion transports.

Transport	Noise Level	Timestep	Time Range	Time Scaling
OT-FM [45, 46]	$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$	$t = \frac{\sigma}{1+\sigma}$	$t \in [0, 1]$	$c_{\text{noise}}(t) = t$
Trigflow [47]	$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$	$t = \arctan(\frac{\sigma}{\sigma_{\text{d}}})$	$t \in [0, \frac{\pi}{2}]$	$c_{\text{noise}}(t) = t$
EDM [37, 38]	$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$	$t = \sigma$	$t \in [0, +\infty)$	$c_{\text{noise}}(t) = \frac{1}{4}\ln(t)$
VP [32, 71]	$\sigma \sim \mathcal{U}(\epsilon_t, 1)$	$t = \sigma$	$t \in [\epsilon_t, 1]$	$c_{\text{noise}}(t) = (T-1)t$
VE [70, 71]	$\sigma \sim \mathcal{U}(\epsilon_t, 1)$	$t = \sigma_{\max}(\frac{\sigma_{\min}^2}{\sigma_{\max}^2})^{\sigma}$	$t \in [\sigma_{\min}, \sigma_{\max}]$	$c_{\text{noise}}(t) = \ln(\frac{1}{2}t)$
TiM (Ours)	$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$	$t = \frac{\sigma}{1+\sigma}$	$t \in [0, 1]$	$c_{\text{noise}}(t) = t$

B.1. Connections with Diffusion Models

TiM generalizes the standard diffusion models. We elucidate the diffusion parameterization and transition parameterization in Tab. 1. As a complement to Tab. 1, the time distribution of different diffusion transports is summarized in Tab. 9. Our TiMs share these parameters with diffusion models, but learn a different objective. We show that the TiM training objective Eq. (13) generalizes the standard diffusion objective Eq. (3). Specifically, the TiM identity equation reduces to the diffusion identity equation in the limit as $t \to r$. Recall that $B_{t,r} = \frac{\sigma_r\alpha_t - \alpha_r\sigma_t}{\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t}$, the training target of TiM when $t \to r$ becomes:

$$
\begin{aligned}
\lim_{t\to r} \hat{\boldsymbol{f}} &= \lim_{t\to r} \left(\hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon} + \frac{B_{t,r}}{\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}} \left(\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} \cdot \boldsymbol{x} + \frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \cdot \boldsymbol{\varepsilon} - \frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,r}}{\mathrm{d}t} \right) \right) \\
&= \hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon} + \lim_{t\to r} \left(\frac{B_{t,r}}{\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}} \left(\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} \cdot \boldsymbol{x} + \frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \cdot \boldsymbol{\varepsilon} - \frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,r}}{\mathrm{d}t} \right) \right) \\
&= \hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon} + \lim_{t\to r} \left(\frac{\frac{\sigma_r\alpha_t - \alpha_r\sigma_t}{\hat{\sigma}_t\alpha_t - \hat{\alpha}_t\sigma_t}}{\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}} \left(\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} \cdot \boldsymbol{x} + \frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \cdot \boldsymbol{\varepsilon} - \frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,r}}{\mathrm{d}t} \right) \right) \\
&= \hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon} + \lim_{t\to r} \left(\frac{0}{\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}} \left(\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} \cdot \boldsymbol{x} + \frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \cdot \boldsymbol{\varepsilon} - \frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,r}}{\mathrm{d}t} \right) \right) \\
&= \hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon}.
\end{aligned}
\tag{29}
$$

The above target is the diffusion target. This target lacks the modeling of state transitions from state to state, thus limiting the arbitrary-step generation capabilities of diffusion models.

EDM parametrization. EDM [37, 38] parameterizes the diffusion model as:

$$
\boldsymbol{D}_\theta(\boldsymbol{x} + t\boldsymbol{\varepsilon}, t) = \frac{\sigma_{\text{d}}^2}{t^2 + \sigma_{\text{d}}^2}(\boldsymbol{x} + t\boldsymbol{\varepsilon}) + \frac{t \cdot \sigma_{\text{d}}}{\sqrt{t^2 + \sigma_{\text{d}}^2}} \boldsymbol{F}_\theta \left(\frac{\boldsymbol{x} + t\boldsymbol{\varepsilon}}{\sqrt{t^2 + \sigma_{\text{d}}^2}}, \frac{1}{4}\ln(t) \right). \tag{30}
$$

It adopts the $\boldsymbol{x}$-prediction in its training and use time weighting $w(t) = \frac{t^2+\sigma_\mathrm{d}^2}{t^2\sigma_\mathrm{d}^2}$, leading to training objective as:

$$\begin{aligned}
\mathcal{L}(\theta) &= \frac{t^2+\sigma_\mathrm{d}^2}{t^2\sigma_\mathrm{d}^2}\|\boldsymbol{D}_\theta(\boldsymbol{x}+t\boldsymbol{\varepsilon},t)-\boldsymbol{x}\|_2^2 \\
&= \frac{t^2+\sigma_\mathrm{d}^2}{t^2\sigma_\mathrm{d}^2}\left\|\frac{\sigma_\mathrm{d}^2}{t^2+\sigma_\mathrm{d}^2}(\boldsymbol{x}+t\boldsymbol{\varepsilon})+\frac{t\cdot\sigma_\mathrm{d}}{\sqrt{t^2+\sigma_\mathrm{d}^2}}\boldsymbol{F}_\theta\left(\frac{\boldsymbol{x}+t\boldsymbol{\varepsilon}}{\sqrt{t^2+\sigma_\mathrm{d}^2}},\frac{1}{4}\ln(t)\right)-\boldsymbol{x}\right\|_2^2 \\
&= \left\|\boldsymbol{F}_\theta\left(\frac{\boldsymbol{x}+t\boldsymbol{\varepsilon}}{\sqrt{t^2+\sigma_\mathrm{d}^2}},\frac{1}{4}\ln(t)\right)-\left(\frac{t}{\sigma_\mathrm{d}\sqrt{t^2+\sigma_\mathrm{d}^2}}\boldsymbol{x}-\frac{\sigma_\mathrm{d}}{\sqrt{t^2+\sigma_\mathrm{d}^2}}\boldsymbol{\varepsilon}\right)\right\|_2^2
\end{aligned} \tag{31}$$

