Self-Evaluation Unlocks Any-Step Text-to-Image Generation

Supplementary Material

Abstract

This supplementary material provides additional details and results complementing the main paper. In Sec. S.1, we provide proofs showing that our proposed self-evaluation loss correctly induces the desired optimization gradients. We distinguish two scenarios: deriving the classifier-score gradient and deriving the full reverse KL divergence gradient. For the latter, we include additional implementation details on training. Sec. S.2 contains extended information on our training and inference implementation. In Sec. S.3, we present additional experimental results and further discussions regarding the choice of the second timestep input. Prompts corresponding to image examples shown in the main paper are provided in Sec. S.4. Finally, in Sec. S.5, we discuss limitations of our method and propose directions for future work.

S.1. Derivation of the Self-Evaluation Loss

Setup. We follow the forward noising in Eq. (1) and the model parameterization in Eqs. (6)–(7) (main paper). Throughout, we use the same network head G_θ as in the main text, and *stop-gradient* is denoted by $\mathrm{sg}[\cdot]$. All gradients are taken w.r.t. $\mathbf{x}_s$ that is obtained by re-noising the model prediction $\hat{\mathbf{x}}_0$ as in Sec. 3.2.

Posterior means. By Tweedie's formula applied to Eq. (1), the (data) conditional and unconditional posterior means, and the (model) conditional posterior mean, satisfy

$$\begin{aligned}\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}] &= \frac{1}{\alpha_s}\big(\mathbf{x}_s+\sigma_s^2\nabla_{\mathbf{x}_s}\log q(\mathbf{x}_s|\mathbf{c})\big)\,,\\ \mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s] &= \frac{1}{\alpha_s}\big(\mathbf{x}_s+\sigma_s^2\nabla_{\mathbf{x}_s}\log q(\mathbf{x}_s)\big)\,,\\ \mathbb{E}_{p_\theta}[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}] &= \frac{1}{\alpha_s}\big(\mathbf{x}_s+\sigma_s^2\nabla_{\mathbf{x}_s}\log p_\theta(\mathbf{x}_s|\mathbf{c})\big)\,.\end{aligned} \tag{s.1}$$

Subtracting the first two lines of (s.1) gives

$$\begin{aligned}&\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s]-\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}]\\ &=\frac{\sigma_s^2}{\alpha_s}(\nabla_{\mathbf{x}_s}\log q(\mathbf{x}_s)-\nabla_{\mathbf{x}_s}\log q(\mathbf{x}_s|\mathbf{c}))\,,\end{aligned} \tag{s.2}$$

and subtracting the first and third lines yields

$$\begin{aligned}&\mathbb{E}_{p_\theta}[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}]-\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}]\\ &=\frac{\sigma_s^2}{\alpha_s}(\nabla_{\mathbf{x}_s}\log p_\theta(\mathbf{x}_s|\mathbf{c})-\nabla_{\mathbf{x}_s}\log q(\mathbf{x}_s|\mathbf{c}))\,.\end{aligned} \tag{s.3}$$

S.1.1. Self-evaluation without auxiliary term

We use the self-evaluation pseudo-target from the main paper, Eq. (14),

$$\mathbf{x}_{\text{self}} := \mathrm{sg}\big[\hat{\mathbf{x}}_0-\big(G_\theta(\hat{\mathbf{x}}_s,s,s,\phi)-G_\theta(\hat{\mathbf{x}}_s,s,s,\mathbf{c})\big)\big]\,,$$

and the per-sample squared loss (whose expectation over $(t,s,\mathbf{x}_0,\boldsymbol{\varepsilon})$ gives Eq. (15)):

$$\mathcal{L}_{\text{self}} := \|\hat{\mathbf{x}}_0-\mathbf{x}_{\text{self}}\|_2^2\,. \tag{s.4}$$

