MANIFOLD-CONSTRAINED INITIAL NOISE OPTIMIZATION FOR EFFICIENT GENERATIVE MODEL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in distillation and flow-map models have enabled deterministic one- or few-step generation for high-quality data, facilitating a new branch of reward alignment approaches that directly optimize the initial noise from a Gaussian distribution. However, most existing initial-noise optimization methods rely on first-order gradient information, which is either inapplicable or suffers from instability and inefficiency in black-box reward scenarios. Here, we introduce ZeNOVA (**Ze**roth-order **N**oise **O**ptimization with **V**alue **A**nnealing), a stable and efficient initial noise alignment method in a gradient-free manner. Specifically, we address existing algorithms' major challenge in black-box scenarios through annealed soft-value guidance, manifold-constrained hyperspherical Langevin dynamics, and Metropolis-Hastings jumping. Extensive experiments on image and video generative models show that ZeNOVA outperforms all evaluated zeroth-order baselines by optimizing the initial noise toward higher rewards substantially more stably while exploiting the geometry of the Gaussian prior, demonstrating its practical applicability to various black-box reward alignment.

1 INTRODUCTION

Reward alignment of a pre-trained generative model aims to align its output distribution to admit higher external reward while minimizing its deviation from the original prior. A wide range of fine-tuning based approaches are presented grounded on several theoretical support, including entropy-regularized reinforcement learning (Black et al., 2024; Fan et al., 2023; Liu et al., 2025a; He et al., 2026; Zheng et al., 2026a), GFlowNets (Zhang et al., 2025; Liu et al., 2025c;b), and stochastic optimal control (Domingo-Enrich et al., 2025). Inference-time methods employing various techniques such as diffusion posterior sampling (Chung et al., 2023; Ye et al., 2024; Yu et al., 2023; He et al., 2023; Song et al., 2023a) and path integral control (Huang et al., 2024; Kim et al., 2025) have also been proposed to implement the same effect without additional training.

However, these works have to assume stochastic sampling trajectories involving many intermediate steps for their methods to analytically admit the reward-tilted distribution and thereby satisfy both the prior and high reward. Consequently, naively extending the results of these works to deterministic generative models is either limited or does not guarantee its behavior as theoretically designed. Moreover, recent advances in generative modeling have led to the development of few-step or one-step models including distilled models (Yin et al., 2024b;a; Sauer et al., 2024) and flow-map models (Song et al., 2023b; Geng et al., 2025), which reduces the effect of control in between intermediate sampling process.

These recent advances facilitate a separate approach of initial-noise optimization as a practical route to inference-time reward alignment; a pretrained deterministic generator can produce different outputs by changing its Gaussian input, without updating its parameters (Eyring et al., 2024; Min et al., 2026; Yoon et al., 2026). Most existing approaches, however, obtain their optimization directions by differentiating through both the generator and the reward computation. This requirement limits their use when the reward is non-differentiable or available only through black-box evaluations, and can impose substantial computational overhead even when gradients are accessible. Estimating directions from reward queries offers a gradient-free alternative, but useful alignment also requires these estimates to support stable and efficient optimization in the high-dimensional noise space. In

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

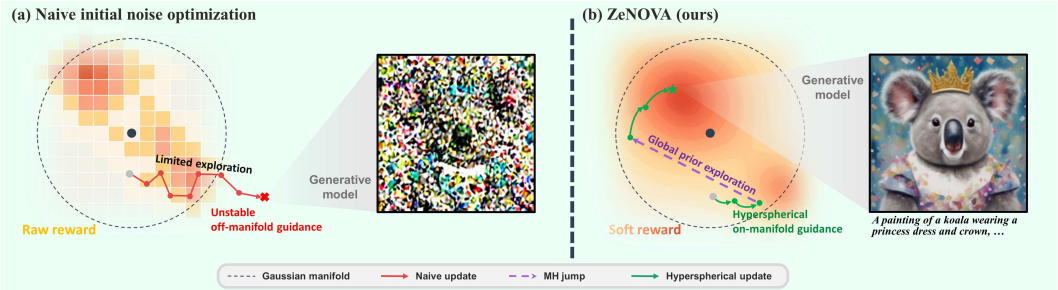

Figure 1: **Overview of ZeNOVA.** Naive initial-noise optimization suffers from limited exploration and unstable updates under inaccurate zeroth-order guidance from raw reward signal. ZeNOVA combines annealed soft-value guidance, hyperspherical updates, and Metropolis-Hastings jumping to enable stable local refinement and global exploration while respecting the Gaussian prior geometry.

particular, seeking higher rewards should account for the Gaussian noise input geometry on which the generator was trained, which still remains challenging with optimizations in gradient-free manner.