Therefore, let $c_\mathrm{noise}(t) = \frac{1}{4}\ln(t)$, the original EDM parameterization can be unified into our parameterization with the following coefficients:

$$\alpha_t = \frac{1}{\sqrt{t^2+\sigma_\mathrm{d}^2}},\quad \sigma_t = \frac{t}{\sqrt{t^2+\sigma_\mathrm{d}^2}},\quad \hat{\alpha}_t = \frac{t}{\sigma_\mathrm{d}\sqrt{t^2+\sigma_\mathrm{d}^2}}\quad \hat{\sigma}_t = -\frac{\sigma_\mathrm{d}}{\sqrt{t^2+\sigma_\mathrm{d}^2}} \tag{32}$$

Therefore, the TiM parameterization is defined as:

$$\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} = -\frac{t^2}{\sigma_\mathrm{d}(t^2+\sigma_\mathrm{d}^2)^{\frac{3}{2}}}+\frac{1}{\sigma_\mathrm{d}\sqrt{t^2+\sigma_\mathrm{d}^2}}, \tag{33}$$

$$\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} = \frac{t\sigma_\mathrm{d}}{(t^2+\sigma_\mathrm{d}^2)^{\frac{3}{2}}}, \tag{34}$$

$$B_{t,r} = \frac{(t-r)\sigma_\mathrm{d}^2\sqrt{t^2+\sigma_\mathrm{d}^2}}{(t^2+\sigma_\mathrm{d}^3)\sqrt{r^2+\sigma_\mathrm{d}^2}}, \tag{35}$$

$$\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t} = \sigma_\mathrm{d}^2\frac{t(t-r)(t^2+\sigma_\mathrm{d}^3)-2t(t-r)(t^2+\sigma_\mathrm{d}^2)+(t^2+\sigma_\mathrm{d}^2)(t^2+\sigma_\mathrm{d}^3)}{(t^2+\sigma_\mathrm{d}^3)^2\sqrt{t^2+\sigma_\mathrm{d}^2}\sqrt{r^2+\sigma_\mathrm{d}^2}}. \tag{36}$$

B.2. Connections to Other training Methods

In this section, we discuss the connections of TiM with other training strategies, including continuous-time consistency models [47, 72], consistency trajectory models [39], phased consistency models [78], Shortcut models [20], and MeanFlow models [26, 55].

Continuous-time consistency models. The TiM objective Eq. (13) generalizes the continuous-time consistency models. Specifically, the CTM objective reduces to the continuous-time CM objective when $r = 0$. For TiM, let $r = 0$ and $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, the training objective becomes:

$$\begin{aligned}
&\nabla_\theta\mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t}\left[\left\|\boldsymbol{f}_\theta(\boldsymbol{x}_t,t,0)-(\hat{\alpha}_t\boldsymbol{x}+\hat{\sigma}_t\boldsymbol{\varepsilon}+\frac{B_{t,0}}{\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}}\left(\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}\cdot\boldsymbol{x}+\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\cdot\boldsymbol{\varepsilon}-\frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,0}}{\mathrm{d}t}\right)\right\|_2^2\right] \\
=&\mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t}\left[[\nabla_\theta\boldsymbol{f}_{\theta,t,0}]^T\left(\boldsymbol{f}_{\theta^-,t,0}-\hat{\alpha}_t\boldsymbol{x}-\hat{\sigma}_t\boldsymbol{\varepsilon}+\frac{B_{t,0}}{\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}}\left(\frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,0}}{\mathrm{d}t}-\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}\cdot\boldsymbol{x}-\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}\cdot\boldsymbol{\varepsilon}\right)\right)\right].
\end{aligned} \tag{37}$$

Continuous-time consistency models [47, 69, 72] are trained to map the noisy input $\boldsymbol{x}_t$ directly to the clean data $\boldsymbol{x}$ in one or a few steps. Given model F_θ, the consistency models are formulated as:

$$\boldsymbol{D}_\theta(\boldsymbol{x}_t,t) = c_\mathrm{skip}(t)\boldsymbol{x}_t + c_\mathrm{out}(t)\boldsymbol{F}_\theta(c_\mathrm{in}(t)\boldsymbol{x}_t, c_\mathrm{noise}(t)). \tag{38}$$

Using the parameters α_t, σ_t, $\hat{\alpha}_t$, and $\hat{\sigma}_t$, consistency parameterization corresponds to the transition from $\boldsymbol{x}_t$ to $\boldsymbol{x}_0$:

$$\boldsymbol{D}_\theta(\boldsymbol{x}_t,t) = \frac{\hat{\sigma}_t\boldsymbol{x}_t-\sigma_t\boldsymbol{F}_\theta(c_\mathrm{in}(t)\boldsymbol{x}_t,c_\mathrm{noise}(t))}{\hat{\sigma}_t\alpha_t-\hat{\alpha}_t\sigma_t}, \tag{39}$$

where $\frac{\hat{\sigma}_t}{\hat{\sigma}_t\alpha_t-\hat{\alpha}_t\sigma_t}=A_{t,0}$ and $-\frac{\sigma_t}{\hat{\sigma}_t\alpha_t-\hat{\alpha}_t\sigma_t}=B_{t,0}$ correspond to TiM parameterizations.

