Imagine Flash: Accelerating Emu Diffusion Models with Backward Distillation

Jonas Kohler*, Albert Pumarola*, Edgar Schönfeld*, Artsiom Sanakoyeu*,
Roshan Sumbaly, Peter Vajda, and Ali Thabet

GenAI, Meta
{jonaskohler,apumarola,edgarschoenfeld,asanakoy}@meta.com

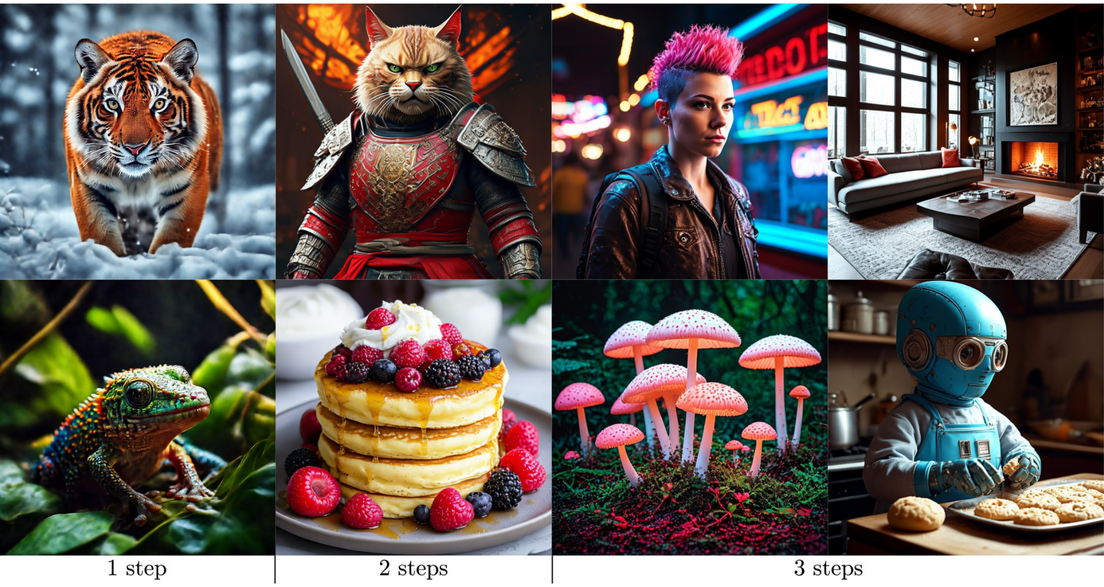


Fig. 1: Imagine Flash generation in 1, 2, and 3 steps. Imagine Flash uses backward distillation to accelerate inference of a baseline diffusion model (Emu). Our distillation framework allows generating high quality images with as few as 1-3 steps.

Abstract. Diffusion models are a powerful generative framework, but come with expensive inference. Existing acceleration methods often compromise image quality or fail under complex conditioning when operating in an extremely low-step regime. In this work, we propose a novel distillation framework tailored to enable high-fidelity, diverse sample generation using just one to three steps. Our approach comprises three key components: (i) Backward Distillation, which mitigates training-inference discrepancies by calibrating the student on its own backward trajectory; (ii) Shifted Reconstruction Loss that dynamically adapts knowledge transfer based on the current time step; and (iii) Noise Correction, an inference-time technique that enhances sample quality by addressing singularities in noise prediction. Through extensive experiments, we demonstrate that our method outperforms existing competitors in quantitative metrics and

[*] Equal Contribution

human evaluations. Remarkably, it achieves performance comparable to the teacher model using only three denoising steps, enabling efficient high-quality generation.

Keywords: Generative AI · Efficient Diffusion · Image Synthesis

1 Introduction

Generative modelling has witnessed a paradigm shift with the advent of Denoising Diffusion Models (DMs) [8, 37]. These models have set new benchmarks across various domains [7, 13, 28], offering an unprecedented combination of realism and diversity, while ensuring stable training. However, the sequential nature of the denoising process presents a significant challenge. Sampling from DMs is a time-consuming and costly process, with the time required largely dependent on two factors: (i) the latency of the per-step neural network evaluation, and (ii) the total number of denoising steps.

Considerable research efforts have been devoted to accelerating the sampling process. For text to image synthesis, the proposed methods span a wide range of techniques, including higher-order solvers [21], modified diffusion formulations for curvature reduction [20], as well as guidance- [24], step- [29] and consistency distillation [36]. These methods have brought impressive improvements, achieving very high quality in the close to 10 step regime. More recently, hybrid methods that leverage both distillation and adversarial losses [17, 31, 39] have pushed the boundary to under five steps. While these methods achieve impressive quality on simple prompts and uncomplicated styles like animation, they suffer from degraded sample quality for photorealistic images, especially for complex text conditioning.

A common theme among the aforementioned methods is the attempt to align the few-step student model with the complex teacher paths, despite endowing the student model with significantly lower capacity (i.e., steps). Recognizing this as a limitation, we invert the process by proposing a novel distillation framework that is designed for the teacher to improve the student along it's own diffusion paths. In summary, our contribution is threefold:

- First, our approach introduces Backward Distillation, a distillation process designed to calibrate the student model on its own upstream backward trajectory, thereby reducing the gap between the training and inference distributions and ensuring zero data leakage during training across all time steps.
- Secondly, we propose a Shifted Reconstruction Loss that dynamically adapts the knowledge transfer from the teacher model. Specifically, the loss is designed to distill global, structural information from the teacher at high time steps, while focusing on rendering fine-grained details and high-frequency components at lower time steps. This adaptive approach enables the student

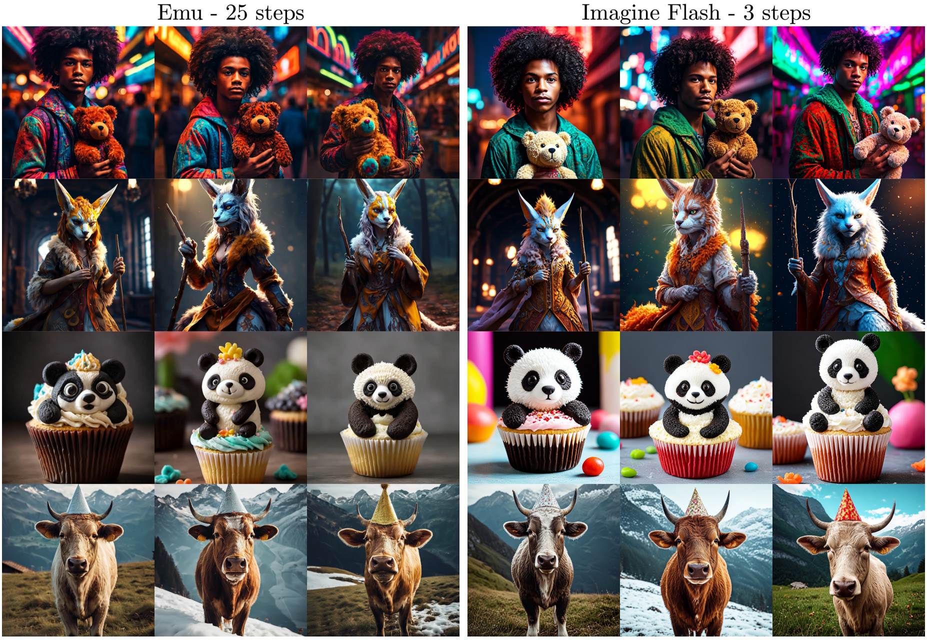


Fig. 2: Imagine Flash vs Emu. Our method reduces the inference time of the baseline by significant amount, and still generates high quality and complex images.

to effectively emulate the teacher's generation process at different stages of the diffusion trajectory.

- Finally, we propose Noise Correction, an inference-time modification that enhances sample quality by addressing singularities present in noise prediction models during the initial sampling step. This training-free technique mitigates degradation of contrast and color intensity that usually arise when operating with an extremely low number of denoising steps.

By synergistically combining these three novel components, we apply our distillation framework to a baseline diffusion model, Emu [4], resulting in Imagine Flash which achieves high-quality generation in the extremely low-step regime without compromising sample quality or conditioning fidelity (Fig. 2). Through extensive experiments and human evaluations, we demonstrate the effectiveness of our approach in achieving favorable trade-offs between sampling efficiency and generation quality across a range of tasks and modalities.

2 Related Work

Diffusion models [8, 25, 34], in contrast to previous generative models (*e.g.*, GANs), approach density estimation and data sampling in an iterative way, by gradually reversing a noising process. This iterative nature translates to multiple queries of a neural network backbone, leading to high inference costs. As a result, a large body of works has focused on producing faster and more efficient ways of sampling from diffusion models. However, enhancing inference speed without sacrificing image quality and text faithfulness continues to present a considerable challenge.