Result 1. Under the posterior-mean approximation $G_\theta(\mathbf{x}_s,s,s,\mathbf{c}) \approx \mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}]$ and $G_\theta(\mathbf{x}_s,s,s,\phi) \approx \mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_s]$, the gradient of (s.4) w.r.t. $\hat{\mathbf{x}}_s$ is

$$\begin{aligned}\nabla_{\hat{\mathbf{x}}_s}\mathcal{L}_{\text{self}} &= \Big(\tfrac{\partial\hat{\mathbf{x}}_0}{\partial\hat{\mathbf{x}}_s}\Big)^\top\nabla_{\hat{\mathbf{x}}_0}\mathcal{L}_{\text{self}} = \frac{2}{\alpha_s}(\hat{\mathbf{x}}_0-\mathbf{x}_{\text{self}})\\ &= \frac{2}{\alpha_s}(G_\theta(\hat{\mathbf{x}}_s,s,s,\phi)-G_\theta(\hat{\mathbf{x}}_s,s,s,\mathbf{c}))\\ &\approx \frac{2}{\alpha_s}(\mathbb{E}_q[\mathbf{x}_0|\hat{\mathbf{x}}_s]-\mathbb{E}_q[\mathbf{x}_0|\hat{\mathbf{x}}_s,\mathbf{c}])\\ &= \frac{2\sigma_s^2}{\alpha_s^2}(\nabla_{\hat{\mathbf{x}}_s}\log q(\hat{\mathbf{x}}_s)-\nabla_{\hat{\mathbf{x}}_s}\log q(\hat{\mathbf{x}}_s|\mathbf{c}))\,,\end{aligned} \tag{s.5}$$

where the last equality uses (s.2). Hence gradient descent on Eq. (15) moves $\hat{\mathbf{x}}_s$ in the direction of the *classifier score* $\nabla_{\hat{\mathbf{x}}_s}\log q(\mathbf{c}|\hat{\mathbf{x}}_s)$. Note that the $\lambda_{s,t}$ has already absorbed these coefficients.

S.1.2. Self-evaluation with auxiliary term

We optionally add a branch prompted by $\mathbf{c}_{\text{fake}}$ to estimate the model posterior mean $G_\theta(\mathbf{x}_s,s,s,\mathbf{c}_{\text{fake}}) \approx \mathbb{E}_{p_\theta}[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}]$. Define

$$\begin{aligned}\Delta_\theta(\mathbf{x}_s,\mathbf{c}) := k(&G_\theta(\mathbf{x}_s,s,s,\phi)-G_\theta(\mathbf{x}_s,s,s,\mathbf{c}))\\ &+(1-k)\,(G_\theta(\mathbf{x}_s,s,s,\mathbf{c}_{\text{fake}})-G_\theta(\mathbf{x}_s,s,s,\mathbf{c}))\,,\end{aligned} \tag{s.6}$$

and the target $\mathbf{x}_{\text{self}} := \mathrm{sg}[\hat{\mathbf{x}}_0-\Delta_\theta(\hat{\mathbf{x}}_s,\mathbf{c})]$, and the per-sample squared loss $\mathcal{L}_{\text{self}} := \|\hat{\mathbf{x}}_0-\mathbf{x}_{\text{self}}\|_2^2$.

Result 2. Proceeding as in (s.5),

$$\begin{aligned}\nabla_{\hat{\mathbf{x}}_s}\mathcal{L}_{\text{self}} &= \frac{2}{\alpha_s}\Delta_\theta(\hat{\mathbf{x}}_s,\mathbf{c})\\ &\approx \frac{2}{\alpha_s}[k(\mathbb{E}_q[\mathbf{x}_0|\hat{\mathbf{x}}_s]-\mathbb{E}_q[\mathbf{x}_0|\hat{\mathbf{x}}_s,\mathbf{c}])\\ &\quad+(1-k)\left(\mathbb{E}_{p_\theta}[\mathbf{x}_0|\hat{\mathbf{x}}_s,\mathbf{c}]-\mathbb{E}_q[\mathbf{x}_0|\hat{\mathbf{x}}_s,\mathbf{c}]\right)]\\ &= \frac{2\sigma_s^2}{\alpha_s^2}[k(\nabla_{\hat{\mathbf{x}}_s}\log q(\hat{\mathbf{x}}_s|\phi)-\nabla_{\hat{\mathbf{x}}_s}\log q(\hat{\mathbf{x}}_s|\mathbf{c}))\\ &\quad+(1-k)\left(\nabla_{\hat{\mathbf{x}}_s}\log p_\theta(\hat{\mathbf{x}}_s|\mathbf{c})-\nabla_{\hat{\mathbf{x}}_s}\log q(\hat{\mathbf{x}}_s|\mathbf{c})\right)],\end{aligned} \tag{s.7}$$