When using loss function $d(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, [72] show that the gradient of continuous-time consistency models is:

$$\begin{aligned}
&\nabla_\theta\mathbb{E}_{\boldsymbol{x}_t,t}\left[\boldsymbol{D}_\theta^T(\boldsymbol{x}_t,t)\frac{\mathrm{d}\boldsymbol{D}_{\theta^-}(\boldsymbol{x}_t,t)}{\mathrm{d}t}\right]\\
=&\nabla_\theta\mathbb{E}_{\boldsymbol{x}_t,t}\left[[B_{t,0}\boldsymbol{f}_\theta^{\mathrm{cm}}(\boldsymbol{x}_t,t)]^T\frac{\mathrm{d}\boldsymbol{D}_{\theta^-}(\boldsymbol{x}_t,t)}{\mathrm{d}t}\right]\\
=&\mathbb{E}_{\boldsymbol{x}_t,t}\left[B_{t,0}[\nabla_\theta\boldsymbol{f}_\theta^{\mathrm{cm}}(\boldsymbol{x}_t,t)]^T\frac{\mathrm{d}\boldsymbol{D}_{\theta^-}(\boldsymbol{x}_t,t)}{\mathrm{d}t}\right]\\
=&\mathbb{E}_{\boldsymbol{x}_t,t}\left[B_{t,0}\nabla_\theta[\boldsymbol{f}_\theta^{\mathrm{cm}}(\boldsymbol{x}_t,t)]^T\left(\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}\boldsymbol{x}_t+A_{t,0}\frac{\boldsymbol{x}_t}{\mathrm{d}t}+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}+B_{t,0}\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}\right)\right],
\end{aligned}\tag{40}$$

where $\boldsymbol{f}_\theta^{\mathrm{cm}}(\boldsymbol{x}_t,t)=\boldsymbol{F}_\theta(c_{\mathrm{in}}(t)\boldsymbol{x}_t,c_{\mathrm{noise}}(t)))$ represents the network in consistency models. As $\boldsymbol{x}_t=\alpha_t\boldsymbol{x}+\sigma_t\boldsymbol{\varepsilon}$, $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}=\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}\boldsymbol{x}+\frac{\sigma_t}{\mathrm{d}t}\boldsymbol{\varepsilon}$, we have:

$$\begin{aligned}
&\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}\boldsymbol{x}_t+A_{t,0}\frac{\boldsymbol{x}_t}{\mathrm{d}t}+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}+B_{t,0}\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}\\
=&\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}(\alpha_t\boldsymbol{x}+\sigma_t\boldsymbol{\varepsilon})+A_{t,0}(\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}\boldsymbol{x}+\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}z)+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}+B_{t,0}\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}\\
=&(\alpha_t\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}+A_{t,0}\frac{\mathrm{d}\alpha_t}{\mathrm{d}t})\boldsymbol{x}+(\sigma_t\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}+A_{t,0}\frac{\mathrm{d}\sigma_t}{\mathrm{d}t})\boldsymbol{z}+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}+B_{t,0}\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}.
\end{aligned}\tag{41}$$

Based on Eqs. (22) and (23), we have:

$$\alpha_t\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}+A_{t,0}\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}=-\hat{\alpha}_t\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}-B_{t,0}\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t},\quad \sigma_t\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}+A_{t,0}\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}=-\hat{\sigma}_t\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}-B_{t,0}\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t}.\tag{42}$$

Therefore, we have:

$$\begin{aligned}
&\frac{\mathrm{d}A_{t,0}}{\mathrm{d}t}\boldsymbol{x}_t+A_{t,0}\frac{\boldsymbol{x}_t}{\mathrm{d}t}+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}+B_{t,0}\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}\\
=&B_{t,0}(\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}-\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}\boldsymbol{x}-\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t})+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}(\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}-\hat{\alpha}_t\boldsymbol{x}-\hat{\sigma}_t\boldsymbol{z}).
\end{aligned}\tag{43}$$

Substituting the above equation to Eq. (40), the gradient of continuous-time consistency models is:

$$\begin{aligned}
&\nabla_\theta\mathbb{E}_{\boldsymbol{x}_t,t}\left[\boldsymbol{D}_\theta^T(\boldsymbol{x}_t,t)\frac{\mathrm{d}\boldsymbol{D}_{\theta^-}(\boldsymbol{x}_t,t)}{\mathrm{d}t}\right]\\
=&\mathbb{E}_{\boldsymbol{x}_t,t}\left[B_{t,0}[\nabla_\theta\boldsymbol{f}_\theta^{\mathrm{cm}}]^T\left(B_{t,0}(\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}-\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}\boldsymbol{x}-\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t})+\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}(\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}-\hat{\alpha}_t\boldsymbol{x}-\hat{\sigma}_t\boldsymbol{z})\right)\right],\\
=&\mathbb{E}_{\boldsymbol{x}_t,t}\left[(B_{t,0}\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t})[\nabla_\theta\boldsymbol{f}_\theta^{\mathrm{cm}}]^T\left(\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}-\hat{\alpha}_t\boldsymbol{x}-\hat{\sigma}_t\boldsymbol{z}+\frac{B_{t,0}}{\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t}}(\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}^{\mathrm{cm}}}{\mathrm{d}t}-\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t}\boldsymbol{x}-\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t})\right)\right],
\end{aligned}\tag{44}$$

Note that TiM network $\boldsymbol{f}_\theta(\boldsymbol{x}_t,t,0)$ corresponds to the network $\boldsymbol{f}_\theta^{\mathrm{cm}}(\boldsymbol{x}_t,t)$ in consistency models. The only difference between Eq. (44) and Eq. (37) is a term $(B_{t,0}\frac{\mathrm{d}B_{t,0}}{\mathrm{d}t})$, which can be bridged by a weighting function.

Consistency trajectory models, phased consistency models, and shortcut models. These models learn to transition from one state to another state in a discrete manner, while our TiM generalizes this to the continuous-time domain. The core of our method is the TiM identity equation Eq. (8), which determines the function for the transition between two arbitrary states. For consistency trajectory models (CTM) [39] and phased consistency models (PCM) [78], they targets at intermediate state $\boldsymbol{x}_r$, where $0\leqslant r\leqslant t_{n-1}$, thus leading to the identity equation:

$$\Psi(\boldsymbol{x}_{t_n},\boldsymbol{f}_\theta(\boldsymbol{x}_{t_n},t_n,r),r)=\Psi(\boldsymbol{x}_{t_{n-1}},\boldsymbol{f}_\theta(\boldsymbol{x}_{t_{n-1}},t_{n-1},r),\tag{45}$$

where $0<t_1<\cdots<t_n<\cdots<t_N=T$ represents the discrete timesteps, and Ψ is an ODE solver to obtain the state at timemstep r. It is noteworthy that PCM splits the entire trajectory into several segments and learns this identity on each segment independently.

