Inference-Time Diffusion Model Distillation

Supplementary Material

The supplementary sections are organized as follows. Section 6 introduces the pseudo training algorithm behind our inference-time diffusion distillation framework. In Section 7, we provide experimental details. Section 8 features additional results. Following this, we delve into the future directions and limitations of the proposed method in Section 9. Code will be released in https://github.com/anony-distillationpp/distillation_pp.

6. Pseudo-code

Algorithm 1 Inference-time Diffusion model distillation

```
1:  Input: Student model $\theta$, Teacher model $\psi$, $N$ sampling steps, $k$ number of steps of teacher guidance, CFG scale $\omega$, Teacher guidance scale $\lambda$.
2:  Output: Improved generation $\boldsymbol{x}_0^*$.
3:
4:  $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{x}_T|0, I), \quad \triangle t = T/N$
5:  for $t = T$ to $\triangle t$ do
6:      Stage 1. Initial student estimation
7:      $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(t) = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_{\theta}^{w}(\boldsymbol{x}_t, \boldsymbol{c})}{\sqrt{\bar{\alpha}_t}}$
8:
9:      Stage 2. Revised teacher estimation
10:     if step $< k$ then
11:         Renoising step $s = t - \triangle t$.
12:         $\boldsymbol{x}_s = \sqrt{\bar{\alpha}_s}\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(t) + \sqrt{1 - \bar{\alpha}_s}\boldsymbol{\epsilon}$. $(\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, I))$
13:         $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\psi}(s) = \frac{\boldsymbol{x}_s - \sqrt{1-\bar{\alpha}_s}\hat{\boldsymbol{\epsilon}}_{\psi}^{w}(\boldsymbol{x}_s, \boldsymbol{c})}{\sqrt{\bar{\alpha}_s}}$.
14:         $\hat{\boldsymbol{x}}_{\text{new},\boldsymbol{c}}^{\theta}(t) = (1 - \lambda)\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(t) + \lambda\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\psi}(s)$
15:     else
16:         $\hat{\boldsymbol{x}}_{\text{new},\boldsymbol{c}}^{\theta}(t) = \hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(t)$
17:     end if
18:     Update $\boldsymbol{x}_{t-\triangle t}$ by forwarding $\hat{\boldsymbol{x}}_{\text{new},\boldsymbol{c}}^{\theta}(t)$.
19: end for
```

We fix typo in (11) with $\tilde{\boldsymbol{x}}_t \to \boldsymbol{x}_t$. Also, we note that $k = 1$ used in every quantitative analysis, ensuring computational efficiency. As the proposed framework revises estimation by interpolation, it can be seamlessly extended with a convex combination of multiple teacher revisions. Moreover, as Algorithm 1 is described with random renoising strategy (Line 12), it is fully compatible with general student models ranging from ones directly predicting the PF-ODE endpoint to the progressive distillation branches.

7. Experimental details

7.1. Extension to other solvers

For completeness, we extend Distillation++ to accommodate a broader range of ODE/SDE solvers. The core principle lies in steering the *denoising* process with teacher models. Specifically, we consider solving the variance-exploding (VE) PF-ODE, commonly employed in standard diffusion model implementations[1], which can be readily derived via reparameterization of VP diffusion models. Following the notation in Lu et al. [26], we consider a sequence of timesteps $\{t_i\}_{i=0}^{M}$, where $t_0 = T$ denotes the initial starting point of the reverse sampling (i.e. Gaussian noise).

Euler [15]. This is in line with DDIM [40] and thus included for completeness:

$$\boldsymbol{x}_{t_{i+1}} = \hat{\boldsymbol{x}}_{\text{new},\boldsymbol{c}}^{\theta}(\boldsymbol{x}_{t_i}) + \frac{\boldsymbol{x}_{t_i} - \hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(\boldsymbol{x}_{t_i})}{\sigma_{t_i}} \cdot \sigma_{t_{i+1}},$$

where $\hat{\boldsymbol{x}}_{\text{new},\boldsymbol{c}}^{\theta}(\boldsymbol{x}_{t_i})$ refers to the revised estimate by interpolation. CFG++ [5] can be integrated by replacing $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(\boldsymbol{x}_{t_i})$ with $\hat{\boldsymbol{x}}_{\varnothing}^{\theta}(\boldsymbol{x}_{t_i})$.