Here, we propose **ZeNOVA**, a stable and efficient gradient-free initial-noise alignment method toward black-box objectives. Specifically, as illustrated in Figure 1, we identify the major challenges faced by existing initial noise optimization algorithms in black-box scenarios, and address them through three contributions of annealed soft-value guidance, manifold-constrained hyperspherical Langevin dynamics, and Metropolis-Hastings jumping. Together, these components require only forward evaluations of the generator and reward model, making ZeNOVA applicable to black-box alignment without backpropagation. Experiments on image and video generative models show that ZeNOVA achieves higher rewards than all evaluated zeroth-order baselines while maintaining more stable initial-noise optimization, supporting its practical use across diverse reward alignment tasks.

2 Preliminaries

Reward-tilted Distribution. When p^{ref} and p^{θ} denote the distribution from the pretrained model and the fine-tuned model with parameters θ, Entropy-regularized RL formulates the alignment objective as balancing expected terminal reward against divergence from the reference process with a balancing parameter $\lambda > 0$:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{x}_0 \sim p_0^{\theta}} \left[r(\boldsymbol{x}_0) \right] - \lambda D_{\mathrm{KL}} \left(p^{\theta} \,\|\, p^{\mathrm{ref}} \right), \tag{1}$$

In many algorithms for solving Eq. (1) in diffusion and flow-matching models (Fan et al., 2023), the trajectory-level KL admits a decomposition into a sum or integral of local policy divergences along the sampling trajectory. When $t \in [0, 1]$ denotes the noise timestep and $\boldsymbol{x}_0$ denotes the clean sample, it is known that the optimal policy for Eq. (1) induces the following marginal distribution (Uehara et al., 2024b) with *Soft value function* $V(\boldsymbol{x}_t)$, which represents the expected reward and KL penalty term over the optimal policy from $\boldsymbol{x}_t$:

$$p_t^{\star}(\boldsymbol{x}_t) \propto p_t^{\mathrm{ref}}(\boldsymbol{x}_t) e^{\frac{1}{\lambda} V(\boldsymbol{x}_t)}, \quad V(\boldsymbol{x}_t) := \lambda \log \mathbb{E}_{p_t^{\mathrm{ref}}} \left[\exp\left(\frac{r(\boldsymbol{x}_0)}{\lambda} \right) \middle| \boldsymbol{x}_t \right] \tag{2}$$

Substituting the relationship between the output of a diffusion/flow-matching model and the score function (Song et al., 2021), Eq. (2) shows that the difference between the optimal and reference models shall be proportional to the gradient of this value function $\nabla_{\boldsymbol{x}_t} V(\boldsymbol{x}_t)$.

Among several mathematical frameworks, Path-Integral Control(PIC) (Kappen, 2005; Huang et al., 2024; Uehara et al., 2024a) allows us to express the optimal policy and soft value function through weighted samples of the reference dynamics:

$$\nabla_{\boldsymbol{x}_t} V(\boldsymbol{x}_t) = \lim_{dt \to 0} \frac{\lambda}{g(t) dt} \frac{\mathbb{E}_{P^{\mathrm{ref}}} \left[e^{r(\boldsymbol{x}_0)/\lambda} d\boldsymbol{w}_t \mid \boldsymbol{x}_t \right]}{\mathbb{E}_{P^{\mathrm{ref}}} \left[e^{r(\boldsymbol{x}_0)/\lambda} \mid \boldsymbol{x}_t \right]} \tag{3}$$

This representation weights sampled noise increments by terminal desirability, allowing the control to be estimated from samples of the reference model and terminal reward evaluations alone.