Shortcut models [20] adopts the OT-flow-matching [45] as the transport, the ODE solver is: $\Psi(\boldsymbol{x}_t, \boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r), r) = \boldsymbol{x}_t - (t-r)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r)$. The original identity equation is: $(t-r)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r) = (t-s)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, s) + (s-r)\boldsymbol{f}_\theta(\boldsymbol{x}_s, s, r)$. This identity equation of shortcut models can be rearranged as:

$$\begin{aligned}
&(t-r)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r) = (t-s)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, s) + (s-r)\boldsymbol{f}_\theta(\boldsymbol{x}_s, s, r) \\
\Longrightarrow &\boldsymbol{x}_t - (t-r)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r) = \boldsymbol{x}_t - (t-s)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, s) - (s-r)\boldsymbol{f}_\theta(\boldsymbol{x}_s, s, r) \\
\Longrightarrow &\boldsymbol{x}_t - (t-r)\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r) = \boldsymbol{x}_s - (s-r)\boldsymbol{f}_\theta(\boldsymbol{x}_s, s, r) \\
\Longrightarrow &\Psi(\boldsymbol{x}_t, \boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r), r) = \Psi(\boldsymbol{x}_s, \boldsymbol{f}_\theta(\boldsymbol{x}_s, s, r), r),
\end{aligned} \tag{46}$$

where $s = \frac{t+r}{2}$. Based on Eq. (45) and Eq. (46), when $t_{n-1} = \frac{t_n + r}{2}$, CTMs are equivalent to shortcut models.

MeanFlow models. We demonstrate that the TiM Eq. (8) generalizes the MeanFlow [26]. In particular, the training objective of TiM reduces to the MeanFlow objective in the OT-FM [45] transport setting.

As in Tab. 1, OT-FM uses the parameterization $\{\alpha_t = 1-t, \sigma_t = t, \hat{\alpha}_t = -1, \hat{\sigma}_t = 1\}$, leading to the TiM parameterization $\{B_{t,r} = r - t, \frac{\mathrm{d}B_{t,r}}{\mathrm{d}t} = -1\}$. Therefore, the TiM training objective becomes:

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t}\left[d\left(\boldsymbol{f}_\theta(\boldsymbol{x}_t, t, r) - \left(\boldsymbol{z} - \boldsymbol{x} - (t-r)\frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,r}}{\mathrm{d}t}\right)\right)\right]. \tag{47}$$

This corresponds to the training objective of MeanFlow.

C. Implementation Details

We present more details of implementation in this section.

C.1. Text-to-Image Training Details

Native-Resolution Training. We adopt VAE-specific native resolution training for text-to-image generation. As we use DC-AE [13] with 32 downsampling scale, an image with shape $H \times W$ is resized to shape $(32 \cdot \lfloor \frac{H}{32} \rfloor) \times (32 \cdot \lfloor \frac{W}{32} \rfloor)$. For example, an image with shape 1025×513 is resized to 1024×512, , preserving resolution and aspect ratio as much as possible. Images of the same size are grouped into resolution buckets for batching.

We set the base batch size as 16 on a single GPU for 1024×1024 resolution bucket, then for $H \times W$ resolution bucket, the minimal batch size is $B = \lfloor \frac{16 \times H \times W}{1024 \times 1024} \rfloor$. For instance, the 512×512 resolution bucket holds the minimal batch size as $B = 64$, while the 2048×2048-resolution bucket holds the minimal batch size as $B = 4$. The maximum resolution is 4096×4096 with $B = 1$. For data parallelism, each device processes distinct buckets with their corresponding batch sizes, maintaining a similar token budget.

Resolution-Dependent Timestep Shifting. Sampling from a single timestep distribution is suboptimal across resolutions ranging from less than 256×256 to 4096×4096 pixels. Intuitively, higher-resolution images require stronger corruption (more noise) to destroy the signal, while lower-resolution images require less noise. Given an image with $n = H_1 \times W_1$ pixels and its high-resolution counterpart with $m = H_2 \times W_2$ pixels, Esser et al. [19] provides an equation to map the the timestep t_n to t_m:

$$t_m = \frac{\sqrt{\frac{m}{n}} t_n}{1 + (\sqrt{\frac{m}{n}} - 1)t_n}. \tag{48}$$

In our practice, we set the base pixel number as $n = 1024 \times 1024$, and apply this mapping to all sampled timesteps.

Model-Guidance Training. Tang et al. [75] propose a model-guidance training target to improve sampling fidelity. We adopt this approach for text-to-image training. Under our formulation, the target becomes:

$$\hat{\boldsymbol{f}} = \hat{\alpha}_t \boldsymbol{x} + \hat{\sigma}_t \boldsymbol{\varepsilon} + \frac{B_{t,r}}{\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t}}\left[\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} \cdot \boldsymbol{x} + \frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} \cdot \boldsymbol{\varepsilon} - \frac{\mathrm{d}\boldsymbol{f}_{\theta^-,t,r}}{\mathrm{d}t}\right] + (\omega - 1)(\boldsymbol{f}^{\text{cond}}_{\theta^*,t,t} - \boldsymbol{f}^{\text{uncond}}_{\theta^*,t,t}), \tag{49}$$

where ω denotes the Classifier-Free Guidance (CFG) scale, θ^* is the Exponential Moving Average (EMA) of θ, $\boldsymbol{f}^{\text{cond}}_{\theta^*,t,t}$ and $\boldsymbol{f}^{\text{uncond}}_{\theta^*,t,t}$ respectively represent the conditional and unconditional outputs.