Euler Ancestral. The Euler Ancestral sampler extends the Euler method by introducing stochasticity, taking larger steps and adding a small random noise. This may potentially improve sampling diversity:

$$\boldsymbol{x}_{t_{i+1}} = \hat{\boldsymbol{x}}_{\text{new},\boldsymbol{c}}^{\theta}(\boldsymbol{x}_{t_i}) + \frac{\boldsymbol{x}_{t_i} - \hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(\boldsymbol{x}_{t_i})}{\sigma_{t_i}} \cdot (\sigma_{t_{d_i}} - \sigma_{t_i}) + \sigma_{t_i}\boldsymbol{\epsilon},$$

where $t_i > t_{d_i} > t_{i+1}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, I)$.

DPM-solver++ 2S [26]. We consider DPM++ 2S with CFG++ [5] in VE-SDE setting [41]. Specifically, define $\sigma_t := e^{-t}$, $h_i := t_i - t_{i-1}$, $r_i := h_{i-1}/h_i$ and initialize $\boldsymbol{x}_{t_0}$ with standard Gaussian noise. DPM-solver++ 2S introduces an additional intermediate time step $\{s_i\}_{i=1}^{M}$ with $t_i > s_{i+1} > t_{i+1}$. Let $r_i = \frac{s_i - t_{i-1}}{t_i - t_{i-1}}$. Then, an iterate of DPM-solver++ 2S with CFG++ reads:

$$\begin{aligned}
\boldsymbol{u}_i &= e^{-r_i h_i}\boldsymbol{x}_{t_{i-1}} + (1 - e^{-r_i h_i})\hat{\boldsymbol{x}}_{\varnothing}^{\theta}(\boldsymbol{x}_{t_{i-1}}), \\
\boldsymbol{x}_{t_i} &= \hat{\boldsymbol{x}}_{\varnothing}^{\theta}(\boldsymbol{x}_{t_{i-1}}) - e^{-h_i}\hat{\boldsymbol{x}}_{\varnothing}^{\theta}(\boldsymbol{x}_{t_{i-1}}) \\
&\quad + \frac{1 - e^{-h_i}}{2r_i}\left(\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(\boldsymbol{u}_i) - \hat{\boldsymbol{x}}_{\varnothing}^{\theta}(\boldsymbol{x}_{t_{i-1}})\right) + e^{-h_i}\boldsymbol{x}_{t_{i-1}},
\end{aligned}$$

where $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(\boldsymbol{u}_i)$ refers to the initial conditional denoised estimate with guidance scale $0 < \lambda < 1$, and the rest terms are related to the higher-order correction of the renoising process. That said, distillation++ modulates the denoising process in (12) by interpolating $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\theta}(\boldsymbol{u}_i)$ with the teacher-

[1] https://github.com/crowsonkb/k-diffusion

revised estimate $\hat{\boldsymbol{x}}^{\psi}_{\boldsymbol{c}}(\boldsymbol{u}_i)$ as $\hat{\boldsymbol{x}}^{\theta}_{\text{new},\boldsymbol{c}}(\boldsymbol{u}_i)$:

$$\begin{aligned}\boldsymbol{u}_i &= e^{-r_i h_i}\boldsymbol{x}_{t_{i-1}} + (1 - e^{-r_i h_i})\hat{\boldsymbol{x}}^{\theta}_{\varnothing}(\boldsymbol{x}_{t_{i-1}}),\\ \boldsymbol{x}_{t_i} &= \hat{\boldsymbol{x}}^{\theta}_{\varnothing}(\boldsymbol{x}_{t_{i-1}}) - e^{-h_i}\hat{\boldsymbol{x}}^{\theta}_{\varnothing}(\boldsymbol{x}_{t_{i-1}}) + \\ &\frac{1-e^{-h_i}}{2r_i}\left(\hat{\boldsymbol{x}}^{\theta}_{\text{new},\boldsymbol{c}}(\boldsymbol{u}_i) - \hat{\boldsymbol{x}}^{\theta}_{\varnothing}(\boldsymbol{x}_{t_{i-1}})\right) + e^{-h_i}\boldsymbol{x}_{t_{i-1}}.\end{aligned} \tag{12}$$

This simple modification implies that Distillation++ is potentially compatible with various solvers, where the core principle is to regularize the denoising path with revised teacher's estimates. We use ancestral variant (DPM++ 2S A) of (12) by adding stochasticity in practice.