Solvers and curvature rectification: Early approaches focus on developing better solvers to the underlining dynamics of the diffusion process. Along this line, several works propose exponential integrators [22, 41], higher order solvers [11, 42] and model-specific bespoke solvers [33, 44]. Other studies investigate reformulations of the diffusion process with the aim of minimizing curvature in both the forward (noising) [1, 19] and backward (de-noising) trajectories [11, 14, 20, 27]. In a nutshell, these approaches aim to linearize the inference path, allowing for larger step sizes, and therefore fewer steps at inference time. Despite the substantial step reduction of these methods, there is a limit on how large the inference step can be, without compromising image quality.

Reducing model size: A series of orthogonal works looks at reducing the per-step cost. In this vein, several works focus on employing smaller backbone architectures [15, 26, 40], and even mobile friendly networks [43]. Reducing per-step cost is also addressed by minimizing the cost of conditional generation at each iteration or caching intermediate activations in the network backbone [38]. To that extent, [24] propose guidance distillation, while [3] present the training-free alternative of truncating guidance. Reducing per-step latency leads to significant gains in inference speed. However, to truly scale inference for real-time applications, these advancements must be coupled with further reductions of the number of steps to a small single-digit number.

Reducing sampling steps: One way to further reduce inference latency is step distillation [24,29]. In these work, the authors propose a progressive approach to distill two or more steps into a single one. The effect is further enhanced when using consistency constraints [23, 36]. While these approaches achieve significant step reductions, substantial quality degradation is evident at low step regimes. To compensate for quality loss, a further line of work proposes additional training enhancements during distillation. Namely, ADD [31], Lightning [17] and UFOGEN [39] add an adversarial loss to increase sample quality.

While the above distillation methods undoubtedly produce impressive results with just a single step generation, these improvements are still not adequate for many practical applications, such as generating photorealistic images or adhering to complex prompts. A more reasonable approach is to control the trade-off between quality and speed. Practically, this translates to methods that allow small step increases (2 to 4 steps), with a major gain in quality. We adopt this approach in our method. To achieve better quality, we propose to distill along the student's *backward* path instead of the forward path. Put differently, rather than having the student mimic the teacher, we use the teacher to improve the student based on its current state of knowledge. We find that this approach leads to competitive results with single inference steps, and substantially improves quality and fidelity with just a slight increase to as few as 3 steps.

3 Background on Diffusion Models

Diffusion models consist of two interconnected processes: forward and backward. The forward diffusion process gradually corrupts the data by interpolating between a sampled data point $\mathbf{x}_0$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. That is

$$\mathbf{x}_t = q(\mathbf{x}_0, \boldsymbol{\epsilon}, t) = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \forall t \in [0, T], \tag{1}$$

where α_t and σ_t define the signal-to-noise ratio (SNR) of the stochastic interpolant $\mathbf{x}_t$. In the following, we opt for coefficients (α_t, σ_t) that result in a variance-preserving process (see e.g. [11]). When viewed in the continuous time limit, the forward process in Eq. 1 can be expressed as the stochastic differential equation (SDE) $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w_t}$, where $f(\mathbf{x}, t) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued drift coefficient, $g(t) : \mathbb{R} \to \mathbb{R}$ is the diffusion coefficient, and $\mathbf{w_t}$ denotes the Brownian motion at time t.

Inversely, the backward diffusion process is intended to undo the noising process and generate samples. According to Anderson's theorem [2], the forward SDE introduced earlier satisfies a reverse-time diffusion equation, which can be reformulated using the Fokker-Planck equations [37] to have a deterministic counterpart with equivalent marginal probability densities, known as the *probability flow ODE*:

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt. \tag{2}$$

As demonstrated in [10,37], this marginal transport map can be learned through maximum likelihood estimation of the perturbation kernel of diffused data samples $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{x}_0)$ in a simulation-free manner. This allows us to estimate $\hat{\boldsymbol{\epsilon}}(\mathbf{x}_t, t)/\sigma_t \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{x}_0)$, usually parameterized by a time-conditioned neural network. Given these estimates, we can sample using an iterative numerical solver f [35]:

$$\mathbf{x}_0 \approx f \circ f \circ \cdots \circ f(\mathbf{x}_T). \tag{3}$$

without loss of generality, in the paper we use the update rule given by first-order solvers like DDIM [35], i.e.:

$$\mathbf{x}_{t-1} = f(\mathbf{x}_t) = \alpha_{t-1} \hat{\mathbf{x}}_0(\mathbf{x}_t, \hat{\boldsymbol{\epsilon}}, t) + \sigma_{t-1} \hat{\boldsymbol{\epsilon}}(\mathbf{x}_t, t), \tag{4}$$

where the sample data estimate $\hat{\mathbf{x}}_0$ at time-step t is computed as:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t, \hat{\boldsymbol{\epsilon}}, t) = \frac{\mathbf{x}_t - \sigma_t \hat{\boldsymbol{\epsilon}}(\mathbf{x}_t, t)}{\alpha_t}. \tag{5}$$

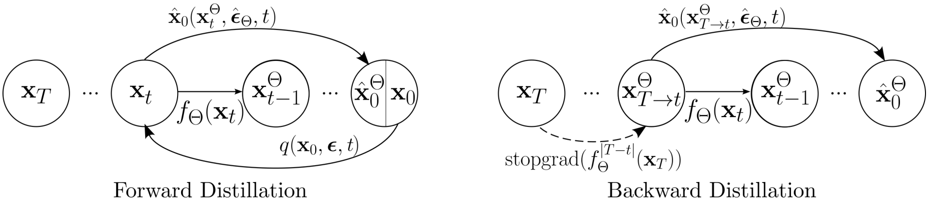

$\hat{\mathbf{x}}_0(\mathbf{x}_t^\Theta, \hat{\boldsymbol{\epsilon}}_\Theta, t)$
$\mathbf{x}_T$
$\cdots$
$\mathbf{x}_t$
$f_\Theta(\mathbf{x}_t)$
$\mathbf{x}_{t-1}^\Theta$
$\cdots$
$\hat{\mathbf{x}}_0^\Theta$
$\mathbf{x}_0$
$q(\mathbf{x}_0, \boldsymbol{\epsilon}, t)$
Forward Distillation
$\hat{\mathbf{x}}_0(\mathbf{x}_{T\to t}^\Theta, \hat{\boldsymbol{\epsilon}}_\Theta, t)$
$\mathbf{x}_T$
$\cdots$
$\mathbf{x}_{T\to t}^\Theta$
$f_\Theta(\mathbf{x}_t)$
$\mathbf{x}_{t-1}^\Theta$
$\cdots$
$\hat{\mathbf{x}}_0^\Theta$
$\text{stopgrad}(f_\Theta^{|T-t|}(\mathbf{x}_T))$
Backward Distillation

Fig. 3: Backward Distillation ensures signal consistency between training and inference during distillation. In standard forward distillation we run training steps starting from the forward-noised latent code $\mathbf{x}_t$. In backward distillation, we instead use the student model's backward iterations to get latent $\mathbf{x}_{T\to t}^{\Theta}$ and use it as the starting code for training steps, where we compute gradients. Backward distillation eliminates information leakage for all ts, and prevents the model from relying on GT signals.

4 Method

We present Imagine Flash, a novel distillation technique designed for fast text-to-image generation that builds upon – but is not exclusive to – Emu [4]. Unlike the original Emu model that requires at least 50 neural function evaluations (NFEs) to produce high-quality samples, Imagine Flash achieves comparable results with just a few NFEs. Our proposed distillation method comprises three novel key components: (i) *Backward Distillation*, a distillation process that ensures zero data leakage during training for all time points t (see Sec. 4.1). (ii) *Shifted Reconstruction Loss (SRL)*, an adaptive loss function designed to maximize knowledge transfer from the teacher (see Sec. 4.2). (iii) *Noise Correction*, a training-free inference modification that improves the sample quality of few-step methods that were trained in noise prediction mode (see Sec. 4.3).

In what follows, we assume access to a pre-trained diffusion model Φ, which predicts noise estimates $\hat{\boldsymbol{\epsilon}}_\Phi$. This *teacher model* can operate in either image or latent space. Our goal is to distill the knowledge of Φ into a *student model* Θ, while reducing the overall number of sampling steps, and providing high quality increases per extra step allowed in Θ. If the Φ model uses classifier-free guidance (CFG), then we also distill this knowledge into our student and eliminate the need for CFG.

4.1 Backward Distillation

It is widely recognized that conventional noise schedulers often fail to achieve zero terminal SNR at $t = T$ [16], thereby creating a discrepancy between training and inference. Specifically, the noise schedule (α_T, σ_T) in Eq. 1 is commonly chosen s.t. $\mathbf{x}_T$ is not pure noise during training, but rather contains low-frequency information leaked from $\mathbf{x}_0$. This discrepancy leads to performance degradation during inference, especially when taking only a few steps. To overcome this issue, Lin *et al.* [16] suggest to rescale existing noise schedules under a variance-preserving formulation to enforce zero terminal SNR.