where we used (s.2) and (s.3). Equation (s.7) is proportional to the full ideal vector field in Eq. (13) once we set $k = (w-1)/w$. In practice, we set $k = 0.9$.

Training. To realize $G_\theta(\mathbf{x}_s, s, s, \mathbf{c}_{\text{fake}}) \approx \mathbb{E}_{p_\theta}[\mathbf{x}_0|\mathbf{x}_s,\mathbf{c}]$, we use model samples and reuse the same conditional FM loss as Eq. (7): draw $\hat{\mathbf{x}}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{c})$ and $\hat{\mathbf{x}}_s \sim p_\theta(\mathbf{x}_s|\mathbf{x}_0,\mathbf{c})$, and $\mathbf{c}_{\text{fake}}$ is constructed by concatenating the phrase 'fake image' with the original prompt, and then minimize

$$\mathcal{L}_{\text{fake}} = \mathbb{E}\left[\|G_\theta(\text{sg}[\hat{\mathbf{x}}_s], s, s, \mathbf{c}_{\text{fake}}) - \text{sg}[\hat{\mathbf{x}}_0]\|_2^2\right]. \tag{s.8}$$

In practice we follow the training schedule in Sec. 4 of the main paper: use only the classifier term early, and enable the auxiliary term later to refine artifacts, while keeping the overall objective identical to Eqs. (16)–(21).

S.2. Implementation Details

We adopt a latent transformer architecture similar to FLUX [1, 2] for our experiments, with minor modifications to accommodate our new s-input. Specifically, the design of the modules handling s mirrors those handling t.

We employ a 2B-parameter model trained on mixed-resolution and varying aspect-ratio text-to-image datasets. Initially, the model is trained at an approximate resolution of 256^2 pixels for 500k iterations with a batch size of 1024. Subsequently, we introduce higher-resolution data of 512^2 pixels, maintaining a balanced batch proportion (1:1) between the lower-resolution and higher-resolution data, with a total batch size of 768, continuing training until reaching 710k iterations. At iteration 550k, we additionally introduce training with the auxiliary term. We use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$, a learning rate warmup for 1000 iterations, and linearly decay the learning rate from 3×10^{-4} to 1×10^{-5}. For model evaluation, we maintain an exponential moving average (EMA) with a decay rate of 0.9999. Additionally, during training in the self-evaluation forward pass, the conditional branch utilizes the EMA model, while the unconditional branch employs the non-EMA model.

Architecture. We adopt a FLUX-style latent transformer and keep the notation consistent with the main paper: the denoiser's raw prediction is $V_\theta(\cdot)$ and the sample head is $G_\theta(\mathbf{x}_t, t, s, \mathbf{c}) = \mathbf{x}_t - t\,V_\theta(\mathbf{x}_t, t, s, \mathbf{c})$ (cf. Eq. (6)–(7) in the main paper). Our implementation consists of four modules: (a) a VAE, (b) a patchifier, (c) frozen text encoders, and (d) a dual-time denoiser.

(a-b) VAE and patch tokens. We use the FLUX.1-dev auto-encoder with z-channels $= 16$, and compression factors $[1, 8, 8]$ for [frames, H, W]. Images are tokenized by a patchifier with patch size $[1, 2, 2]$. Thus, each image produces a sequence of $L_{\text{img}} = (H/16)\times(W/16)$ tokens, each of dimension $d_{\text{img}} = 16\times2\times2 = 64$.