Reward-tilted Noise Distribution. Consider reward alignment of the input initial noise fed to the generative models, without assuming access to reward gradients. Let $\boldsymbol{z} \sim p_0 := \mathcal{N}(0, \mathbf{I}_d)$ denote the d-dimensional input noise, and $G_\theta(\boldsymbol{z}, \boldsymbol{c})$ be a generator conditioned on $\boldsymbol{c}$, leading to the reward in noise space and the reward-titled noise distribution:

$$R(\boldsymbol{z}) := r(G_\theta(\boldsymbol{z}, \boldsymbol{c})), \quad p^\star(\boldsymbol{z}) \propto p_0(\boldsymbol{z}) \exp\left(R(\boldsymbol{z})/\lambda\right) \tag{4}$$

To perform optimization of the initial noise $\boldsymbol{z}$, the simplest approach is to search for a maximizer of $R(\boldsymbol{z})$ over the entire support $\mathbb{R}^d$. But this is not only prone to mode collapse toward a few high-reward regions, it also disregards the Gaussian prior distribution of $\boldsymbol{z}$ on which the latter generative model G_θ is trained, leading to reward hacking and degradation of the image quality. While several empirical regularizers were suggested to mitigate this (Eyring et al., 2024; Shi et al., 2026), when R is differentiable with respect to its input $\boldsymbol{z}$, sampling from (4) is straightforward because its score function is tractable. Hence, the following Langevin dynamics has $p^\star$ as its stationary distribution.

$$d\boldsymbol{z} = \left(-\boldsymbol{z} + \frac{1}{\lambda}\nabla_{\boldsymbol{z}} R(\boldsymbol{z})\right) dt + \sqrt{2}\, d\boldsymbol{w} \tag{5}$$

This observation is utilized in several gradient-based initial-noise alignment algorithms (Min et al., 2026; Yoon et al., 2026).

3 METHOD

Our interest is a more challenging setting in which R can be evaluated but $\nabla_{\boldsymbol{z}} R$ is unavailable. Here, we first show why a direct zeroth-order extension of gradient-based Langevin optimization is inadequate, and then describe the proposed contributions ZeNOVA to address those challenges.

3.1 CHALLENGES FOR BLACK-BOX OBJECTIVES.

A natural zeroth-order baseline. For a small perturbation scale $\delta > 0$, a classic gradient estimation method with function evaluations according to Stein's Lemma (Stein, 1981) is

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{\delta} R(\boldsymbol{z} + \delta\boldsymbol{\epsilon})\boldsymbol{\epsilon}\right] = \nabla_{\boldsymbol{z}} \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}[R(\boldsymbol{z} + \delta\boldsymbol{\epsilon})] \approx \nabla_{\boldsymbol{z}} R(\boldsymbol{z}), \tag{6}$$

which can be estimated from N particles as

$$\hat{\boldsymbol{g}}(\boldsymbol{z}) = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{\delta}\left(R(\boldsymbol{z} + \delta\boldsymbol{\epsilon}^{(n)}) - \bar{R}\right)\boldsymbol{\epsilon}^{(n)}\right] \tag{7}$$

where $\bar{R}$ is an optional control variate to reduce variance, typically set as the mean of R. This is the familiar Gaussian-search estimator used in evolution strategies such as NES (Wierstra et al., 2014). Therefore, the most direct gradient-free extension of (5) is to replace $\nabla_{\boldsymbol{z}} R(\boldsymbol{z})$ by $\hat{\boldsymbol{g}}(\boldsymbol{z})$. We refer to this construction as *naive zeroth-order(ZO) Langevin*.

However, while this naive alternative seems theoretically sound, it performs poorly when applied to practical initial-noise optimization. We analyze the reasons for this failure through the following three failure modes.

Failure mode I: Nature of Black-box rewards. Many black-box rewards can be discrete, quantized, thresholded, ranking-based, or simply locally constant. In this case, the gradient (including its estimator) becomes 0 and does not carry any signal for the optimization. Naively increasing δ in Eq. (7) can search a broader range to recover reward signal, but this makes the estimator more deviate from the ground truth gradient.

Failure mode II: Destabilization from few-particle estimates. To make the zeroth-order estimator accurate, the required number of particles N has to grow proportionally to the number of dimensions d. However, in most high-dimensional noise spaces for recent generative models, a limited computational budget usually admits $N \ll d$, which introduces unavoidable inaccuracy in the estimator.

Empirically, we observe that this inaccuracy goes beyond making the optimization process less efficient, and frequently drives $\boldsymbol{z}$ completely out of distribution, leading to sample collapse. Figure 2

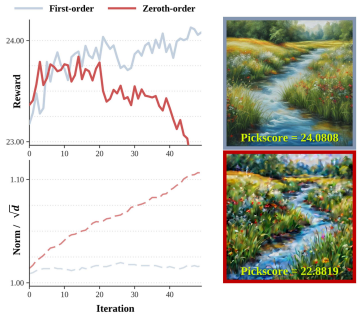

Figure 2: Replacing real gradients in Langevin dynamics(grey) with zeroth-order estimates(red).