However, we argue that this solution is not sufficient as information leakage occurs not only at $t = T$, but at all t's via the forward diffusion Eq. 1. Recall that the distillation loss gradient is computed at every training step as follows:

$$\nabla_{\Theta} \left\| \mathbf{x}^{\Phi}_{t\to 0} - \hat{\mathbf{x}}_0(\mathbf{x}_t, \hat{\boldsymbol{\epsilon}}_{\Theta}, t) \right\|^2 , \tag{6}$$

where $\hat{\mathbf{x}}_0(\cdot)$ is defined in Eq. 5 and $\mathbf{x}^{\Phi}_{t\to 0} = f^{|k|}_{\Phi}(\mathbf{x}_t)$ is the k step teacher prediction. Now, since $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \mathbf{x}_T$ even when enforcing zero *terminal* SNR ($\mathbf{x}_T = \boldsymbol{\epsilon}$) as suggested by [16], any stochastic interpolant $\mathbf{x}_t, t < T$ still contains information from the ground-truth sample via the first summand $\alpha_t \mathbf{x}_0$. As a result, the model learns to denoise taking into account information from the ground-truth signal. The smaller the t, the stronger the presence of the signal, and thus the more it will learn to preserve it. Let $\mathbf{x}^{\Theta}_{T\to t} = f^{|T-t|}_{\Theta}(\mathbf{x}_T)$ be the student's estimate at time t starting from pure noise at T in $|T - t|$ steps (see Eq. 4). During inference, the signal contained in $\mathbf{x}^{\Theta}_{T\to t}$ is *no longer ground-truth* signal $\mathbf{x}_0$, but rather the student's own best guess $\mathbf{x}^{\Theta}_0 := \hat{\mathbf{x}}_0(\mathbf{x}_{t+1}, \hat{\boldsymbol{\epsilon}}_{\Theta}, t + 1)$ from the previous step (see Eq. 3). As a result, models that have been trained to preserve a given signal will continue to propagate errors from previous steps instead of correcting them.

We propose a solution to ensure signal consistency between training and inference at all times. This is achieved by simulating the inference process during training, a method we term *backward distillation*. Unlike standard forward distillation, during training we do not begin sampling from the forward-noised latent code $\mathbf{x}_t = q(\mathbf{x}_0, \boldsymbol{\epsilon}, t)$. Instead, we first perform backward iterations of the student model to obtain $\mathbf{x}^{\Theta}_{T\to t} = \text{stopgrad}(f^{|T-t|}_{\Theta}(\mathbf{x}_T))$, and then use this as input for both the student and teacher models during training (see Fig. 3). The training gradients are then computed as

$$\nabla_{\Theta} \left\| \hat{\mathbf{x}}_0(f^k_{\Phi}(\mathbf{x}^{\Theta}_{T\to t}), \hat{\boldsymbol{\epsilon}}_{\Phi}, t/k) - \hat{\mathbf{x}}_0(\mathbf{x}^{\Theta}_{T\to t}, \hat{\boldsymbol{\epsilon}}_{\Theta}, t) \right\|^2 , \tag{7}$$

where $\mathbf{x}^{\Phi}_0 := \hat{\mathbf{x}}_0(f^k_{\Phi}(\mathbf{x}^{\Theta}_{T\to t}), \hat{\boldsymbol{\epsilon}}_{\Phi}, t/k)$ represents the target produced by running the teacher for k time-uniform denoising steps (from timestep t to t/k) with CFG, starting from the current latent code $\mathbf{x}^{\Theta}_{T\to t}$.

In summary, *backward distillation* eliminates information leakage at all time steps t, preventing the model from relying on a ground-truth signal. This is achieved by simulating the inference process during training, which can also be interpreted as calibrating the student on its own upstream backward path.[1]

[1] Note that the computation of $\mathbf{x}^{\Theta}_{T\to t}$ is cheap, as we distill for few steps only and gradient computation are omitted $\forall t' > t$.

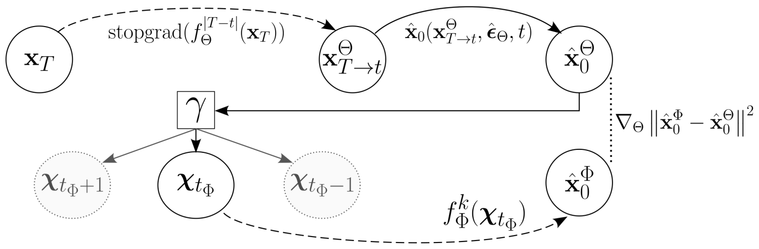

stopgrad($f_{\Theta}^{|T-t|}(\mathbf{x}_T)$)
$\hat{\mathbf{x}}_0(\mathbf{x}_{T\to t}^{\Theta}, \hat{\boldsymbol{\epsilon}}_{\Theta}, t)$
$\mathbf{x}_T$
$\mathbf{x}_{T\to t}^{\Theta}$
$\hat{\mathbf{x}}_0^{\Theta}$
γ
$\nabla_{\Theta} \left\|\hat{\mathbf{x}}_0^{\Phi} - \hat{\mathbf{x}}_0^{\Theta}\right\|^2$
$\boldsymbol{\chi}_{t_\Phi+1}$
$\boldsymbol{\chi}_{t_\Phi}$
$\boldsymbol{\chi}_{t_\Phi-1}$
$\hat{\mathbf{x}}_0^{\Phi}$
$f_{\Phi}^{k}(\boldsymbol{\chi}_{t_\Phi})$

Fig. 4: Shifted Reconstruction Loss (SRL). We propose a new distillation loss to improve the structure and adherence to detail of the student's predictions. $\mathbf{x}_{T\to t}^{\Theta}$ is the current noisy latent code at timestep t in the context of backward distillation. $\hat{\mathbf{x}}_0^{\Theta}$ is the image predicted by the student in one step from $\mathbf{x}_{T\to t}^{\Theta}$. SRL then entails noising $\hat{\mathbf{x}}_0^{\Theta}$ again to a t_{Φ} specified by the shifting function γ, followed by k uniformly sampled denoising steps from the teacher. The shifts are designed to adapt the type of knowledge distilled from the teacher for different t's in order to maximize efficacy.

4.2 SRL: Shifted Reconstruction Loss

In the process of image generation through backward diffusion, the early stages (where t is close to T) are instrumental in formulating the overall structure and composition of the image. Conversely, the later stages (where t is near 0) are essential for adding high-level details [3]. Drawing from this observation, we devise enhancements to the default knowledge distillation loss (Eq. 6), that encourage the student model to learn both the structural composition and detail-rendering capabilities of the teacher model. This involves shifting starting points for the teacher denoising away from the student's starting point t, hence we refer to this method as shifted reconstruction loss (SRL). Fig. 4 provides an overview of our proposed loss.

To obtain a target in SRL, instead of running the teacher model from the current noisy latent code $\mathbf{x}_{T\to t}^{\Theta}$ as in Eq. 7, we generate the target from the student's prediction $\mathbf{x}_0^{\Theta} = \hat{\mathbf{x}}_0(\mathbf{x}_{T\to t}^{\Theta}, \hat{\boldsymbol{\epsilon}}_{\Theta}, t)$ noised to $t_{\Phi} = \gamma(t)$, which we term $\boldsymbol{\chi}_{t_{\Phi}} = q(\mathbf{x}_0, \boldsymbol{\epsilon}, \gamma(t))$. As a result, the gradient updates are computed as

$$\nabla_{\Theta} \left\|\hat{\mathbf{x}}_0(f_{\Phi}^{k}(\boldsymbol{\chi}_{t_{\Phi}}), \hat{\boldsymbol{\epsilon}}_{\Phi}, \gamma(t)/k) - \hat{\mathbf{x}}_0(\mathbf{x}_{T\to t}^{\Theta}, \hat{\boldsymbol{\epsilon}}_{\Theta}, t)\right\|^2 . \tag{8}$$

Unlike conventional step distillation methods [24] where both the teacher and student begin with the same latent code, in SRL the mapping function $\gamma : [0, T] \to [0, T]$ is not defined as the identity function $\gamma(t) := t$. Instead, it is designed such that for higher values of t, the target produced by the teacher model displays global content similarity with the student output but with improved semantic text alignment; and for lower values of t, the target image features enhanced fine-grained details while maintaining the same overall structure as the student's prediction. This approach encourages the student to prioritize distilling structural knowledge during the early backward steps and focus on generating more refined details towards the final backward steps.

4.3 Noise Correction

The most common diffusion models are trained in noise prediction mode [8,28], which, according to Eq. 1, tasks the network with separating noise from signal given a randomly corrupted image. However, the process of sampling from diffusion model naturally starts from a point of pure noise, i.e. $\mathbf{x}_T = \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Consequently, there is no signal to be found in $\mathbf{x}_T$, and hence noise prediction at T becomes trivial but completely uninformative for the image generation process. To circumvent this singularity, existing works modify the noise schedule in Eq. 1, s.t. $\alpha_T = 0$ and $\sigma_t = T$ and switch to velocity prediction [16, 29]. Together, these changes ensure that the first update step at T is informative and unbiased.