Method	ϵ	Speed	1-step	4-step	50-step
JVP	n.a.	1.8 iter/s	49.75	26.22	18.11
DDE	0.0001	2.4 iter/s	111.25	23.34	18.38
DDE	0.0002	2.4 iter/s	80.14	23.83	17.58
DDE	0.0005	2.4 iter/s	67.09	24.33	**16.93**
DDE	0.001	2.4 iter/s	**48.83**	**24.73**	17.03
DDE	0.002	2.4 iter/s	49.07	25.54	17.59
DDE	0.005	2.4 iter/s	49.91	26.09	17.99
DDE	0.01	2.4 iter/s	50.05	26.53	18.33
DDE	0.02	2.4 iter/s	49.72	26.67	18.33
DDE	0.05	2.4 iter/s	49.90	27.05	18.79

Table 10. **The impacts of DDE ϵ on generation performance.**

Transport	***Diffusion Parameterization***				***Transition Parameterization***				***FID*↓**		
	$\alpha_t =$	$\sigma_t =$	$\hat{\alpha}_t =$	$\hat{\sigma}_t =$	$\frac{\mathrm{d}\hat{\alpha}_t}{\mathrm{d}t} =$	$\frac{\mathrm{d}\hat{\sigma}_t}{\mathrm{d}t} =$	$B_{t,r} =$	$\frac{\mathrm{d}B_{t,r}}{\mathrm{d}t} =$	1-NFE	8-NFE	50-NFE
OT-FM [45]	$1-t$	t	-1	1	0	0	$r-t$	-1	49.91	26.09	17.99
TrigFlow [47]	$\cos(t)$	$\sin(t)$	$-\sin(t)$	$\cos(t)$	$-\cos(t)$	$-\sin(t)$	$\sin(r-t)$	$-\cos(r-t)$	67.32	25.14	18.28
EDM [37]	$\frac{1}{t^2+\sigma_\mathrm{d}^2}$	$\frac{t}{\sqrt{t^2+\sigma_\mathrm{d}^2}}$	$\frac{t}{\sigma_\mathrm{d}\sqrt{t^2+\sigma_\mathrm{d}^2}}$	$-\frac{\sigma_\mathrm{d}}{\sqrt{t^2+\sigma_\mathrm{d}^2}}$	Eq. (33)	Eq. (34)	Eq. (35)	Eq. (36)	53.64	37.01	24.06
VP-SDE [32]	$\frac{1}{\beta_t^2+1}$	$\frac{\beta_t}{\sqrt{\beta_t^2+1}}$	0	1	0	0	$\frac{\beta_r-\beta_t}{\sqrt{\beta_r^2+1}}$	$\frac{-1}{\sqrt{\beta_r^2+1}}\cdot\frac{\mathrm{d}\beta_t}{\mathrm{d}t}$	78.98	37.44	35.72
VE-SDE [71]	1	t	0	-1	0	0	$t-r$	1	65.68	65.57	68.04
TiM (Ours)	$\cos(\frac{\pi}{2}t)$	$\sin(\frac{\pi}{2}t)$	-1	1	0	0	$\frac{\sin(\frac{\pi}{2}(r-t))}{\sin(\frac{\pi}{2}t)+\cos(\frac{\pi}{2}t)}$	$-\frac{\sqrt{2}\sin(\frac{\pi}{2}r+\frac{\pi}{4})}{2\sin(\pi t)+2}$	**46.38**	**23.54**	**15.06**
Option-1	$\cos(\frac{\pi}{2}t)$	$\sin(\frac{\pi}{2}t)$	$-\frac{\pi}{2}\sin(\frac{\pi}{2}t)$	$\frac{\pi}{2}\cos(\frac{\pi}{2}t)$	$-\frac{\pi^2}{4}\cos(\frac{\pi}{2}t)$	$-\frac{\pi^2}{4}\sin(\frac{\pi}{2}t)$	$\frac{2}{\pi}\sin(\frac{\pi}{2}(r-t))$	$-\cos(\frac{\pi}{2}(r-t))$	58.54	24.94	16.05
Option-2	$\cos(\frac{\pi}{2}t)$	$\sin(\frac{\pi}{2}t)$	$-\sin(\frac{\pi}{2}t)$	$\cos(\frac{\pi}{2}t)$	$-\frac{\pi}{2}\cos(\frac{\pi}{2}t)$	$-\frac{\pi}{2}\sin(\frac{\pi}{2}t)$	$\sin(\frac{\pi}{2}(r-t))$	$-\frac{\pi}{2}\cos(\frac{\pi}{2}(r-t))$	57.62	24.77	16.09
Option-3	$1-t^2$	t^2	-1	1	0	0	r^2-t^2	$-2t$	55.36	40.52	27.24
Option-4	$1-t^2$	t^2	$-2t$	$2t$	-2	2	$\frac{r^2-t^2}{2t}$	$-\frac{t^2+r^2}{2t^2}$	57.14	41.19	27.59
Option-5	$1-t^2$	t^2	$-t$	t	-1	1	$\frac{r^2-t^2}{t}$	$-\frac{t^2+r^2}{t^2}$	59.46	40.42	26.53
Option-6	$1-\sqrt{t}$	$\sqrt{t}$	-1	1	0	0	$\sqrt{r}-\sqrt{t}$	$-\frac{1}{2\sqrt{t}}$	97.65	38.62	31.11

Table 11. **Transition parameterization for different diffusion transports**. For VP-SDE, T is set to 1000, and $\beta_t = \sqrt{e^{\frac{1}{2}\beta_\mathrm{d}t^2+\beta_\mathrm{min}t}-1}$, where $\beta_\mathrm{d} = 19.9$ and $\beta_\mathrm{min} = 0.1$ by default. Song et al. [71] proves that VP-SDE equals DDPM [32] while VE-SDE equals score matching [70]. The EDM transition parameterization is too complex, so we provide them in Eqs. (33) to (36) in appendix.