(c) Text and global conditioning. We use a frozen T5-XXL encoder to obtain token embeddings of dimension $d_{\text{txt}} = 4096$. Additionally, we compute a global pooled CLIP embedding (ViT-L/14) of dimension $d_{\text{vec}} = 768$. Both encoders are kept frozen during training, and their outputs are linearly projected to $\mathbb{R}^{d_{\text{model}}}$ before entering the denoiser.

(d) Denoiser. Our 2B model has a model width $d_{\text{model}} = 2048$, head size $d_{\text{head}} = 128$ (thus 16 heads), and a total of 8 *Double-Stream* blocks followed by 16 *Single-Stream* blocks. Positional encoding uses multi-axis RoPE over (t, y, x) with axis dimensions $[16, 56, 56]$, whose sum matches d_{head} and whose three axes correspond to time and the two spatial directions. Inputs are linearly projected to d_{model}: text via $\mathbb{R}^{4096}\to\mathbb{R}^{2048}$, image tokens via $\mathbb{R}^{64}\to\mathbb{R}^{2048}$, and the global CLIP vector via a two-layer MLP $\mathbb{R}^{768}\to\mathbb{R}^{2048}$. For ablations, we also train a smaller 0.5B variant with $d_{\text{model}} = 1024$, $d_{\text{head}} = 64$, and RoPE axis dimensions $[8, 28, 28]$, while keeping all other components identical.

Particularly, the denoiser has two time inputs: the primary time t and an auxiliary time s used by the self-evaluation mechanism (Sec. 3.2). In practice, we encode t and the gap $t - s$ with sinusoidal features followed by small MLPs:

$$e_t = \text{MLP}_t(\text{Sinusoid}(t)), e_s = \text{MLP}_s(\text{Sinusoid}(t-s)), \tag{s.9}$$

and form a combined time embedding

$$\tilde{e}_t = e_t + e_s. \tag{s.10}$$

This combined embedding $\tilde{e}_t$ simply replaces the original single-time embedding in the backbone: every module that previously consumed e_t now receives $\tilde{e}_t$. Consequently, the only architectural change relative to FLUX is the additional auxiliary term e_s added on top of e_t, while all downstream conditioning and modulation remain unchanged.

Timestep Scheduler. We first sample the primary time t from a logit-normal distribution defined on $(0, 1)$:

$$t_{\text{raw}} = \sigma(z), \quad z \sim \mathcal{N}(0, 1), \tag{s.11}$$

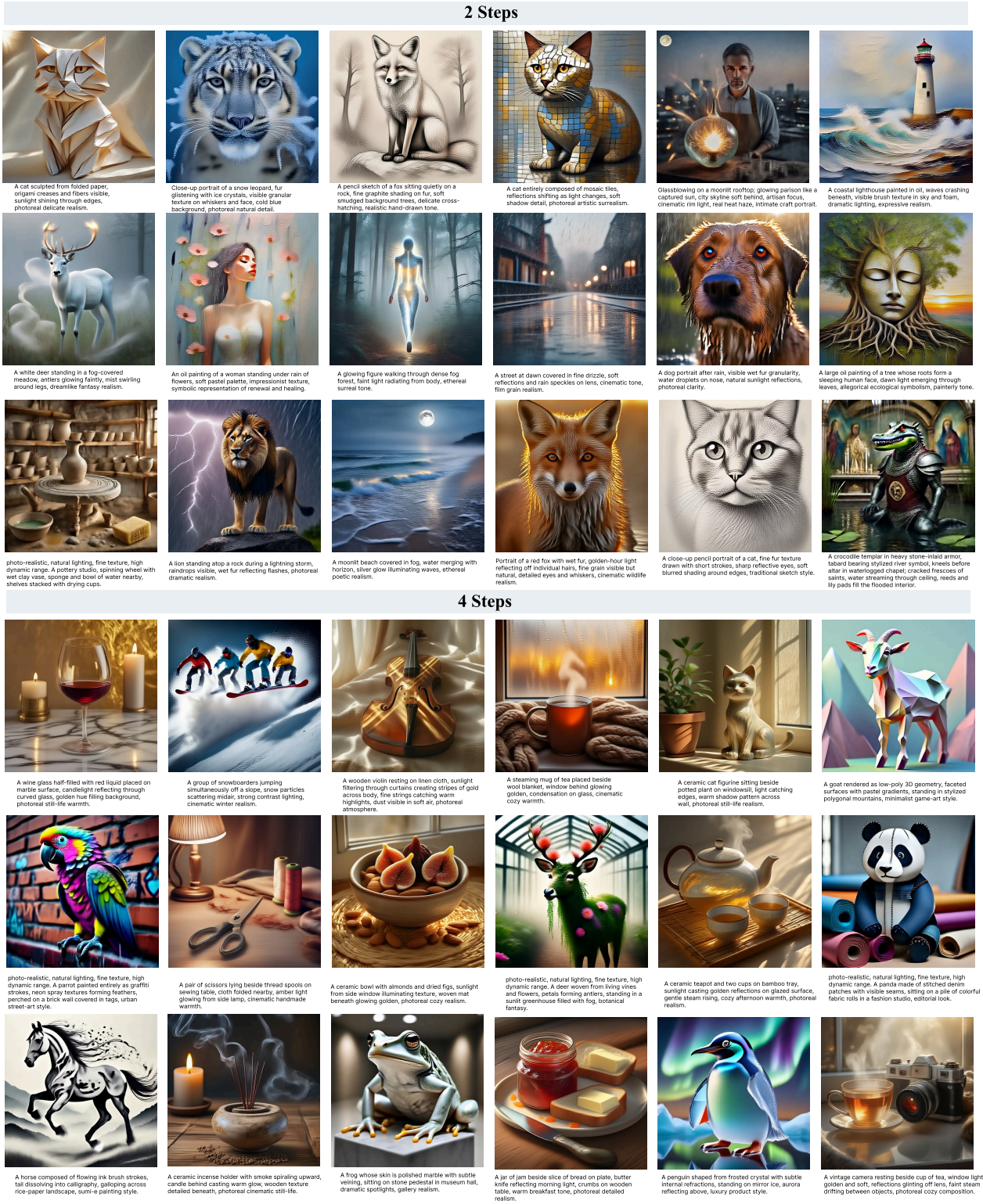

Figure S.1. **More results with 2 and 4 steps.** We showcase diverse text-to-image results from our model at 2 and 4 inference step counts, demonstrating coherent semantics, strong text alignment.