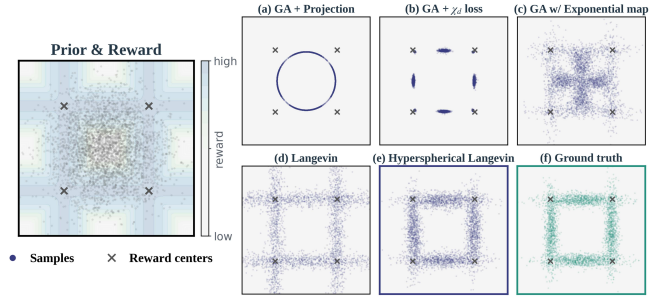

Figure 3: 2D toy experiment for sampling from reward-tilted target with zeroth-order gradient estimation.

shows an example of reward-tilted Langevin dynamics with Pickscore (Kirstain et al., 2023) failing when the actual reward gradient is replaced by $\hat{\boldsymbol{g}}(\boldsymbol{z})$. Note that when this happens, the norm of $\boldsymbol{z}$ drifts away from the expected value for a Gaussian distribution, rapidly degrading the model output.

Failure mode III: Hazard of bad initial states. Even if the aforementioned issues are resolved, a sequential and incremental motion of Langevin dynamics suffers from moving between separated modes slowly, as it requires traversing intermediate low-likelihood regions. This makes the finite-budget trajectories sensitive to their initialization.

3.2 ZeNOVA: Noise Optimization for Black-box Rewards with Value Annealing

Zeroth-order value-gradient estimator. Note that when the reward gradient control term is removed from Eq. (5), the remainder is an Ornstein–Uhlenbeck (OU) process (Øksendal, 2003) that preserves $p_0 = \mathcal{N}(0, \mathbf{I})$. We take this process as a reference and consider a soft value that aggregates future terminal rewards. The OU process admits an exact Gaussian transition over any time interval $h > 0$: setting $\sigma^2 = 1 - e^{-2h}$, we can express its transition using a single kernel,

$$q_\sigma(\boldsymbol{z}' \mid \boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z}'; \sqrt{1-\sigma^2}\boldsymbol{z}, \sigma^2\mathbf{I}\right), \qquad 0 < \sigma < 1. \tag{8}$$

The parameter σ specifies the *horizon* of the OU process: large σ nearly forgets the current state and broadly explores the reference prior, whereas small σ concentrates around the current state.

Following the previous work in the soft RL (Uehara et al., 2024b) and PIC (Kappen, 2005), we define the soft value as the log-expectation of the terminal reward $\exp(R/\lambda)$ under this transition kernel:

$$V_\sigma(\boldsymbol{z}) := \lambda \log \mathbb{E}_{\boldsymbol{z}' \sim q_\sigma(\cdot|\boldsymbol{z})}\left[\exp\left(\frac{R(\boldsymbol{z}')}{\lambda}\right)\right] \tag{9}$$

Then, we propose to sample from an intermediate distribution obtained by exponentially tilting the prior with this soft value, instead of the raw reward R:

$$p^\star_\sigma(\boldsymbol{z}) \propto p_0(\boldsymbol{z}) \exp\left(\frac{V_\sigma(\boldsymbol{z})}{\lambda}\right), \tag{10}$$

This choice is motivated by two properties of the value function: First, even when the reward function is poorly behaved (*e.g.*, discrete or piecewise constant), the value function is defined as its expectation over a Gaussian kernel, and therefore continuous and well-behaved (Jensen et al., 2026). Moreover, this intermediate distribution induced by the soft-value function has an exact relationship to $p^\star$:

Proposition 1 (Interpolation of target distribution)**.** *Let reward-tilted distribution $p^\star$ be defined as Eq. (4) and soft-value-tilted distribution $p^\star_\sigma$ as Eq. (10). Then $p^\star_\sigma$ is equal to $p^\star$ passed through a Gaussian kernel of Eq. (8):*

$$p^\star_\sigma(\boldsymbol{z}) \propto \int q_\sigma(\boldsymbol{z} \mid \boldsymbol{z}')\, p^\star(\boldsymbol{z}')\, d\boldsymbol{z}'. \tag{11}$$

This implies that $p^\star_\sigma \to p_0$ as $\sigma \to 1$, and $p^\star_\sigma \to p^\star$ as $\sigma \to 0$, where the proof is in Appendix A.1. Motivated by this interpolation, we use a decreasing horizon schedule $1 > \sigma_1 > \sigma_2 > \cdots > \sigma_K > 0$

with the aim of tracking the sequence $p^\star_{\sigma_1}, p^\star_{\sigma_2}, \ldots, p^\star_{\sigma_K}$ in a coarse-to-fine manner; larger horizons aggregate reward information over a broad neighborhood, while smaller horizons bring the target closer to the original reward tilt and support local refinement.