Weighting	$\mathtt{transform}(t)=t$			$\mathtt{transform}(t)=t/(1-t)$			$\mathtt{transform}(t)=\tan(t)$		
	NFE=1	NFE=8	NFE=50	NFE=1	NFE=8	NFE=50	NFE=1	NFE=8	NFE=50
(a) Reciprocal	48.47	23.80	15.06	62.66	23.74	14.71	49.97	23.20	15.08
(b) SMS	**47.97**	24.85	16.24	78.87	24.14	**14.60**	47.84	23.23	14.57
(c) Sqrt	48.02	24.78	16.15	**48.38**	23.29	14.49	**46.31**	22.55	**14.21**
(d) Square	48.33	**23.64**	**14.89**	90.73	**22.99**	15.06	49.76	**22.31**	14.32

Table 12. **Ablation studies on different time weighting functions.**

From-Scratch Training. The TiM-T2I model contains $865M$ parameters with the patch size of 1. We train from scratch for about 30 days across 16 NVIDIA-A100 GPUs with a constant learning rate of 4×10^{-4}, using PyTorch-FSDP [88] and half-precision (`torch.bfloat16`) for memory efficiency. Following Tang et al. [75], we use model-guidance target Eq. (49) with CFG scale $\omega = 1.75$ after $100K$ iterations.

D. Additional Results

We provide additional results on class-guided image generation in this section.

D.1. Additional Ablations

We provide additional ablation results in this section.

Differential Derivation Equation Calculation. We incorporate a small quantity ϵ into Eq. (10) to calculate the time derivative of the network. Using the TiM-B/4 model (Config (c)), we systematically evaluated the impact of ϵ on numerical precision in Tab. 10 and observed that $\epsilon \in [0.001, 0.01]$ produces high precision. For training stability, we adopt $\epsilon = 0.005$ in all experiments.

Transport Comparison. We conduct ablation studies on different transports in Tab. 11 using the TiM-B/4 model (Config (c)). Our proposed TiM method demonstrates a consistent and significant performance advantage across all settings. Notably, in the high-quality 50-NFE regime, TiM achieves a state-of-the-art FID of 15.06, surpassing the strong baselines of OT-FM (17.99) and TrigFlow (18.28). The superiority of our method is particularly evident in the highly efficient 1-NFE (one-step) setting, where TiM (46.38) substantially outperforms the next-best method, OT-FM (49.91). Furthermore, the ablation study confirms our design: Options 1 and 2, which share TiM's trigonometric diffusion parameterization, are the most competitive variants, but TiM's novel transition parameterization provides a definitive performance edge. The comparatively poor results of Options 3-6, which employ different diffusion parameterizations, underscore the criticality of our selected formulation.

Time Weighting. Using the TiM-B/4 model (Config (f)), we provide a systematic analysis of time weighting $w(t,r) = k(\tau(t),\tau(r))$. We study three types of transformations: (1) $\tau(t) = t$, (2) $\tau(t) = \frac{t}{1-t}$, (3) $\tau(t) = \tan(t)$; and four types of weighting functions: (1) Reciprocal: $k(t,r) = \frac{1}{\sigma_\text{d}+t-r}$, (2) Soft-Min-SNR (SMS): $k(t,r) = \frac{1}{\sigma_\text{d}^2+(t-r)^2}$, (3) SQRT: $k(t,r) = \frac{1}{\sqrt{\sigma_\text{d}+t-r}}$, (4) Square: $k(t,r) = \frac{1}{(\sigma_\text{d}+t-r)^2}$, where σ_d is the standard deviation of clean data $\boldsymbol{x}$ ($\sigma_\text{d} = 1$ in our dataset). Empirically, the combination $w(t,r) = \frac{1}{\sqrt{\sigma_\text{d}+\tan(t)-\tan(r)}}$ achieves the best performance, slightly exceeding the best results reported in Tab. 8.

D.2. Class-Guided Image Generation

We provide the results of class-guided image generation in this section.

Setup. We use SD-VAE [58] for ImageNet-256 $\times$ 256 and DC-AE [13] for ImageNet-512 $\times$ 512, with patch sizes of 2 and 1, respectively. We train an XL-model with $664M$ parameters for $750K$ iterations with a batch size of 512 (300 epochs), using a constant learning rate of 2×10^{-4} and AdamW optimizer. We report FID [31], sFID [52], Inception Score (IS) [60], Precision and Recall [40] using ADM evaluation suite [18].

Performance Analysis. We provide the results on ImageNet-256 $\times$ 256 and ImageNet-512 $\times$ 512 in Tabs. 13 and 14 respectively. Across both ImageNet-256$\times$256 and ImageNet-512$\times$512, TiM-XL demonstrates strong performance-efficiency trade-offs: at low NFE (1 to 8), it can compete with few-step consistency models, achieving comparable FID with fewer training epochs and smaller model size. When increasing NFEs, TiM-XL matches or surpasses many multi-step diffusion models in FID, despite training for only 300 epochs. Notably, TiM demonstrates remarkable generation quality and shows stable gains as NFE increases.

E. Qualitative Results

We provide the qualitative results in Figs. 4 to 6.