DPM-solver++ 2M [26]. While many student models support only first-order solvers, customized distillation models in the open-source community[2] are compatible with higher-order solvers like DPM-Solver++ 2M. Using an iterative process initialized with Gaussian noise, DPM-solver++ refines the sampling trajectory with higher-order corrections, enabling precise updates. Similarly as DPM-solver++ 2S, define $\sigma_t := e^{-t}$, $h_i := t_i - t_{i-1}$, and $r_i := h_{i-1}/h_i$. Given $\boldsymbol{x}_{t_0}$ initialized as Gaussian noise, the first iteration reads:

$$\boldsymbol{x}_{t_1} = \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_0}) + e^{-h_1}(\boldsymbol{x}_{t_0} - \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_0})).$$

Then, the following provides higher-order correction:

$$\boldsymbol{D}_i = \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}}) + \frac{1}{2r_i}\left(\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}}) - \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-2}})\right), \tag{13}$$

$$\boldsymbol{x}_{t_i} = e^{-h_i}\boldsymbol{x}_{t_{i-1}} - (e^{-h_i} - 1)\boldsymbol{D}_i. \tag{14}$$

Rearranging (13), (14), we can rewrite the update steps as

$$\begin{aligned}\boldsymbol{x}_{t_i} &= \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}}) - e^{-h_i}\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}})\\ &+ \frac{1-e^{-h_i}}{2r_i}\left(\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}}) - \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-2}})\right) + e^{-h_i}\boldsymbol{x}_{t_{i-1}}.\end{aligned}$$

As we are interested in modulating the final form of denoised estimates, Distillation++ can be applied as follows:

$$\begin{aligned}\boldsymbol{x}_{t_i} &= \hat{\boldsymbol{x}}^{\theta}_{\text{new},\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}}) - e^{-h_i}\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}})\\ &+ \frac{1-e^{-h_i}}{2r_i}\left(\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-1}}) - \hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_{t_{i-2}})\right) + e^{-h_i}\boldsymbol{x}_{t_{i-1}}.\end{aligned}$$

Ancestral variants (DPM-solver++ 2M A) can be readily derived by similarly adding a random noise.

7.2. Experiment Setup

In this work, we employ several diffusion distillation models: DMD2 [47], SDXL-Turbo [38], SDXL-Lightning [21], LCM [28], LCM-LoRA [29], and SDXL-Lightning LoRA. These models rely on classifier-free guidance (CFG) with a fixed guidance scale during training. We use $w = 7.5$ for CFG with teacher models $\hat{\boldsymbol{x}}^{\psi}_{\boldsymbol{c}}(s)$.

Baselines using 4 sampling steps include SDXL-Lightning, DMD2, and SDXL-Turbo, while LCM and LCM-LoRA use 8 sampling steps. Specifically, for Table 1, we use 4 step Euler sampling for SDXL-Lightning and its LoRA variant, 4 step iterative random sampling for DMD2, LCM, and LCM-LoRA (incompatible with conventional solvers), and 4 step DPM++ 2S Ancestral sampling [26] for SDXL-Turbo, utilizing DreamShaper [31], an open-source customized model from the community.

All quantitative results presented in the main paper are obtained using Algorithm 1, which employs a fully random renoising process (Line 12) to ensure generality. For distillation models that directly approximate the score function (e.g., SDXL-Lightning, SDXL-Turbo), the renoising strategy can be extended by incorporating the predicted epsilon from the previous time step. While the performance gains are marginal, we observed some improvement, likely due to adherence to the fundamental refinement principle of reverse diffusion sampling, as studied in [17].