Crucially, ∇V_σ can be estimated without differentiating R. Applying the Gaussian score identity to (9), we obtain

$$\nabla_{\boldsymbol{z}} V_\sigma(\boldsymbol{z}) = \frac{\lambda\sqrt{1-\sigma^2}}{\sigma} \frac{\mathbb{E}_{\boldsymbol{\epsilon}}\left[\exp(R(\sqrt{1-\sigma^2}\boldsymbol{z}+\sigma\boldsymbol{\epsilon})/\lambda)\,\boldsymbol{\epsilon}\right]}{\mathbb{E}_{\boldsymbol{\epsilon}}\left[\exp(R(\sqrt{1-\sigma^2}\boldsymbol{z}+\sigma\boldsymbol{\epsilon})/\lambda)\right]}. \tag{12}$$

See Appendix A.3 for the derivation. With N reward evaluations, we use

$$\widehat{\boldsymbol{g}}_\sigma(\boldsymbol{z}) := \frac{\lambda\sqrt{1-\sigma^2}}{\sigma} \sum_{n=1}^{N} \frac{\exp(R(\sqrt{1-\sigma^2}\boldsymbol{z}+\sigma\boldsymbol{\epsilon}^{(n)})/\lambda)}{\sum_{j=1}^{N}\exp(R(\sqrt{1-\sigma^2}\boldsymbol{z}+\sigma\boldsymbol{\epsilon}^{(j)})/\lambda)}\boldsymbol{\epsilon}^{(n)} \qquad \boldsymbol{\epsilon}^{(n)} \sim \mathcal{N}(0,\mathbf{I}). \tag{13}$$

The normalized exponential weights are implemented with the usual log-sum-exp stabilization. For a fixed σ, we can replace the reward gradient in Eq. (5) with Eq. (13).

Hyperspherical Langevin dynamics. Even for soft value gradient functions, their Monte Carlo estimate can still be noisy for small N. Meanwhile, the standard Gaussian prior is rotationally symmetric, and its density depends only on the radius. Moreover, in high dimensions, Gaussian samples exhibit a uniform likelihood for any direction, whereas their norm concentrates within a narrow range around $\sqrt{d}$. Indeed, as described in Figure 2, failure of the update with zeroth-order estimation is often accompanied by a drift of the noise norm away from this Gaussian typical range.

This observation suggests treating the two degrees of freedom separately and formulating the Langevin dynamics in hyperspherical coordinates, separating the prior-sensitive radial motion from the angular motion. Note that sampling $\boldsymbol{z}$ from the standard Gaussian can be decomposed as:

$$\boldsymbol{z} = \rho\boldsymbol{u} \quad \text{where} \quad \rho \sim \chi_d, \boldsymbol{u} \in \text{Unif}(\mathbb{S}^{d-1}), \tag{14}$$

Therefore, for a given current state $\boldsymbol{z}_k$, we decompose it into its norm ρ_k and its unit direction vector $\boldsymbol{u}_k$, and separately perform *angular* Langevin update and *radial* Langevin update.

Angular Langevin on the sphere. Let $\mathbf{P}^\perp_{\boldsymbol{u}} := \mathbf{I} - \boldsymbol{u}\boldsymbol{u}^\top$ denote the projection onto the tangent space of $\boldsymbol{u}$. For $\boldsymbol{g}_\sigma(\boldsymbol{z}) := \nabla_{\boldsymbol{z}} V_\sigma(\boldsymbol{z})$, the angular update uses a gradient and noise on a tangent space, followed by spherical exponential map (Lee, 2018):

$$\boldsymbol{v}_k = \mathbf{P}^\perp_{\boldsymbol{u}_k}\left(\frac{\eta_A\rho_k}{\lambda}\widehat{\boldsymbol{g}}_{\sigma_k}(\boldsymbol{z}_k) + \sqrt{2\eta_A}\,\boldsymbol{\xi}_k\right), \qquad \boldsymbol{\xi}_k \sim \mathcal{N}(0,\mathbf{I}) \tag{15}$$

$$\boldsymbol{u}_{k+1} = \text{Exp}_{\boldsymbol{u}_k}(\boldsymbol{v}_k) \tag{16}$$

where $\text{Exp}(\cdot)$ denotes an exponential mapping as follows:

$$\text{Exp}_{\boldsymbol{u}}(\boldsymbol{v}) = \cos(\|\boldsymbol{v}\|_2)\boldsymbol{u} + \sin(\|\boldsymbol{v}\|_2)\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \tag{17}$$