Unfortunately, converting a model to velocity prediction requires extra training efforts. Hence, state-of-the-art few-step methods instead decide to remain in noise prediction mode, but compute loss on $\hat{\mathbf{x}}_0$ [17, 31]. While this circumvents the triviality problem of noise prediction at T, we argue that it also introduces a bias in the first update step. To see this, consider the first-order update in Eq. 4. The update step $f(\mathbf{x}_t)$ constitutes as a weighted sum of the current estimated signal $\hat{\mathbf{x}}_0$ and the model output $\boldsymbol{\epsilon}_\Theta$. Importantly, for noise prediction models, the estimated signal is a function of $\boldsymbol{\epsilon}_\Theta$ itself (Eq. 5). Now, since only the former ($\hat{\mathbf{x}}_0$) goes into the loss (see Eq. 6) and since there is no signal whatsoever in $\mathbf{x}_T$, the network is explicitly tasked *not* to predict $\boldsymbol{\epsilon}_\Theta = \boldsymbol{\epsilon}$ (which would give an all black image and hence high loss). As a result, using $\boldsymbol{\epsilon}_\Theta$ for the second part the update step in Eq. 3 biases the denoising process which leads to error accumulations.

To overcome this issue, we present a simple, training-free alternative to switching to zero-SNR velocity prediction [16] that allows the usage of noise prediction models without the aforementioned bias. Specifically, by treating $t = T$ as a unique case and replacing $\boldsymbol{\epsilon}_\Theta$ with the true noise $\mathbf{x}_T$, the update f is corrected:

$$f_\Theta(\mathbf{x}_t) = \begin{cases} \alpha_{T-1}\hat{\mathbf{x}}_0(\mathbf{x}_T, \hat{\boldsymbol{\epsilon}}_\Theta, T) + \sigma_{T-1}\boldsymbol{\epsilon} & \text{if} t = T, \\ \alpha_{t-1}\hat{\mathbf{x}}_0(\mathbf{x}_t, \hat{\boldsymbol{\epsilon}}_\Theta, t) + \sigma_{t-1}\hat{\boldsymbol{\epsilon}}_\Theta(\mathbf{x}_t, t) & \text{if} t < T. \end{cases} \tag{9}$$

We observed that this small modification can significantly improve the estimated colors, resulting in more vibrant and more saturated hues. This effect is particularly pronounced when the number of inference steps is low. We delve further into the effect of noise correction in Sect. 5.5 and provide qualitative comparisons in Figure 7 and Appendix D.

5 Experiments

To ensure fairness, we use the Emu model [4] as a base for all experiments. Emu is a state-of-the-art model with 2.7B parameters and resolution 768×768. We compare our results to previous distillation methods, such as Step Distillation [24], LCM [23], and ADD [31], by applying them directly on Emu. All models are replacement trained on a commissioned dataset of images. Since there is no publicly available code for ADD training, we implemented it ourselves based on the details provided in the paper [31].

5.1 Implementation Details

Like ADD [31], UFOGen [39] and Lightning [17], we train our model with an additional adversarial loss for improved image quality. Following ADD, we use the StyleGAN-T discriminator [30]. For single step models, we observe better image quality with a U-Net-based discriminator [32] crafted from the teacher U-Net, in line with UFOGen [39] and Lightning [17]. We choose timesteps $t \in \{999, 750, 500\}$ and $t \in \{999, 666\}$ for our 3-step and 2-step models, respectively. For SRL, we set $\gamma(t > 900) := 990$; $\gamma(900 \geq t > 500) := 950$; and $\gamma(t \leq 500) := 200$. From there, the teacher model $\varPhi$ takes $k = 8$ uniformly spaced time steps. Training was conducted for 15k iterations on 8 NVIDIA A100 GPUs, using the Adam [12] optimizer with a learning rate of 5e−6 for the U-Net and 1e−4 for the discriminator. For Emu Baseline model we run DPM++ [22] solver with 25 steps if not stated otherwise.

5.2 Quantitative Comparison to State of the Art

We compare Imagine Flash to previous methods using the FID [6], CLIP score [5], and CompBench [9]. FID and CLIP measure image quality and prompt alignment, and are evaluated on a split of 5k samples from COCO2017 [18], following the evaluation protocoll from [31]. CompBench is a benchmark that separately measures attribute binding (color, shape, and texture) and object relationships (spatial, non-spatial and complex). We generate 2 images per each prompt in CompBench validation set (300 prompts in total). For LCM and Imagine Flash, we compute the metrics for 1, 2, and 3 steps. For ADD, we compute the metrics for 4 steps, since this method was specifically tuned and configured for 4-step inference, ensuring a fair comparison. We also evaluate Step Distillation for 4 steps to provide a more direct comparison.

Table 1 shows the results. Our 3-step Imagine Flash outperforms Step Distillation and ADD in FID, even while using one less step. It also achieves lower FID than LCM for 1, 2, and 3 steps. The CLIP score of our 3-step model is higher than all variants of ADD and LCM and matches the score (30.2) of the 4-step Step Distillation model. Unlike Step Distillation and ADD, which degrade FID by 10.1 and 3.4 correspondingly compared to the Emu baseline, our 3-step and 2-step Imagine Flash preserve FID with slight improvements. For CompBench,

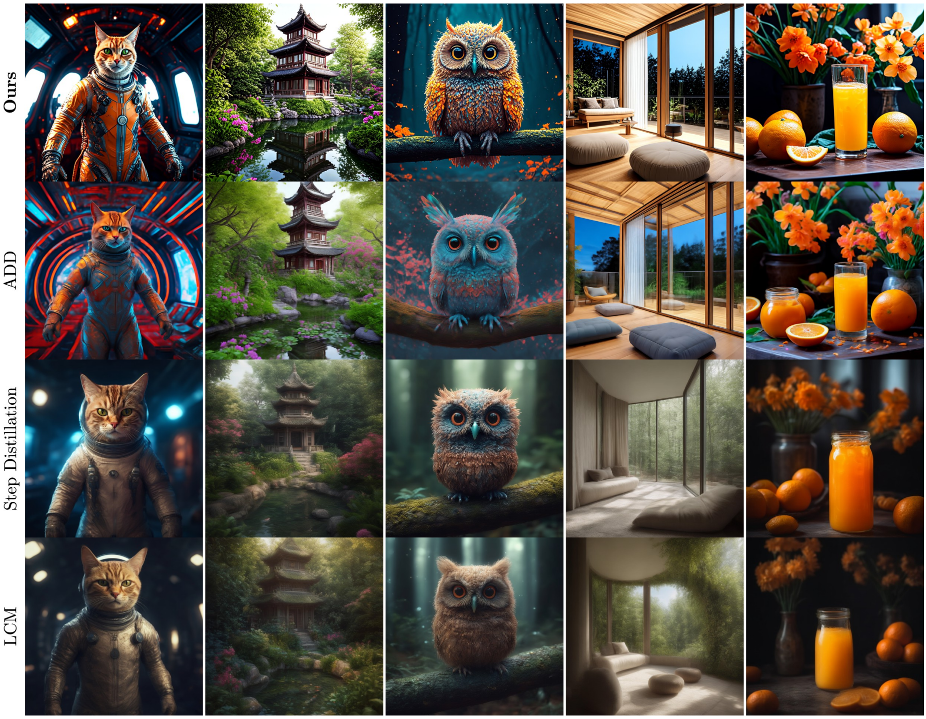


Fig. 5: Imagine Flash vs. SOTA. We show image generations of Imagine Flash and SOTA methods, all applied to Emu baseline. Every column is generated using the same random seed. Imagine Flash provides better realism, sharper images, and a higher level of detail. We attribute these gains to the proposed distillation method, which includes backward distillation, SRL, and noise correction.

any of our 1-, 2-, or 3-step Imagine Flash outperforms previous methods in all categories, except color, where 4-step Step Distillation and ADD score similarly to ours. This highlights the superior prompt alignment of Imagine Flash.

5.3 Qualitative Comparison to State of the Art

In Fig. 5, we show qualitative comparison of Imagine Flash to the current state-of-the-art (SOTA): Step Distillation, LCM, and ADD, all distilling the same baseline Emu model for fair comparison. We observe that ADD images are more crisp than those generated by step distillation and LCM due to the use of an adversarial loss. While both Imagine Flash and ADD use a discriminator, Imagine Flash produces sharper and more detailed images than ADD. The enhanced sharpness and detail of Imagine Flash are due to our proposed SRL, which effectively refines high-frequency details of the student prediction as depicted in the last row of Fig. 6.

On the other hand, for ADD, the target images may display a significantly different color spectrum, exhibit color artifacts (see Fig. 6), and the colors may

Table 1: Imagine Flash vs. SOTA - Quantitative. We compare 1-, 2-, and 3-step versions of Imagine Flash against the Emu Baseline and other SOTA distillation methods – Step Distillation, ADD, and LCM, all using the same teacher model and initialization (Emu Baseline). Not only does Imagine Flash outperform other methods, but it also excels at preserving baseline FID metric: it remains the same between the baseline and Imagine Flash with 3 steps, while for the other methods it decreases significantly.