For all quantitative analyses, we fix the number of teacher guidance steps at $k = 1$. Our approach remains simple with minimal hyperparameters, where the teacher guidance scale λ is the primary parameter. Specifically, we set $\lambda = 0.02$ for LCM, LCM-LoRA, and DMD2, and $\lambda = 0.1$ for SDXL-Turbo, SDXL-Lightning, and SDXL-Lightning LoRA. The same λ values are used in Fig. 7 to evaluate the teacher guidance approximation.

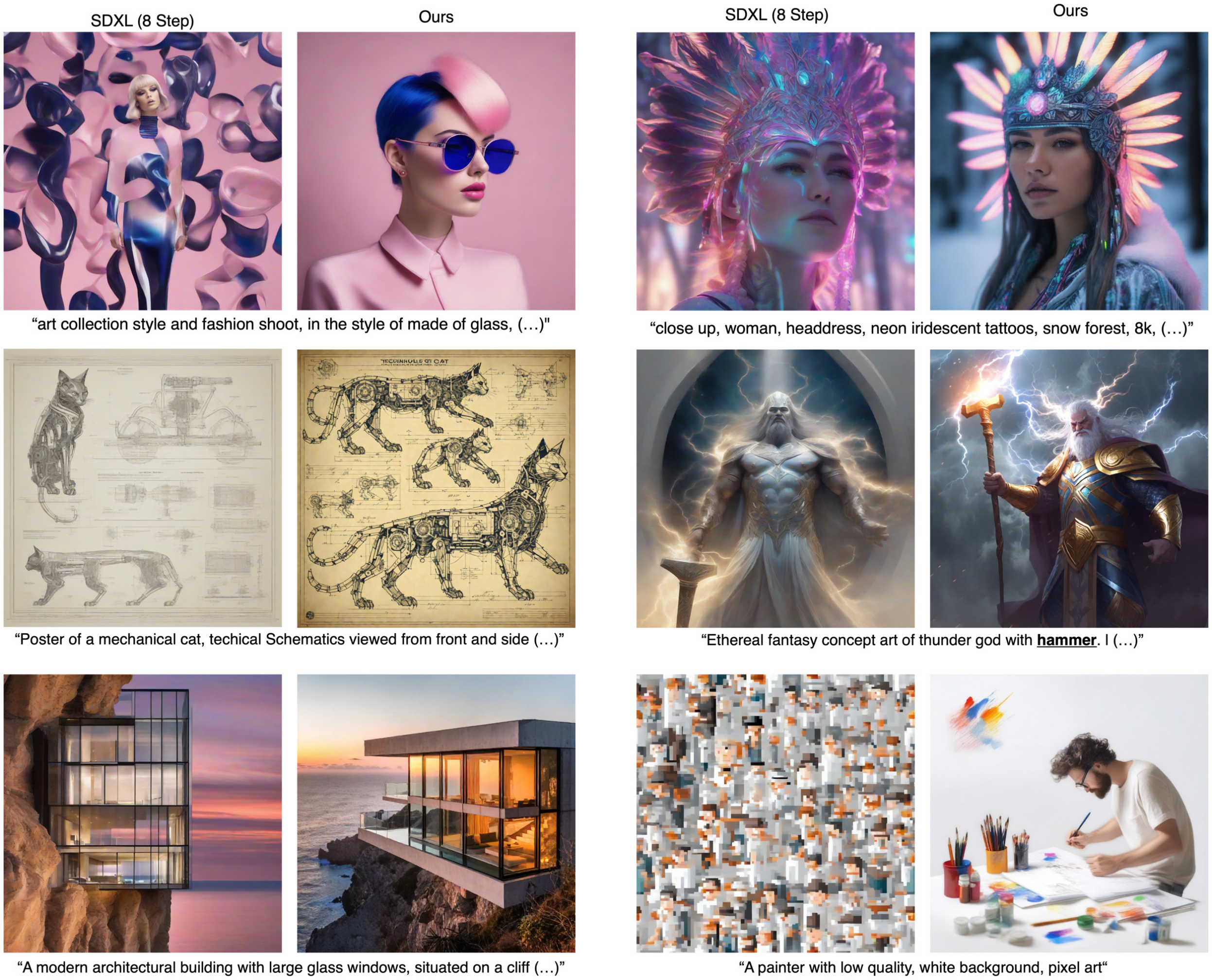
8. Additional results

In Fig. 9 and 10, we demonstrate the effectiveness of the proposed inference-time distillation with various student models. This advances stem from the guidance of teacher model, whereas the teacher model itself does not guarantee high-quality samples with few sampling steps, e.g. 8 steps (Fig. 8). That said, our work fosters a synergistic collaboration between two kinds of diffusion models: fast but sub-optimal student models, and high-quality buy computationally expensive teacher models.

9. Discussions and Limitations