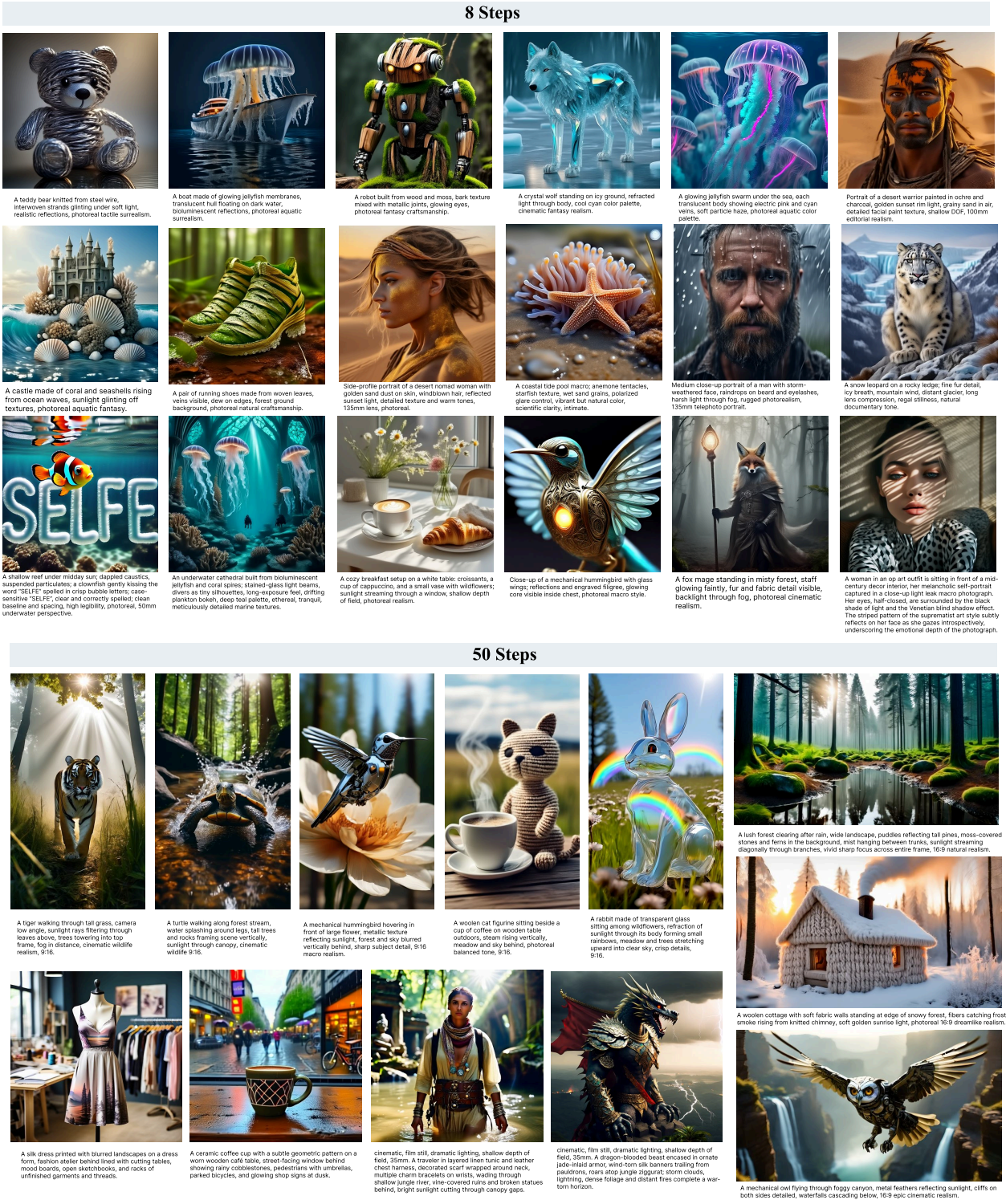

Figure S.2. **More results with 8 and 50 steps.** We showcase diverse text-to-image results from our model at 8 and 50 inference step counts, demonstrating coherent semantics, strong text alignment.

where $\sigma(\cdot)$ denotes the sigmoid function. This raw time is further adjusted by a length-dependent warping function. Specifically, given the latent patch length L, we define a linear shift $\mu(L)$ interpolating between 0.5 at length 512 and

1.15 at length 4096, then compute the warped primary time as:

$$t = \frac{e^{\mu(L)}}{e^{\mu(L)} + (1/t_{\text{raw}} - 1)}. \tag{s.12}$$

For the secondary time s, we set $s = t$ with probability $p = 0.5$. For the remaining half of the cases, we sample s uniformly from the interval:

$$s \sim \mathcal{U}((1-\tau)\, t,\, t), \tag{s.13}$$

where τ is a linear annealing weight, transitioning from 0 to 1 over the first $300,000$ training iterations. As a result, the effective lower bound $(1-\tau)t$ decreases gradually from approximately t towards 0 during training. For the weighting function $w_{s,t}$ in Eq. (20), we set it to $1/t^2$.

Inference. For inference, we employ an initially linear timestep scheduler with a length-dependent warping function, same with Eq. (s.12). We use a DDIM-style update with an η-controlled noise level, following Song et al. [3]; setting $\eta = 0$ recovers deterministic DDIM, while $\eta = 1$ corresponds to the original DDPM ancestral sampling. In our case, we use $\eta = 1$.