Method	Steps	FID ↓	CLIP ↑	CompBench ↑ (%)					
				Color	Shape	Texture	Complex	Spatial	Non-spatial
Emu Baseline	25	35.7	30.8	51.8	39.8	53.5	46.0	60.3	67.9
Step Distill. Emu	4	45.8	**30.2**	**44.1**	26.9	40.0	39.6	48.7	54.3
ADD-Emu	4	39.1	29.6	43.3	32.9	44.5	36.5	42.6	56.2
LCM-Emu	1	123.3	24.0	29.1	20.2	25.0	29.4	26.6	28.9
	2	67.4	28.44	32.6	18.9	25.7	33.1	39.3	43.1
	3	59.1	28.94	33.8	19.0	25.4	32.5	40.8	44.2
Imagine Flash	1	40.4	29.0	40.9	**37.5**	46.5	41.8	53.1	59. 5
	2	**34.7**	30.1	44.0	36.3	**49.3**	42.4	**57.9**	61.4
	3	35.5	**30.2**	42.7	36.2	47.9	**42.8**	57.4	**62.9**

fluctuate unpredictably during training iterations. We hypothesize that, to minimize the L2 reconstruction loss in expectation, the ADD model fares best by predicting color values close to zero, leading to pale images and blurry outlines.

In addition to improving local details, SRL can also correct text-alignment mistakes of the student, as can be seen on the right side of Fig. 6 (1-step), where the small panda is transformed back into a dog.

5.4 Comparison to Public Models

We also compare the performance of Imagine Flash to the public models released by ADD-LDMXL [31] and Lightning-LDMXL [17].[2] To this end, we compute CLIP and FID scores as detailed in Section 5.2 and compare the relative gain/drop versus the base model. Please find the table in Appendix A. Our method retains similar text alignment capacity to ADD and Lightning but shows a much more favourable FID increase, especially for two and three steps.

In addition we conducted extensive human evaluations. To this end, we generate images for all methods using three inference steps on 1,000 randomly sampled prompts from the OUI dataset [4]. Paired images are presented to a subset of five from a total of 42 trained human annotators, who are tasked with voting for the more visually appealing image. The results, aggregated by majority voting, are displayed in Table 2, where a clear preference for Imagine Flash is evident.

[2] This comparison should be interpreted with caution as these methods start from a different base model.

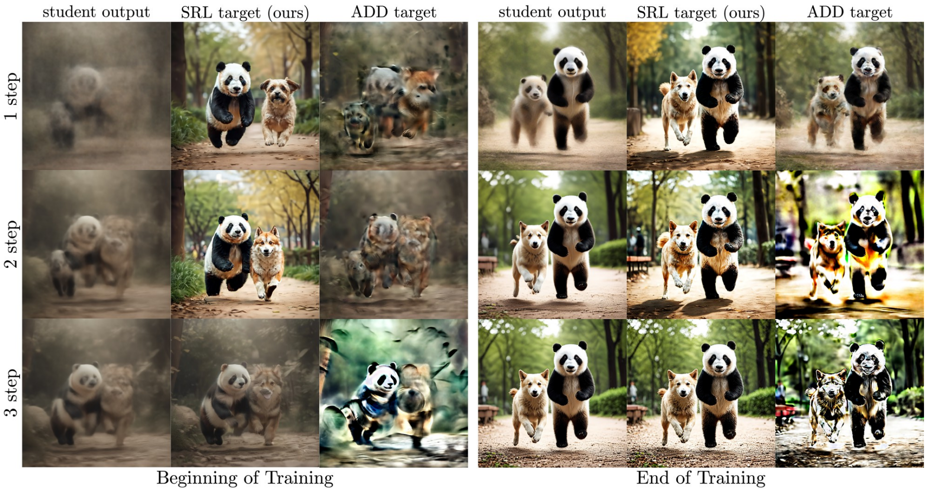


Fig. 6: Qualitative Effect of SRL. The SRL distillation target consistently encourages the generation of better semantics (step 1), richer colors (step 2), and more high-frequency details (step 3) at all training stages of Imagine Flash. In contrast, we do not observe this effect when using ADD.

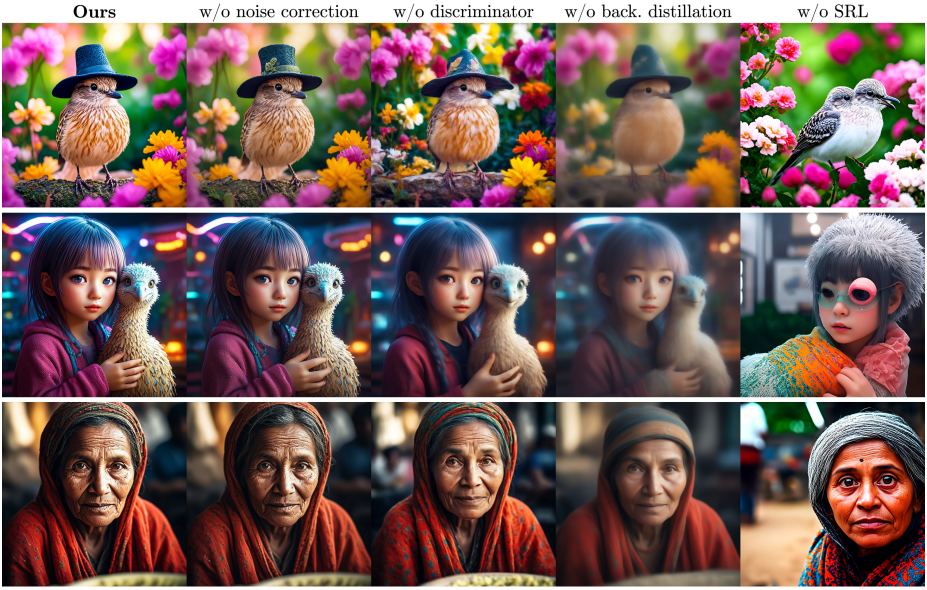


Fig. 7: Qualitative Ablations. We present visual ablation of the components of Imagine Flash. Noise correction improves colors and saturation. The discriminator improves colors and sharpness. However, the biggest impact on sharpness comes from backward distillation. Without the SRL reconstruction loss, the images suffer from artifacts and poor prompt alignment.

5.5 Ablations

We conduct quantitative and qualitative ablations on Imagine Flash to evaluate the effect of the proposed *backward distillation*, *SRL*, and *noise correction.* A quantitative evaluation is shown in Table 3, while a complementary visual ablation is provided in Figure 7.

Backward distillation: Without backward distillation, the CLIP score reduces by two points and FID deteriorates from 35.5 to 44.2 (Tab. 3), demonstrating the strong positive impact of backward distillation on image quality. As seen in Fig. 7, this quality degradation is characterized by increased blurriness.
SRL: Without the structure-aware reconstruction loss (i.e. using only the discriminator), FID increases by more than 10 points and CLIP decreases by 4.5 (Tab. 3). Fig. 7 shows an increase in artifacts, e.g., a bird with two heads, and worse prompt alignment.

Noise correction: While FID and CLIP are almost not affected by noise correction (Tab. 3), the effects are clearly seen in the visual ablation in Fig. 7. Noise correction results in more vivid colors and higher saturation.

Discriminator: Removing the discriminator slightly blurs the images and reduces color saturation (Fig. 7). Note that in contrast to previous works, which relay on the same discriminator [31], in our case the discriminator contributes much less to the image quality thanks to our backward distillation which largely addresses the blurriness problem.

With all components combined, Imagine Flash achieves an FID and CLIP close to the original Emu model (see Tab. 3).

Table 2: Human Evaluations on 1000 randomly sampled prompts from OUI. We report majority voting of 5 human annotators (in %). The annotators showed a clear preference for our model over ADD-LDMXL [31] and Lightning-LDMXL [17].

Method	Win	Tie	Loss
Ours vs ADD-LDMXL	**71.2**	1.2	27.6
Ours vs Lightning-LDMXL	**60.6**	21.0	18.4

Table 3: Ablations. We show the individual effects of every components of Imagine Flash. While noise correction and discriminator have a positive effect on performance, the main gains come from backward distillation and SRL.

Method	FID ↓	CLIP ↑
Emu Baseline DPM 25 Steps	35.7	30.8
Emu Baseline DPM 3 Steps	66.5	25.6
Ours	**35.5**	**30.2**
w/o Noise Correction	35.5	30.0
w/o Discriminator	39.0	29.0
w/o Backward Distillation	44.2	28.2
w/o SRL	45.9	25.7

6 Limitations

Constraints of Human Evaluation. Our human evaluation is based on a relatively large set of 1000 Open User Input prompts. Each pair of images is shown to a random set of five out of 42 human annotators and results are aggregated by majority voting. However, this approach may not entirely represent the real-world application of the models. The human evaluation of text-to-image models, particularly in terms of aesthetics, is inherently subjective and prone to variability. Consequently, evaluations conducted with a different set of prompts, annotators, or guidelines may yield varying results.