Beyond the image domain, diffusion models have become a cornerstone of high-dimensional visual generative modeling, including applications such as video generation [13] and multi-view synthesis [44]. While computational efficiency is critical for modeling in these high-dimensional spaces, recent studies highlight the challenge of reducing inference steps for video generation. Compared to the image generation, the quality and prompt alignment of generated motion are more dependent on the number of inference steps [35]. Although an increasing number of video diffusion distillation models [20, 49] have emerged, a significant gap remains between student and teacher video diffusion models. Apply-

[2] https://civitai.com/



“art collection style and fashion shoot, in the style of made of glass, (…)"

“close up, woman, headdress, neon iridescent tattoos, snow forest, 8k, (…)”

“Poster of a mechanical cat, techical Schematics viewed from front and side (…)”

“Ethereal fantasy concept art of thunder god with **hammer**. I (…)”

“A modern architectural building with large glass windows, situated on a cliff (…)”

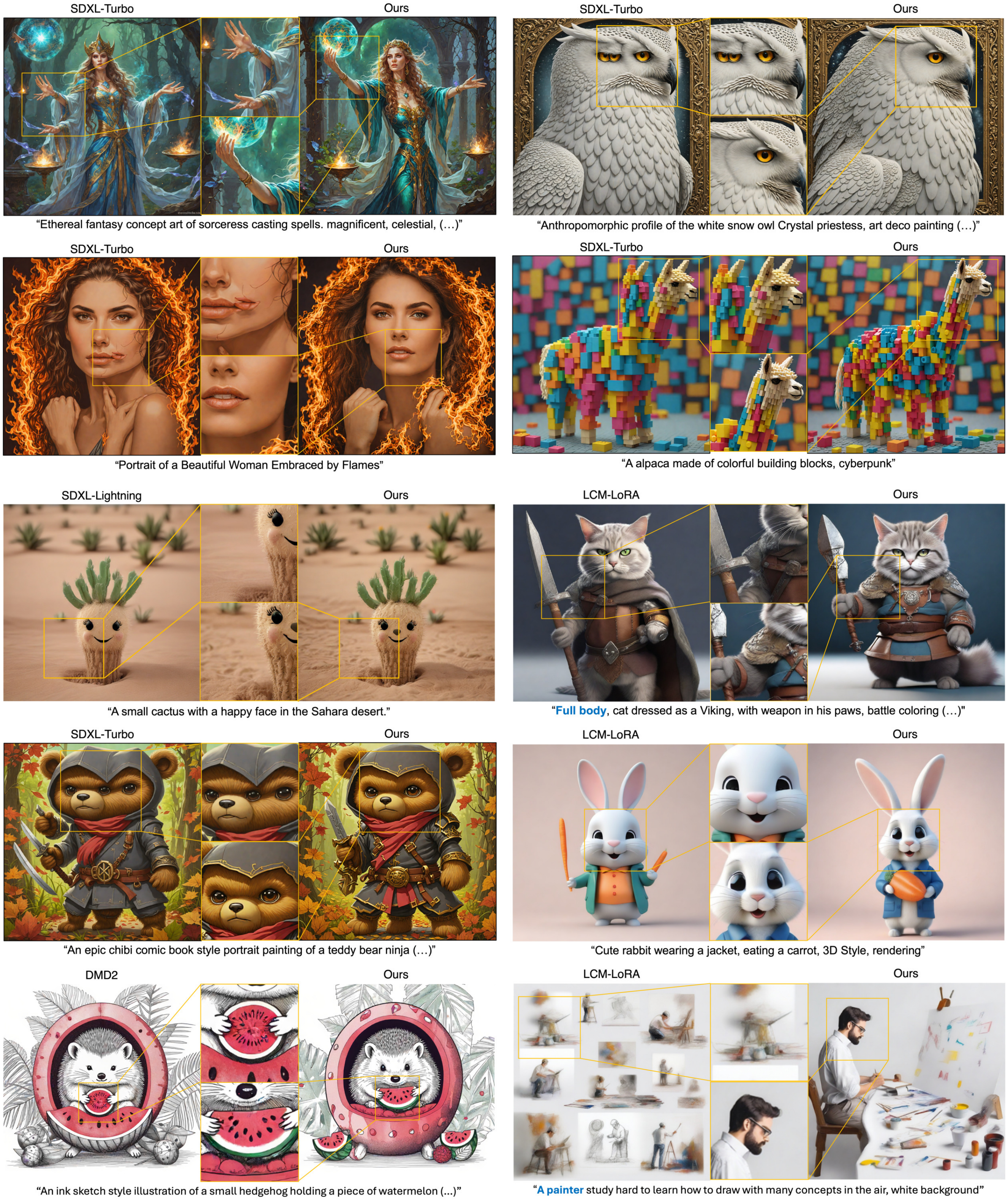
“A painter with low quality, white background, pixel art“