Radial Langevin. Since the reference radial distribution is $\rho \sim \chi_d$, a radial update step is

$$\rho_{k+1} = \rho_k + \eta_R\left(\frac{d-1}{\rho_k} - \rho_k + \frac{1}{\lambda}\boldsymbol{u}_k^\top\widehat{\boldsymbol{g}}_{\sigma_k}(\boldsymbol{z}_k)\right) + \sqrt{2\eta_R}\,\xi_k, \qquad \xi_k \sim \mathcal{N}(0,1). \tag{18}$$

The radius remains stochastic and reward-adaptive, while noisy high-dimensional guidance is regularized by an analytic one-dimensional prior score rather than by an explicit norm constraint.

For any fixed σ, this continuous spherical-radial Langevin dynamics has (10) as its stationary distribution of $\rho\boldsymbol{u}$. This spherical-radial Langevin update allows us to design their dynamics separately(*e.g.*, small step size for radial update), which stabilizes noisy zeroth-order guidance without freezing the radius or relying on an explicit shell projection.

Metropolis-Hastings Jumping. Since incremental updates may be a somewhat inevitable limitation of Langevin dynamics to simulate an SDE, we briefly examine a fundamentally different alternative of utilizing the Metropolis–Hastings(MH) algorithm:

Proposition 2 (Metropolis-Hastings algorithm for reward-tilted distribtuion)**.** *The Metropolis–Hastings chain with proposal* $\boldsymbol{z}' \sim p_0$ *and acceptance probability* $\alpha(\boldsymbol{z}, \boldsymbol{z}') = \min\{1, \exp((R(\boldsymbol{z}') - R(\boldsymbol{z}))/\lambda)\}$ *converges to the reward-tilted distribution* $p^\star(\boldsymbol{z}) \propto p_0(\boldsymbol{z})e^{R(\boldsymbol{z})/\lambda}$.

See Appendix A.2 for the derivation. MH chain targets the same distribution with Langevin dynamics and provides global exploration, but it is inefficient as a stand-alone sampler: Namely, for a finite number of proposals, the MH algorithm is strictly worse than Best-of-K(BoK) sampling. Moreover, as the sample reward increases, every proposal is drawn afresh regardless of the current high-reward state, and the MH algorithm rejects almost all proposals. Nonetheless, this has a complementary behavior against Langevin dynamics, allowing occasional global moves to new reward modes with much smaller computation than gradient estimation. We therefore interleave a small number of independent MH transitions with Langevin updates, dubbed as *MH jumping*. MH jumping is effective in the early stage of the optimization, after which the annealed soft-value and hyperspherical Langevin updates perform coarse-to-fine refinement.

ZeNOVA: The Algorithm. Combining the three components yields **ZeNOVA**(**Ze**roth-order **N**oise **O**ptimization with **V**alue **A**nnealing), where the complete procedure is summarized in Algorithm 1 of Appendix B.1. The two principal hyperparameters governing inference-time computation are the number of particles N and the number of sampling iteration K. Increasing N improves the Monte Carlo estimate of the soft value gradient, while increasing K allows the sampler to traverse the interpolation more gradually. Importantly, ZeNOVA requires only forward evaluations of the generator and reward model.

4 Experiments

We evaluate ZeNOVA across a range of deterministic generators and reward functions to demonstrate its effectiveness, scalability, and broad applicability.