S.3. Additional Experimental Results

S.3.1. Alternative s-Scheduler

We investigate alternative strategies for selecting the secondary timestep s_k during inference, given a transition from t_k to t_{k+1}. During training, the selection of s_k affects two aspects simultaneously: it determines the noise level for the smoothed data distribution used in the reverse KL divergence, and it specifies the self-evaluation weighting factor $\lambda_{s,t}$. These dual roles suggest alternative choices for s_k might yield intermediate and potentially improved behaviors. An intriguing direction for future work would be decoupling the dependence between s_k and the weighting factor $\lambda_{s,t}$, making $\lambda_{s,t}$ independently tunable.

We illustrate our empirical observations in Fig. S.3, highlighting two notable special cases:

1. When $s_k = t_k$, the model utilizes only the flow matching loss. Consequently, its behavior closely resembles standard Flow Matching, performing poorly at very low inference steps but improving significantly with more steps.
2. When $s_k = t_{k+1}$, the model excels in few-step generation. However, as the number of inference steps increases (e.g., at 50 steps), we occasionally observe it underperforms compared to $s_k = t_k$ (see the last two examples in Fig. S.3).

Additionally, we explore a special inference setting—*one-step* generation without classifier-free guidance. As shown in Fig. S.4, we interpolate between $s_k = t_k$ (represented as $s = 1$) and $s_k = t_{k+1}$ (represented as $s = 0$). Both extreme cases fail to yield meaningful images, whereas the midpoint choice $s = 0.5$ achieves a favorable balance between texture detail and overall image coherence.

Figure S.3. **Visualization of two special cases for choosing the secondary timestep s_k during inference.** Top rows: $s_k = t_k$, bottom rows: $s_k = t_{k+1}$.

S.3.2. More Results

We present more results at different inference budgets in Fig. S.1 and Fig. S.2.

S.4. Prompts of Results

We provide the text prompts used for the qualitative results shown in the main paper.

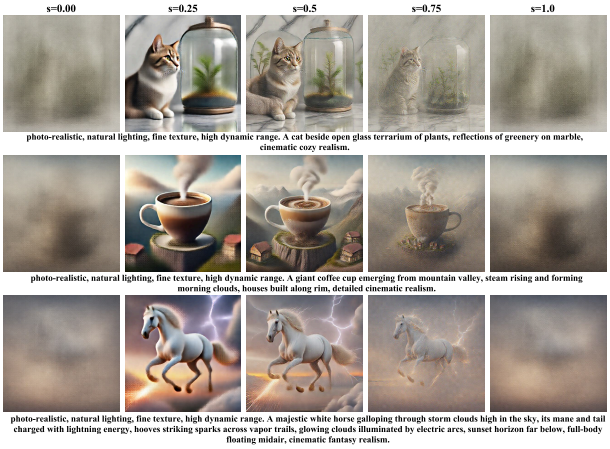

Figure S.4. **One-step generation without classifier-free guidance**. We show results of when selecting different s.

S.4.1. Prompts of Figure 1.

2-step:

- The word "Self-E" appearing faintly through condensation on a train window, blurred landscape passing behind, city lights refracting, cinematic melancholic tone.
- Portrait of a wolf under snowfall, frost collecting on its muzzle and fur, visible texture and natural grain, calm expression, photoreal cold-environment realism.
- An oil painting of a woman with her hair turning into waves, seascape blending with portrait, tactile brushwork, painterly surreal tone.

4-step:

- A volcano erupting with petals instead of lava, clouds of color drifting across the sky, surreal cinematic beauty.
- A cat composed of smoke sitting on a rooftop, its form dissolving into the night air, glowing eyes reflecting city lights, detailed cinematic surrealism.
- A plate of pastries beside a teacup, sunlight highlighting golden crusts, powdered sugar shimmering under a warm glow, photoreal comforting realism.