Figure 8. Comparisons between SDXL teacher model (8 steps) and ours. Our results are from the Fig. 3 of the main paper.

ing inference-time diffusion distillation to the video domain offers a promising avenue for improving temporal consistency, addressing issues that text-to-video (T2V) models–often adapted from text-to-image (T2I) models–frequently encounter.

Additionally, flow-based generative models generalize diffusion models and similarly rely on off-the-shelf ODE solvers. Thus, extending inference-time distillation to bridge the gap between flow-based teacher models (e.g., [7]) and student models (e.g., [39]) presents an intriguing direction for future research.

One limitation of the proposed framework is that both student and teacher (latent) diffusion models must operate within a shared latent space to enable interpolation between denoised estimates. To address this, one potential approach could involve mapping latent estimates back to pixel space, refining the student’s pixel estimates, and subsequently re-encoding them. Furthermore, the interplay between students and diverse open-source customized teacher models, which exhibit varying styles and aesthetics, represents another compelling avenue for exploration.



"Ethereal fantasy concept art of sorceress casting spells. magnificent, celestial, (…)"

"Anthropomorphic profile of the white snow owl Crystal priestess, art deco painting (…)"

"Portrait of a Beautiful Woman Embraced by Flames"

"A alpaca made of colorful building blocks, cyberpunk"

"A small cactus with a happy face in the Sahara desert."

"**Full body**, cat dressed as a Viking, with weapon in his paws, battle coloring (…)"

"An epic chibi comic book style portrait painting of a teddy bear ninja (…)"

"Cute rabbit wearing a jacket, eating a carrot, 3D Style, rendering"

"An ink sketch style illustration of a small hedgehog holding a piece of watermelon (...)"

"**A painter** study hard to learn how to draw with many concepts in the air, white background"

Figure 9. Qualitative comparisons against state-of-the-art distillation baselines. Baselines using 4 sampling steps: SDXL-Lightning, DMD2, SDXL-Turbo. Baselines using 8 sampling steps: LCM, LCM-LoRA.