General Limitations of Text-to-Image Models. Similar to other text-to-image models, our models may occasionally produce biased, misleading, or offensive outputs. We have made substantial efforts to ensure the fairness and safety of our models. These efforts include the construction of balanced datasets, dedicated evaluation for high-risk categories, and extensive red teaming. Despite these measures, potential risks and biases may still exist.

7 Conclusion

We presented Imagine Flash, a novel distillation framework enabling high-fidelity few-step image generation with diffusion models. Our approach comprises three key components: Backward Distillation to reduce train-inference discrepancy, a Shifted Reconstruction Loss (SRL) dynamically adapting knowledge transfer per time step, and Noise Correction to enhance initial sample quality.

Through extensive experiments, Imagine Flash achieves remarkable results, matching the performance of the pre-trained teacher model using only three denoising steps and consistently surpassing existing methods. This unprecedented sampling efficiency combined with high sample quality and diversity makes our model well-suited for real-time generative applications.

Our work paves the way for ultra-efficient generative modeling. Future directions include extending to other modalities like video and 3D, further reducing the sampling budget, and combining our approach with complementary acceleration techniques. By enabling on-the-fly high-fidelity generation, Imagine Flash unlocks new possibilities for real-time creative workflows and interactive media experiences.

8 Acknowledgement

This work would not have been possible without a large group of collaborators who helped with the underlying base model, infrastructure, data, privacy, and the evaluation framework. We extend our gratitude to the following people for their contributions (alphabetical order): Eric Alamillo, Andres Alvarado, Giri Anantharaman, Stuart Anderson, Snesha Arumugam, Christine Awad, Chris Bray, Matt Butler, Vijaya Chandra, Li Chen, Lawrence Chen, Anthony Chen, Jessica Cheng, Vincent Cheung, Michael Cohen, Lauren Cohen, Xiaoliang Dai, Andrew Denyes, Abhimanyu Dubey, Cristian Canton Ferrer, Assaf Gelber, Jort Gemmeke, Freddy Gottesman, Freddy Gottesman, Nader Hamekasi, Jack Hanlon, Zecheng He, Zijian He, Tarek Hefny, Ji Hou, Shuming Hu, Jiabo Hu, Ankit Jain, Kinnary Jangla, Abhishek Kadian, Daniel Kreymer, Praveen Krishnan, Carolyn

Krol, Kunpeng Li, Tianhe Li, Kevin Chih-Yao Ma, Chih-Yao Ma, Dhruv Mahajan, Dhruv Mahajan, Mo Metanat, Dan Moskowitz, Simran Motwani, Vivek Pai, Mitali Paintal, Guan Pang, Devi Parikh, Hyunbin Park, Vladan Petrovic, Filip Radenovic, Vignesh Ramanathan, Ankit Ramchandani, Stephen Roylance, Kalyan Saladi, Mayank Sanganeria, Dev Satpathy, Alex Schneidman, Shubho Sengupta, Hardik Shah, Shivani Shah, Yaser Sheikh, Mitesh Kumar Singh, Animesh Sinha, Karthik Sivakumar, Yiwen Song, Lauren Spencer, Ragavan Srinivasan, Harihar Subramanyam, Fei Sun, Khushboo Taneja, Kate Trono, Sam Tsai, Mor Tzur, Simon Vandenhende, Jialiang Wang, Mike Wang, Rui Wang, Xiaofang Wang, Mack Ward, Yi Wen, Bichen Wu,Felix Xu, Seiji Yamamoto, Licheng Yu, Matthew Yu, Hector Yuen, Luxin Zhang, Peizhao Zhang, Yinan Zhao, Yue Zhao, Jessica Zhong, and Tali Zvi.

Finally, thank you Manohar Paluri, Ryan Cairns, Connor Hayes and Ahmad Al-Dahle for your support and leadership.

References

1. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)
2. Anderson, B.D.: Reverse-time diffusion equation models. Stochastic Processes and their Applications **12**(3), 313–326 (1982)
3. Castillo, A., Kohler, J., Pérez, J.C., Pérez, J.P., Pumarola, A., Ghanem, B., Arbeláez, P., Thabet, A.: Adaptive guidance: Training-free acceleration of conditional diffusion models. arXiv preprint arXiv:2312.12487 (2023)
4. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
5. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
7. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
10. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**(4) (2005)
11. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

13. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
14. Lee, S., Kim, B., Ye, J.C.: Minimizing trajectory curvature of ode-based generative models. arXiv preprint arXiv:2301.12003 (2023)
15. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. arXiv preprint arXiv:2306.00980 (2023)
16. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5404–5411 (2024)
17. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
21. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems **35**, 5775–5787 (2022)
22. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022)
23. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
24. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14297–14306 (2023)
25. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
27. Pooladian, A.A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., Chen, R.: Multisample flow matching: Straightening flows with minibatch couplings. arXiv preprint arXiv:2304.14772 (2023)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
30. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515 (2023)
31. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023)

32. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8207–8216 (2020)
33. Shaul, N., Perez, J., Chen, R.T., Thabet, A., Pumarola, A., Lipman, Y.: Bespoke solvers for generative flow models. arXiv preprint arXiv:2310.19075 (2023)
34. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
36. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
38. Wimbauer, F., Wu, B., Schoenfeld, E., Dai, X., Hou, J., He, Z., Sanakoyeu, A., Zhang, P., Tsai, S., Kohler, J., et al.: Cache me if you can: Accelerating diffusion models through block caching. arXiv preprint arXiv:2312.03209 (2023)
39. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. arXiv preprint arXiv:2311.09257 (2023)
40. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22552–22562 (2023)
41. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: The Eleventh International Conference on Learning Representations (2022)
42. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. arXiv preprint arXiv:2302.04867 (2023)
43. Zhao, Y., Xu, Y., Xiao, Z., Hou, T.: Mobilediffusion: Subsecond text-to-image generation on mobile devices. arXiv preprint arXiv:2311.16567 (2023)
44. Zheng, K., Lu, C., Chen, J., Zhu, J.: Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)

A Quantitative Results

In Table 4 we compare 1, 2 and 3 step versions of Imagine Flash against the publically available models of SOTA few-step methods – LDMXL-Turbo [31] and LDMXL-Lightning [17]. Since our method has a different base model, we report numbers in relative terms to the base model. As can be seen, Imagine Flash shows similar CLIP score preservation as Lightning. Turbo's CLIP scores improve over the teacher for the one step model, which is likely to be an artifact of the metric itself rather than a sign of superior text alignment. Regarding image quality, we find that Imagine Flash shows significantly less FID score degradation than both Turbo and Lightning. In fact, our 2 and 3 step model show even slight improvements in FID.

Table 4: Imagine Flash vs. public SOTA - Quantitative.

Method	FID ↓	CLIP ↑
LDMXL Baseline (25 Steps)	100% (24.50)	100% (32.01)
LDMXL-Turbo (3 Steps)	129.9% (31.80)	99.2% (31.76)
LDMXL-Turbo (2 Steps)	133.5% (32.72)	99.8% (31.94)
LDMXL-Turbo (1 Step)	132.3% (32.40)	**100.5**% (32.17)
LDMXL-Lightning (3 Steps)	132.8% (32.50)	98.2% (31.44)
LDMXL-Lightning (2 Steps)	130.9% (32.09)	97.1% (31.09)
LDMXL-Lightning (1 Step)	125.1% (30.60)	95.2% (30.46)
EMU Baseline (25 Steps)	100% (35.74)	100% (30.76)
Imagine Flash (3 Steps)	99.4% (35.53)	98.2% (30.19)
Imagine Flash (2 Steps)	**98.1**% (34.70)	98.0% (30.14)
Imagine Flash (1 Step)	116.3% (40.37)	94.4% (29.03)

B Generation Variance

In Figures 8 and 9, we demonstrate the proficiency of our model in generating a diverse array of high-quality samples, all derived from the same text prompt, with the sole variation being the initial noise.

C Generated Images

In Figures 10 and 11, we provide further qualitative examples demonstrating the performance of our model.

D Noise Correction

In Figure 12, we present additional examples illustrating the impact of our proposed noise correction technique. Note the enhancement in image quality, characterized by improved contrast and color saturation. This method further intensifies the dark colors while simultaneously amplifying the brightness of light colors.