Toy Experiment. We design a reward-aligned sampling problem of 2-D data to identify the effect of existing sampling algorithms under zeroth-order gradient estimation. Starting from $p_0 = \mathcal{N}(0, \mathbf{I})$, we use a synthetic reward with narrow high-reward bands and aim to sample from $p^\star(\boldsymbol{z}) \propto p_0(\boldsymbol{z})e^{R(\boldsymbol{z})}$. Figure 3 compares (a) Gradient Ascent(GA) with fixed-radius projection (Kim et al., 2026), GA with radial χ_d regularization loss (Eyring et al., 2024), GA with exponential-map updates (Shi et al., 2026), (d) Cartesian Langevin updates, and (e) our hyperspherical Langevin updates against (f) ground-truth samples obtained by reward-weighted resampling of prior samples. Unlike the other baselines (a)-(d) whose samples concentrate in partial high-reward regions due to their empirical regularizers, our hyperspherical update faithfully reproduces the ground-truth distribution, without artifacts such as samples far along the reward bands in the naive cartesian update.

Text-to-Image Generation. We evaluate initial noise alignment for pretrained text-to-image generation with SDXL-turbo (Sauer et al., 2024) and DMD2 (Yin et al., 2024a) on five tasks: ImageReward (Xu et al., 2024), PickScore (Kirstain et al., 2023), GenEval (Ghosh et al., 2023), Query object counting (Qian et al., 2025), and OCR (Cui et al., 2025). The generator parameters remain fixed, and alignment is performed by updating the initial noise. ImageReward and PickScore provide differentiable preference rewards, while GenEval, Counting, and OCR evaluate compositional alignment, object counting, and text rendering through black-box reward evaluations. We additionally report aesthetic score (Schuhmann, 2022) as a separate evaluation metric to assess how task-specific alignment affects visual quality. See Appendix B.3 for the detailed settings on the reward functions and utilized text prompt datasets.

We include uncontrolled generation and Best-of-K, which selects the best sample among K independently drawn initial noises. The zeroth-order baselines comprise ZO Langevin, exponential-map updates with a ZO gradient estimator, and Natural Evolution Strategies (NES) (Wierstra et al., 2014). For differentiable rewards, we also include ReNO (Eyring et al., 2024), ORIGEN (Min et al., 2026), and MoNO (Shi et al., 2026) as baselines that utilize backpropagation through the generator and reward model.

Table 1 shows that ZeNOVA improves over uncontrolled generation, Best-of-K, and every reported zeroth-order optimization baseline on all five task metrics for both generative models. Specifically, in black-box reward tasks, ZeNOVA achieves performance gains of $+0.046$ on GenEval and $+0.088$

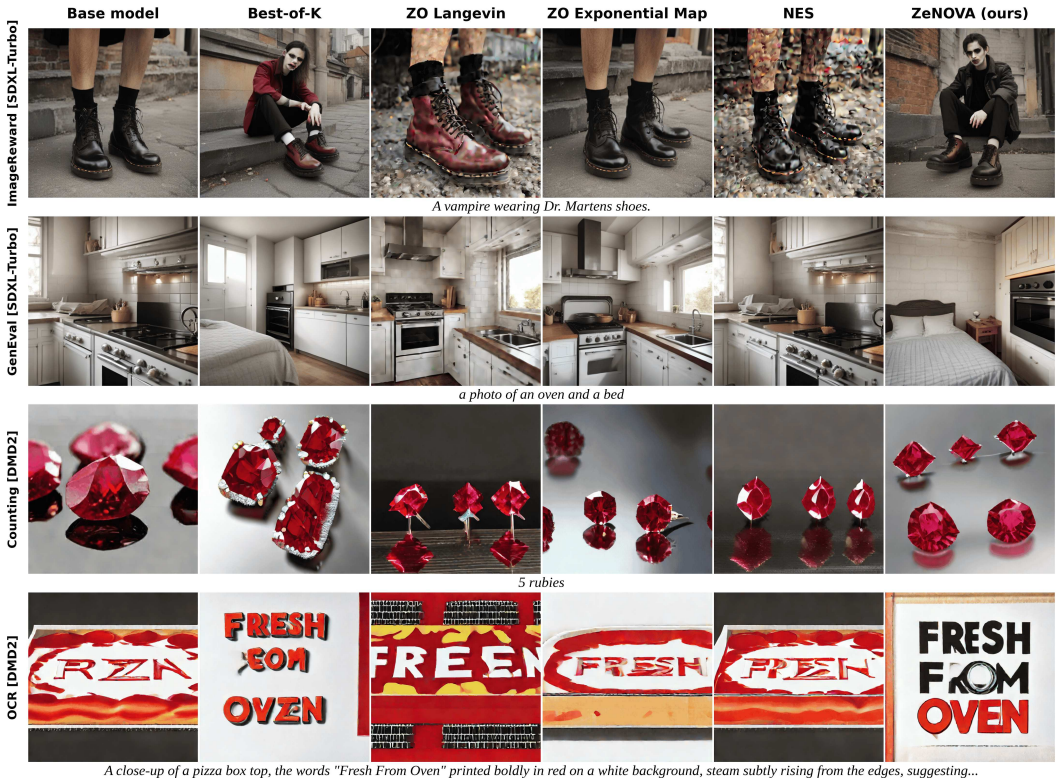

Figure 4: Qualitative image results from SDXL-Turbo and DMD2.