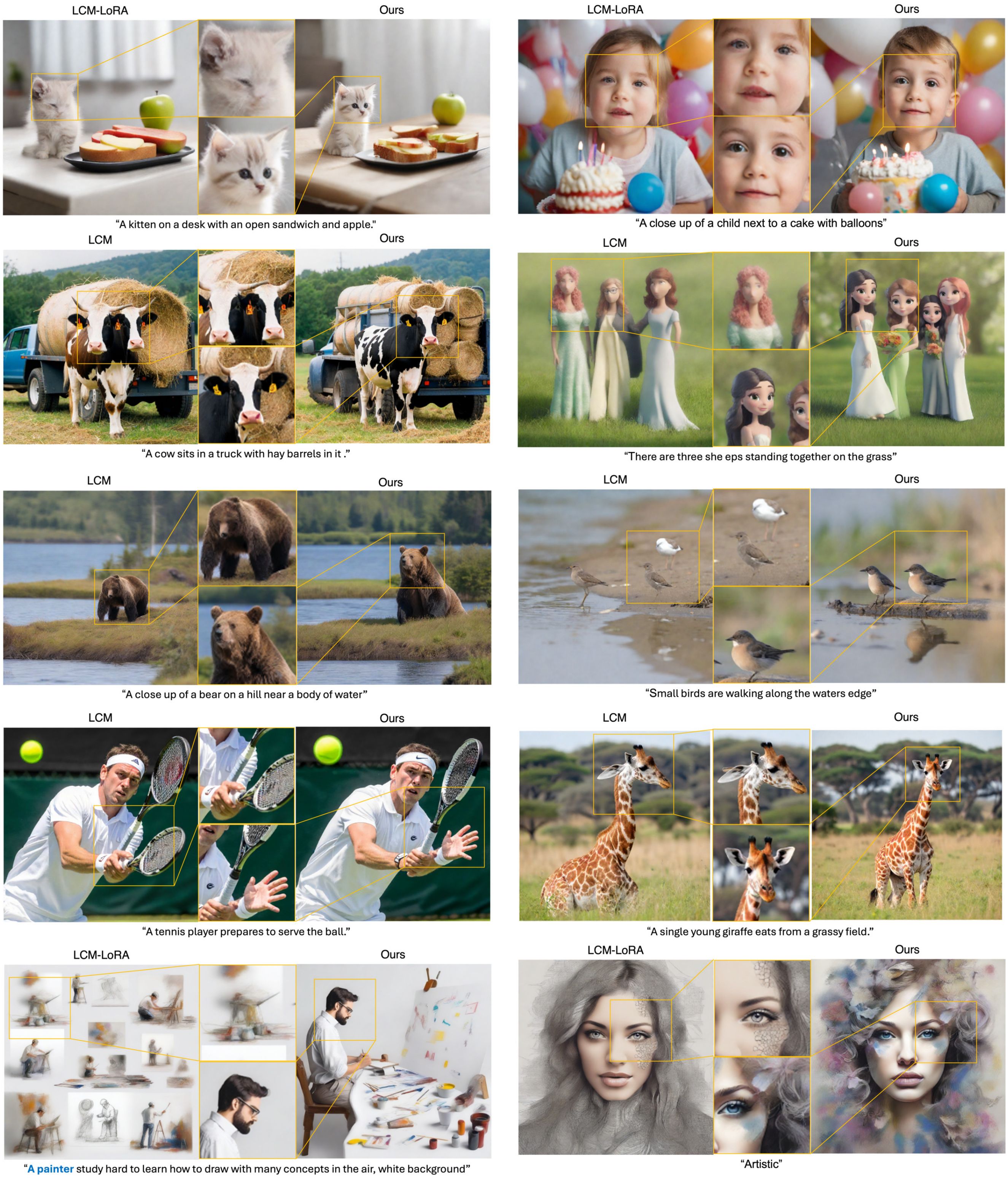

LCM-LoRA
Ours
“A kitten on a desk with an open sandwich and apple."
LCM-LoRA
Ours
“A close up of a child next to a cake with balloons”
LCM
Ours
“A cow sits in a truck with hay barrels in it .”
LCM
Ours
“There are three she eps standing together on the grass”
LCM
Ours
“A close up of a bear on a hill near a body of water”
LCM
Ours
“Small birds are walking along the waters edge”
LCM
Ours
“A tennis player prepares to serve the ball.”
LCM
Ours
“A single young giraffe eats from a grassy field.”
LCM-LoRA
Ours
“A painter study hard to learn how to draw with many concepts in the air, white background”
LCM-LoRA
Ours
“Artistic”

Figure 10. Additional qualitative comparisons against state-of-the-art distillation baselines.