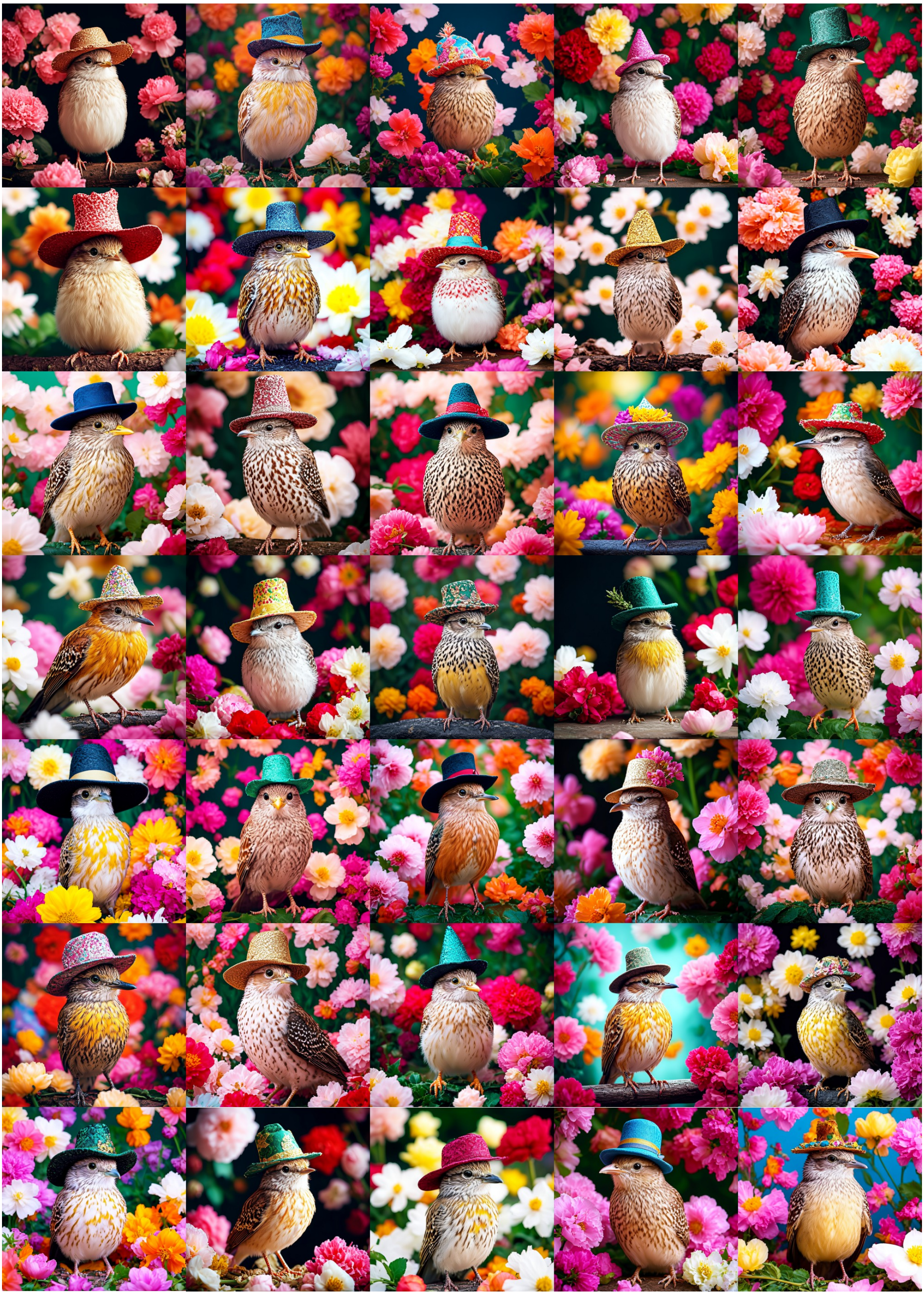


Fig. 8: Prompt Variance. Images generated with the proposed model given the prompt 'Cute bird in fancy hat with flowers on the background' with different initial noise.

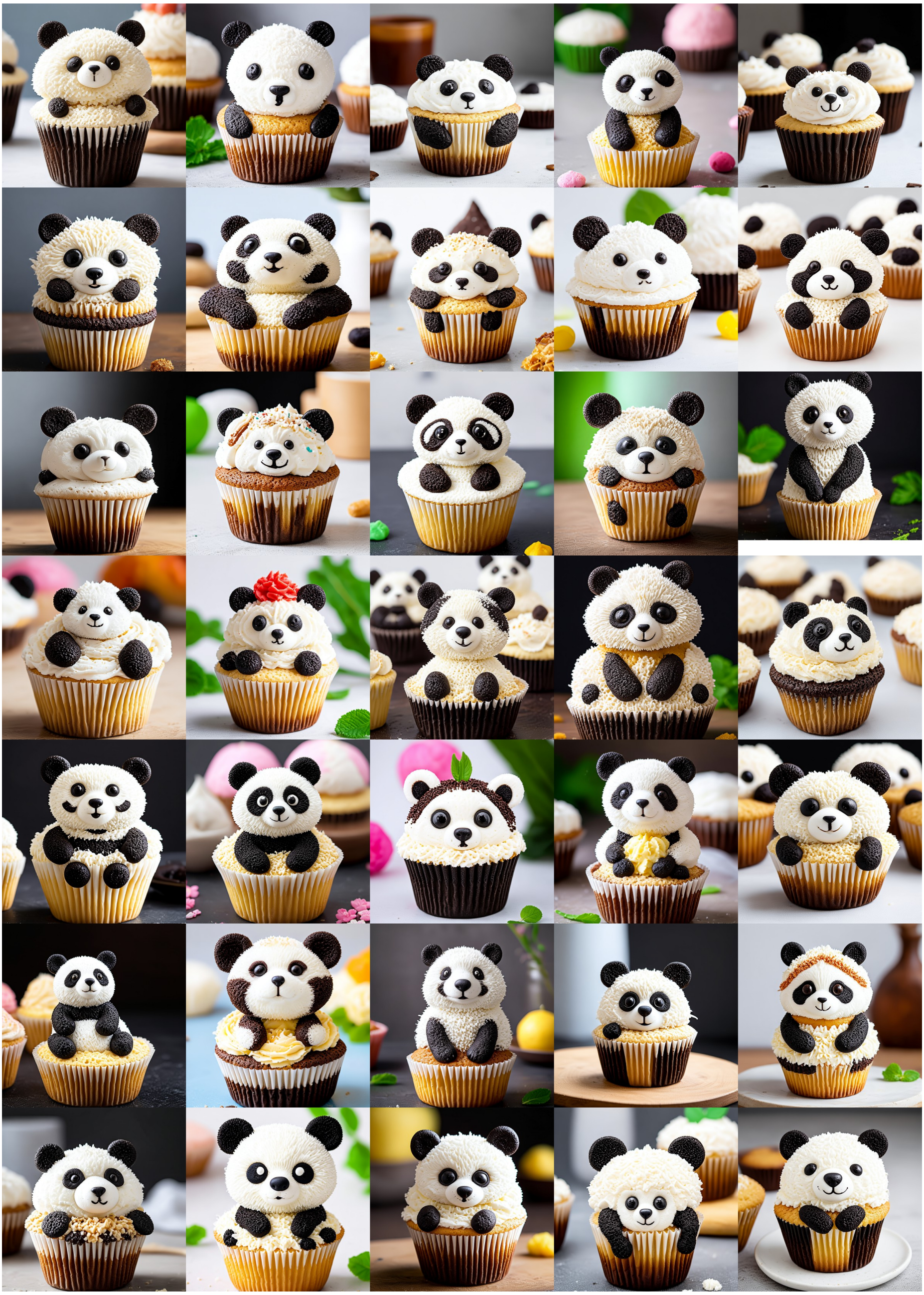
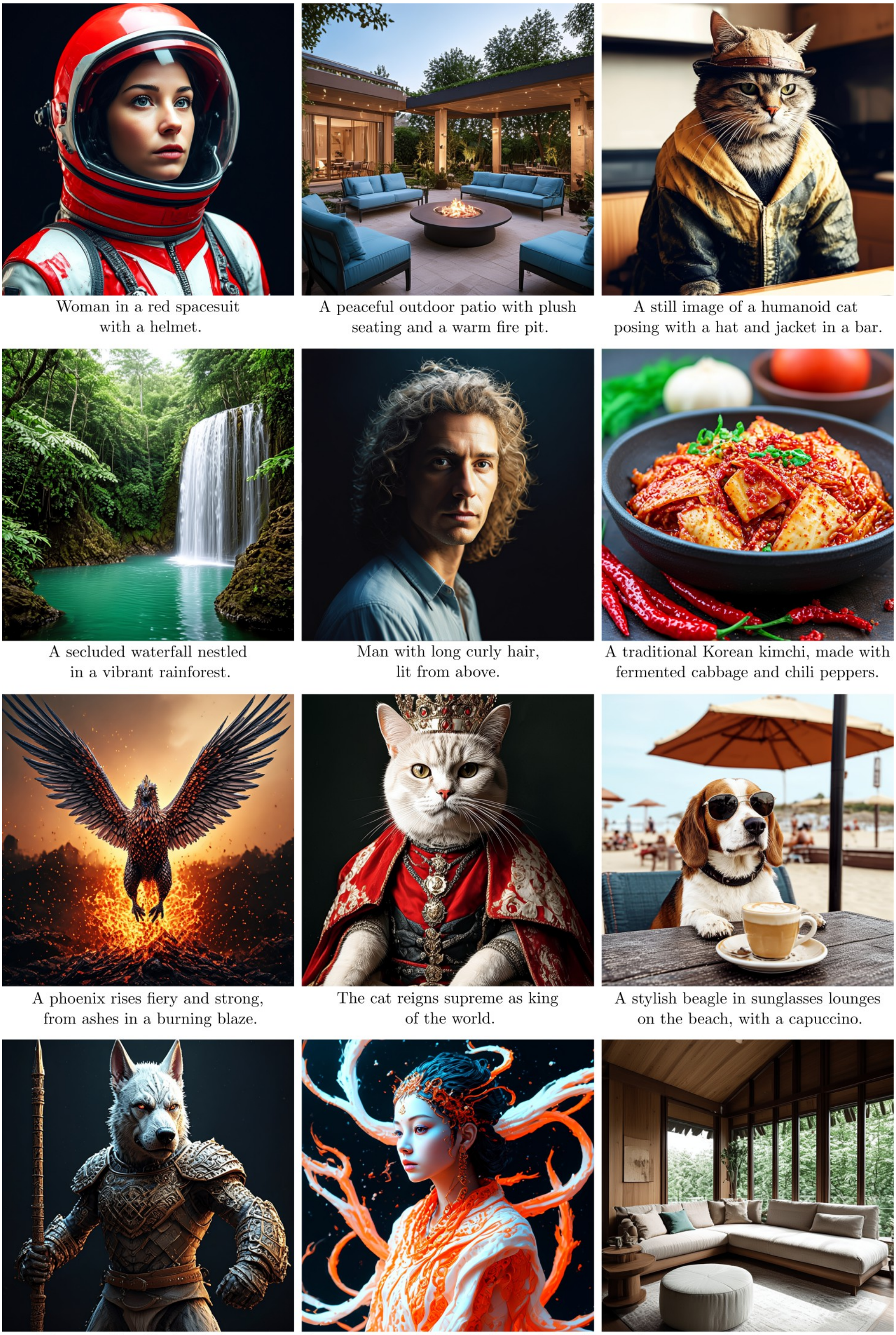