Method	Epochs	Params	NFE	**FID↓**	**sFID↓**	**IS↑**	**Prec.↑**	**Rec.↑**
Generative Adversarial Networks								
BigGAN [8]	-	112M	1	6.95	7.36	171.4	0.87	0.28
StyleGAN-XL [62]	-	166M	1	2.30	4.02	265.12	0.78	0.53
GigaGAN [36]	-	569M	2	3.45	-	225.52	0.84	0.61
Masked and Autoregressive Models								
Mask-GIT [10]	555	-	-	6.18	-	182.1	-	-
MagViT-v2 [86]	1080	307M	-	1.78	-	319.4	-	-
LlamaGen-XL [73]	300	775M	-	2.62	5.59	244.08	0.81	0.58
LlamaGen-XXL [73]	300	1.4B	-	2.34	5.97	253.90	0.81	0.59
LlamaGen-3B [73]	300	3.1B	-	2.18	5.97	263.3	0.81	0.58
VAR [77]	350	2.0B	-	1.80	-	365.4	0.83	0.57
MAR [42]	800	943M	-	1.55	-	303.7	0.81	0.62
RandAR-XL [53]	300	775M	-	2.22	-	314.21	0.80	0.60
RandAR-XXL [53]	300	1.4B	-	2.15	-	321.97	0.79	0.62
Multi-step Diffusion Models								
LDM-4-G [58]	170	395M	500	3.60	5.12	247.67	0.87	0.48
SimpleDiffusion [33]	800	2B	500	2.44	-	256.3	-	
Flag-DiT-3B* [21]	200	4.23B	500	1.96	4.43	284.8	0.82	0.61
Large-DiT-3B* [21]	340	4.23B	500	2.10	4.52	304.36	0.82	0.60
MDT [22]	1300	676M	500	1.79	4.57	283.01	0.81	0.61
MDTv2 [23]	700	676M	500	1.63	4.45	311.73	0.79	0.65
DiT-XL [54]	1400	675M	500	2.27	4.60	278.24	0.83	0.57
SiT-XL [51]	1400	675M	500	2.06	4.49	277.50	0.83	0.59
FlowDCN-XL [79]	400	675M	500	2.00	4.37	263.16	0.82	0.58
SiT-REPA-XL [87]	800	675M	500	1.42	4.70	305.7	0.80	0.65
Few-step Consistency Models								
MeanFlow-XL [26]	1000	675M	1	3.43	-	-	-	-
iCT-XL [69]	-	675M	2	20.30	-	-	-	-
Shortcut-XL [20]	250	675M	2	10.60	-	-	-	-
			8	7.80	-	-	-	-
IMM-XL [90]	3840	675M	2	7.77	-	-	-	-
			4	3.99	-	-	-	-
			8	2.51	-	-	-	-
Any-step Transition Models								
TiM-XL	300	675M	1	3.15	8.49	220.90	0.75	0.61
			2	2.47	6.50	221.41	0.79	0.58
			8	2.06	5.67	239.62	0.79	0.62
			100	1.68	5.22	257.79	0.79	0.66

Table 13. **Performance comparison on ImageNet-**256×256 **class-guided generation.** *: Flag-DiT-3B and Large-DiT-3B actually have 4.23B parameters.

Method	Epochs	Params	NFE	**FID↓**	**sFID↓**	**IS↑**	**Prec.↑**	**Rec.↑**
Generative Adversarial Networks								
BigGAN [8]	-	160M	1	7.5	-	152.8	-	-
StyleGAN-XL [61]	-	168M	2	2.41	4.06	267.75	0.77	0.52
Masked and Autoregressive Models								
VAR [77]	1080	307M	-	2.63	-	303.2	-	-
MAGVITv2 [86]	350	2.3B	-	1.91	-	324.3	-	-
Multi-step Diffusion Models								
SimpleDiffusion [33]	800	2B	500	3.02	-	248.7	-	-
DiffiT [29]	-	561M	500	2.67	-	252.12	0.83	0.55
MaskDiT [89]	800	-	500	2.50	5.10	256.27	0.83	0.56
Large-DiT-3B [21]	368	4.23B	500	2.52	5.01	303.70	0.82	0.57
ADM-G,U [18]	400	774M	500	3.85	5.86	221.72	0.84	0.53
U-ViT-H/2 [4]	400	501M	500	4.05	6.44	263.79	0.84	0.48
DiT-XL [54]	600	675M	500	3.04	5.02	240.82	**0.84**	0.54
EDM2-L [38]	1468	778M	64	1.87	-	-	-	-
EDM2-XL [38]	1048	1.1B	64	1.80	-	-	-	-
EDM2-XXL [38]	734	1.5B	64	1.73	-	-	-	-
SiT-XL [51]	600	675M	500	2.62	4.18	252.21	0.84	0.57
FlowDCN-XL [79]	-	675M	500	2.44	4.53	252.8	0.84	0.54
SiT-REPA-XL [87]	200	675M	500	2.08	4.19	274.6	0.83	0.58
Few-step Consistency Models								
sCT-L [47]	1273	778M	1	5.15	-	-	-	-
			2	4.65	-	-	-	-
sCT-XL [47]	1117	1.1B	1	4.33	-	-	-	-
			2	3.73	-	-	-	-
sCT-XXL [47]	762	1.5B	1	4.29	-	-	-	-
			2	3.76	-	-	-	-
Any-step Transition Models								
TiM-XL	300	675M	1	4.32	4.34	175.15	0.79	0.61
			2	2.77	4.22	206.13	0.80	0.61
			8	2.31	4.17	218.62	0.82	0.59
			64	1.76	4.12	246.59	0.81	0.63

Table 14. **Performance comparison on ImageNet-**512×512 **class-guided generation.**