8-step:

- Portrait of a jungle guardian with vine tattoos and green-gold war paint, wet skin glistening under filtered sunlight, 85mm, macro detail on skin texture, cinematic naturalism.
- A bison standing in a foggy grassland at dawn, dew on tall grass, sun barely visible through haze, fur glistening with moisture, cinematic atmospheric realism.
- A cozy cottage built entirely from red and white yarn, knitted walls and woven roof shingles, soft texture visible in each thread, golden sunlight casting gentle shadows, photoreal tactile realism.

50-step:

- A human face emerging from cracked porcelain, half side smooth and half crumbled revealing crystalline interior, emotional surreal realism.
- A queen in jeweled crown standing under golden archway, sunlight refracting through gems, detailed embroidery on gown, distant cityscape visible behind, regal photoreal tone, 9:16.
- A rabbit made of transparent glass jumping across a shallow creek, sunlight refracting rainbow light through its body, ripples and stones visible beneath, forest on both sides, 16:9 photoreal wide scene.
- A close-up underwater portrait of a woman leaning forward on a large rectangular glowing sign that reads "Self-E," the sign filling the lower part of the frame like a real physical board. Neon hues of cyan, pink, and gold from the illuminated surface ripple through the clear turquoise water, casting colorful reflections across her face. She smiles brightly, blue eyes open with confidence, freckles and natural skin texture visible under shifting light. Transparent fish swim nearby among coral branches, tiny bubbles rising through the calm cinematic 9:16 scene.
- A valley full of blooming lupines and daisies, 16:9 panoramic view, rolling hills leading toward mountain horizon, warm afternoon light highlighting color contrast, photoreal cinematic realism.

S.4.2. Prompts of Figure 4.

- A colorful chalkboard artwork spelling "SELFE" in bright pastel colors—blue, pink, yellow, and green—each letter outlined softly, chalk dust particles floating through air, faint eraser marks around, warm nostalgic classroom atmosphere.
- A small home bar setup with wine bottles, glass of whiskey half full, sliced lemon on napkin, reflections on wooden counter, photoreal cinematic tone.
- A cat sleeping on cloud drifting above mountain range, soft pink sunrise illuminating fur, photoreal dreamlike realism.
- A royal guard in ornate jade armor, sword reflecting sunlight, palace gardens behind full of flowers and fountains, silk banners waving in soft breeze, cinematic elegant realism.

S.4.3. Prompts of Figure 5.

- A high-altitude thunderhead above a wheat plain; sculpted cumulonimbus, sunlit anvil, tiny barn for scale, global contrast, 24mm vastness, dramatic meteorological realism.
- A house constructed from luminous jelly bricks glowing at night, detailed transparency and refraction, cinematic realism.

S.5. Limitations and Future Work

While our method significantly surpasses existing from-scratch training methods in few-step generation, it still has some limitations. Notably, our current approach, although effective in significantly reducing the number of inference steps, cannot fully compete with the quality obtained by

50-step inference when employing extremely few steps (e.g., 1–2 steps). In these cases, the generated images may lack sufficiently sharp details.

Additionally, given that our proposed paradigm fundamentally differs from existing consistency-based methods, it remains at an early stage of exploration. Several critical design choices, such as loss weighting schemes and inference strategies, have not yet been thoroughly optimized. We believe further systematic exploration of these aspects could lead to considerable improvements.

Nonetheless, we emphasize that our method introduces a genuinely novel training paradigm, distinct from the consistency-training family. Empirically, we observe that our method inherently produces robust structure and semantic coherence, exhibiting a clear trend of generating coherent structures first, followed by iterative refinement of details.

Looking forward, we identify several promising avenues for future work:

1. Improving training strategies and inference-time scheduling to further enhance generation quality.
2. Investigating the efficacy of our approach for downstream task fine-tuning.
3. Exploring scalability and potential adaptations of the proposed paradigm to video generative models.
4. Extending our method to unconditional generative settings, as the current approach relies on conditional guidance to derive the classifier scores.

References

[1] Black-Forest-Labs. Flux.1 [dev], 2024. 12B parameter rectified flow transformer, text-to-image. 2

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5