Fig. 9: Prompt Variance. Images generated with the proposed model given the 'A cupcake in the shape of a panda' with different initial noises.



Woman in a red spacesuit with a helmet.

A peaceful outdoor patio with plush seating and a warm fire pit.

A still image of a humanoid cat posing with a hat and jacket in a bar.

A secluded waterfall nestled in a vibrant rainforest.

Man with long curly hair, lit from above.

A traditional Korean kimchi, made with fermented cabbage and chili peppers.

A phoenix rises fiery and strong, from ashes in a burning blaze.

The cat reigns supreme as king of the world.

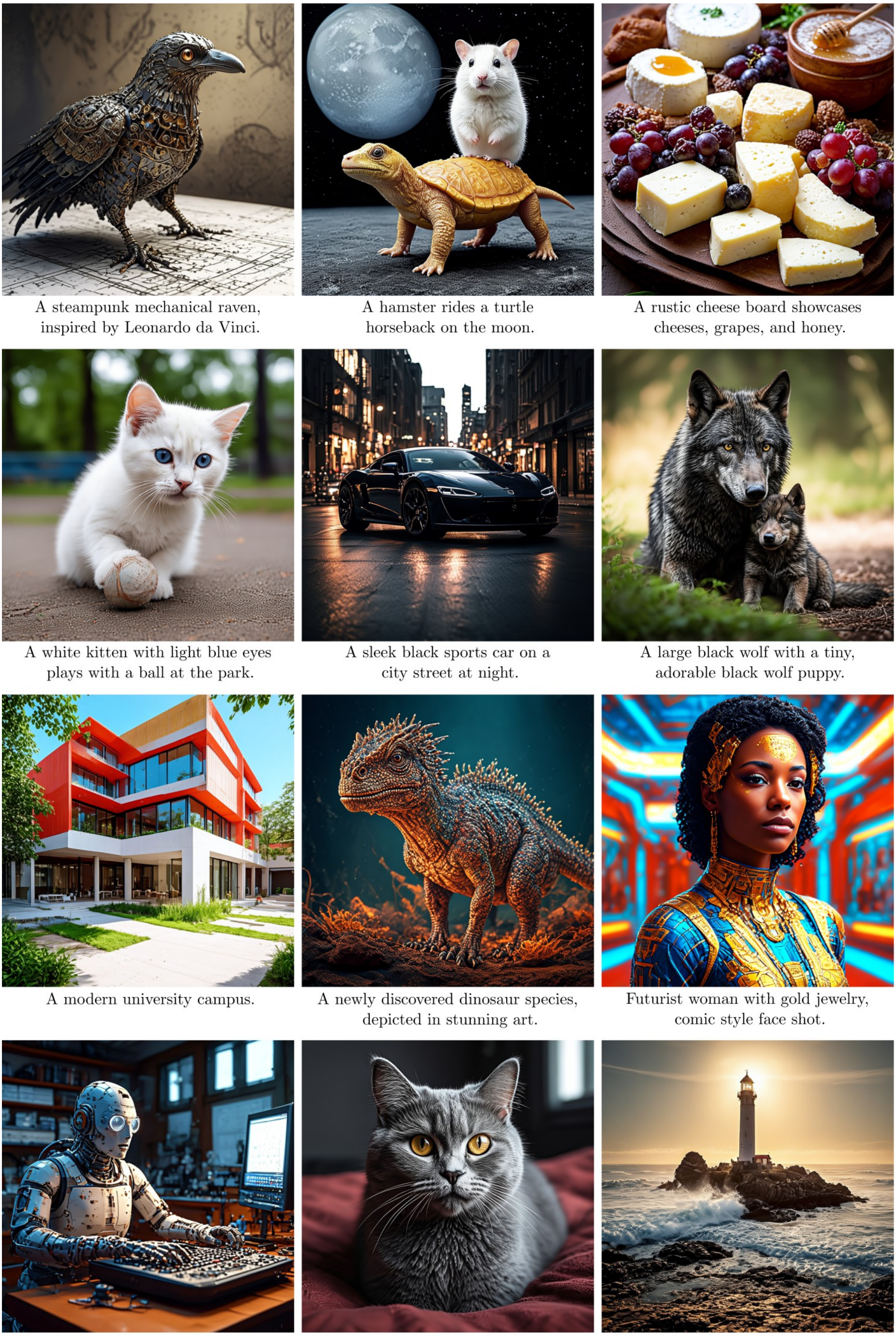
A stylish beagle in sunglasses lounges on the beach, with a capuccino.

A humanoid wolf, clad in heavy armor, wields a massive spear.

Concept art depicting Hindu mythology's wisps and tendrils in wlop style.

A serene meditation space with plush cushions and natural accents.

Fig. 10: Generated Images. Images generated with the proposed model. Each with the corresponding text prompt.



A steampunk mechanical raven, inspired by Leonardo da Vinci.

A hamster rides a turtle horseback on the moon.

A rustic cheese board showcases cheeses, grapes, and honey.

A white kitten with light blue eyes plays with a ball at the park.

A sleek black sports car on a city street at night.

A large black wolf with a tiny, adorable black wolf puppy.

A modern university campus.

A newly discovered dinosaur species, depicted in stunning art.

Futurist woman with gold jewelry, comic style face shot.

A robot is working in a laboratory, solving complex problems.

A gray cat with big, sad yellow eyes looks up from a bed in a dimly lit room.

A lighthouse on a cliffside, surrounded by crashing waves, during golden hour.

Fig. 11: Generated Images. Images generated with the proposed model. Each with the corresponding text prompt.

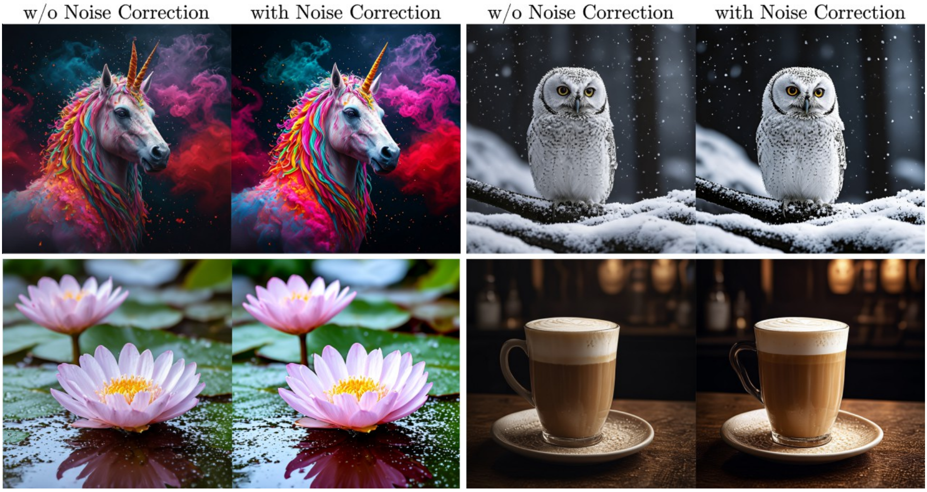


Fig. 12: Visual demonstration of the effect of Noise Correction. We observe an enhancement in image quality through improved contrast and color saturation. Best viewed on screen.