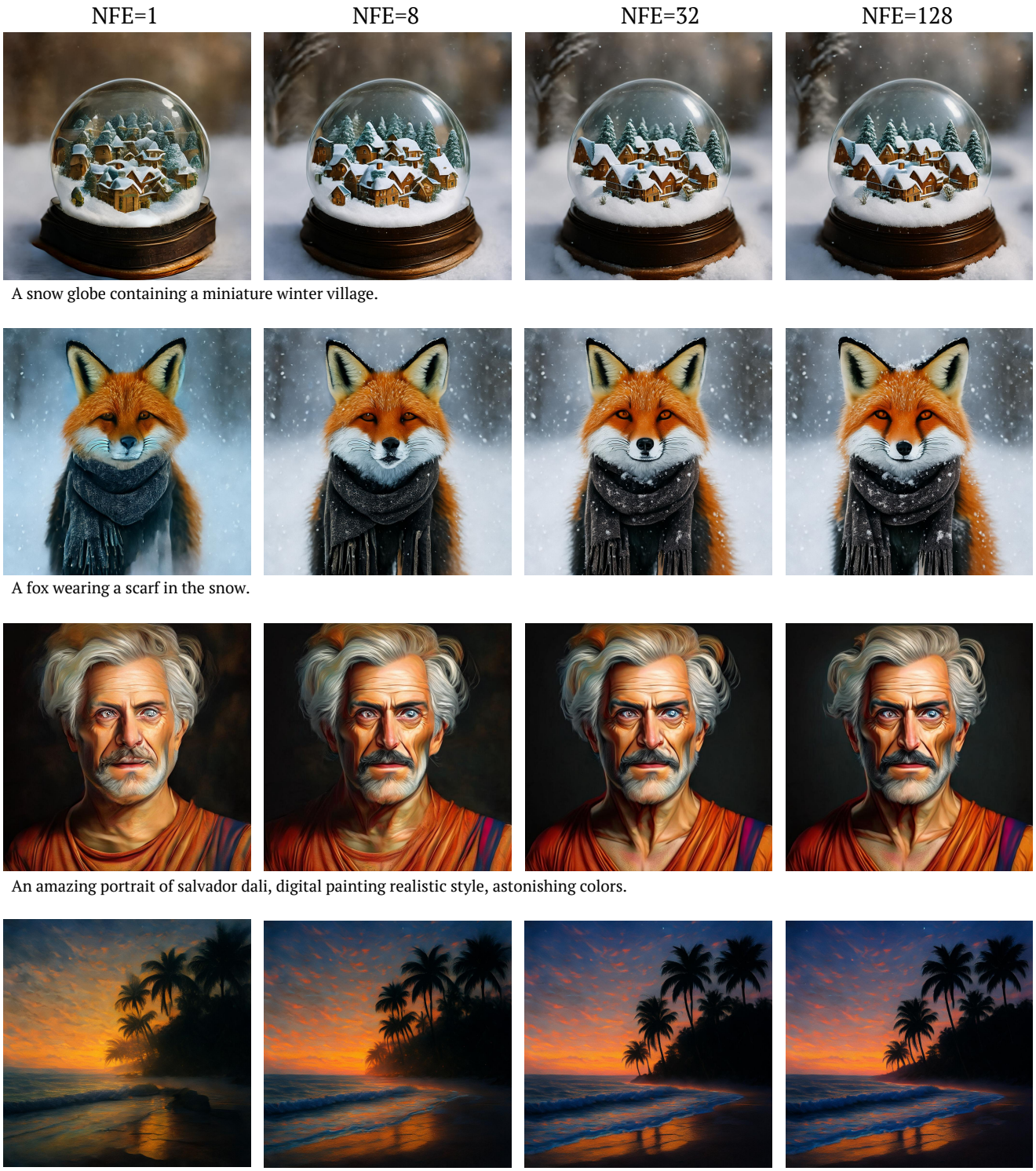

Figure 4. 1024×1024**-resolution generations from TiM with varing NFEs.** TiM supports arbitrary sampling steps and consistently delivers higher-quality images with finer details as the sampling budget increases.

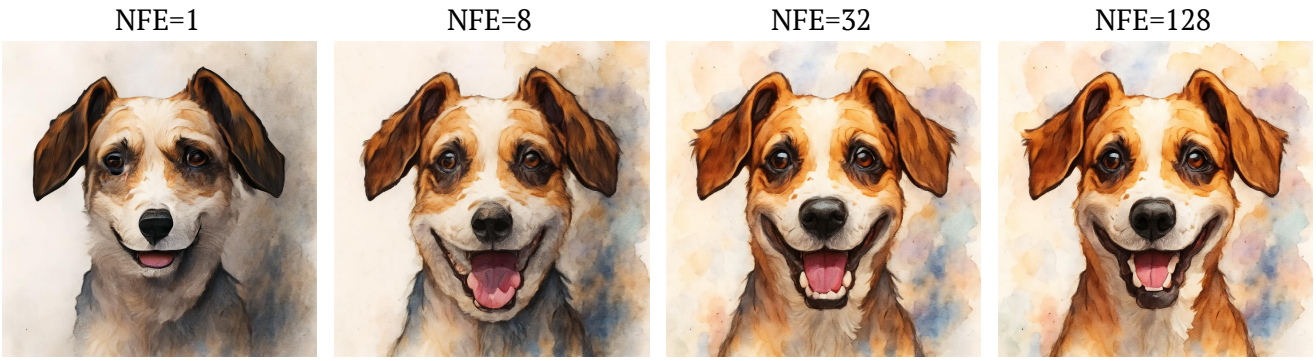

Figure 5. 1024×1024**-resolution generations from TiM with varing NFEs.**

An oil-on-canvas masterpiece that captures the dynamic essence of a blue night sky, bursting with the energy of swirling azure hues and speckled with stars that seem to explode in shades of yellow. Dominating the celestial scene is a luminous, fuzzy-the canvas. Below this cosmic display, a tranquil village is depicted to the right, its quaint houses huddled in repose. On the left, a towering cypress tree stretches upwards, its branches undulating like flames, creating a stark contrast against the serene sky. In the distance, amidst the gentle roll of blue hills, the spire of a church stands tall, a silent sentinel overlooking the sleepy hamlet.

realistic, male, surfer, 80s, chill, relaxed, stoner, aviators, beach, illustration, shaggy hair, round face, round chin, brown hair, photorealistic

knight in full silver armor with a red scarf towering over the camera, sunset in the background, fantasy, mountains, clouds seeming to protude from his figure, photorealistic

portrait of wolf, lots of colour, pen and soft watercolour in style of sarah taylor art

A close-up image of an intricately designed lotus flower, which appears to be crafted entirely from crystal-clear water droplets. The flower is set against a backdrop of soft green lily pads floating on a tranquil pond. Sunlight filters through the scene, highlighting the delicate texture and the shimmering surface of the water-formed petals.

a puppy and a young cat in a cozy room, close up photorealistic image with high details. Picture shows sun flares with warm light

fantasy, a majestic sky filled with stars and galaxies, over looking a serene lake.

Figure 6. **High-resolution and multi-aspect generations from TiM (128 NFEs).** TiM attains up to 4096×4096 resolution generation capability and reliably handles multiple aspect ratios, including 1024×4096 and 2560×1024.