Table 1: Initial noise reward alignment performance on text-to-image models. **Bold**: best performance, "–": Not available.

	ImageReward			PickScore			GenEval		Counting		OCR	
Method	Reward ↑	Aes. ↑	Val. ↑	Reward ↑	Aes. ↑	Val. ↑	Reward ↑	Aes. ↑	Reward ↓	Aes. ↑	Reward ↑	Aes. ↑
SDXL-Turbo												
base model	1.0505	6.0282	22.8809	21.5378	5.4020	0.5321	0.5375	5.3853	28.375	5.6965	0.1368	5.4256
Best-of-K	1.6932	6.1148	23.2829	23.1925	5.5747	0.8949	0.7232	5.3717	2.9	5.7096	0.4784	5.3303
ReNO	1.7378	6.0395	22.9585	**25.1031**	5.5721	0.7755	–	–	–	–	–	–
ORIGEN	**1.8198**	5.1640	21.1789	24.2321	5.5748	0.9286	–	–	–	–	–	–
MoNO	1.6781	6.0558	23.0221	24.6250	5.5579	0.7675	–	–	–	–	–	–
ZO Langevin	1.3968	5.8129	22.8452	22.2734	5.5092	0.7402	0.6875	5.4123	6.025	5.7078	0.3823	5.3439
ZO Exponential Map	1.6452	6.0507	23.2201	22.6230	5.5437	0.7880	0.6673	5.3912	3.325	5.6777	0.4239	**5.4677**
NES	1.6354	6.0795	23.0850	22.3546	5.5050	0.7583	0.6613	5.4006	5.8	5.6811	0.4195	5.3516
ZeNOVA (ours)	1.8078	**6.1279**	**23.3584**	23.4564	**5.5957**	**1.0329**	**0.7332**	**5.4278**	**0.55**	**5.7222**	**0.5114**	5.3864
DMD2												
base model	0.5709	5.4908	21.4339	20.4482	5.3677	0.3438	0.4547	5.1387	14.575	5.1468	0.0921	**5.3661**
Best-of-K	1.6456	5.6970	22.1676	22.0510	5.5511	0.8915	0.7224	5.1387	1.85	5.1557	0.4646	5.0526
ReNO	1.3789	5.4232	20.9871	**23.4025**	5.4159	0.5440	–	–	–	–	–	–
ORIGEN	1.2708	4.9731	19.6959	21.8586	5.3021	0.2489	–	–	–	–	–	–
MoNO	1.2816	5.3627	21.0667	23.1507	5.4140	0.3623	–	–	–	–	–	–
ZO Langevin	1.4898	5.4877	21.8468	21.8061	5.4043	0.7811	0.6597	5.1388	9.775	4.6922	0.3961	5.3055
ZO Exponential Map	1.4100	5.5935	21.9041	21.9611	5.5548	0.6782	0.6383	5.1436	1.35	**5.2610**	0.2351	5.3417
NES	1.2996	5.5763	21.7994	21.6149	5.5007	0.6193	0.6251	5.1531	1.975	5.0784	0.2073	5.3637
ZeNOVA (ours)	**1.7683**	**5.7210**	**22.4022**	22.6398	**5.6130**	**0.9187**	**0.7275**	**5.1703**	**0.075**	5.1902	**0.5237**	5.1575

on OCR from the best zeroth-order baseline, and reduces MAE of counting by around 40%. These results show improvements over the alternatives on objectives that do not provide gradients. Note that on ImageReward and PickScore where the reward gradient is accessible, gradient-based algorithms of ReNO and ORIGEN often attain higher reward than ours. Here, leveraging the fact that ImageReward and PickScore both measure the same aspect of human preference, we evaluated both runs with the reward not employed during optimization as a validation metric. We observe that gradient-based methods are prone to reward hacking, allowing our method to achieve the highest validation metric.

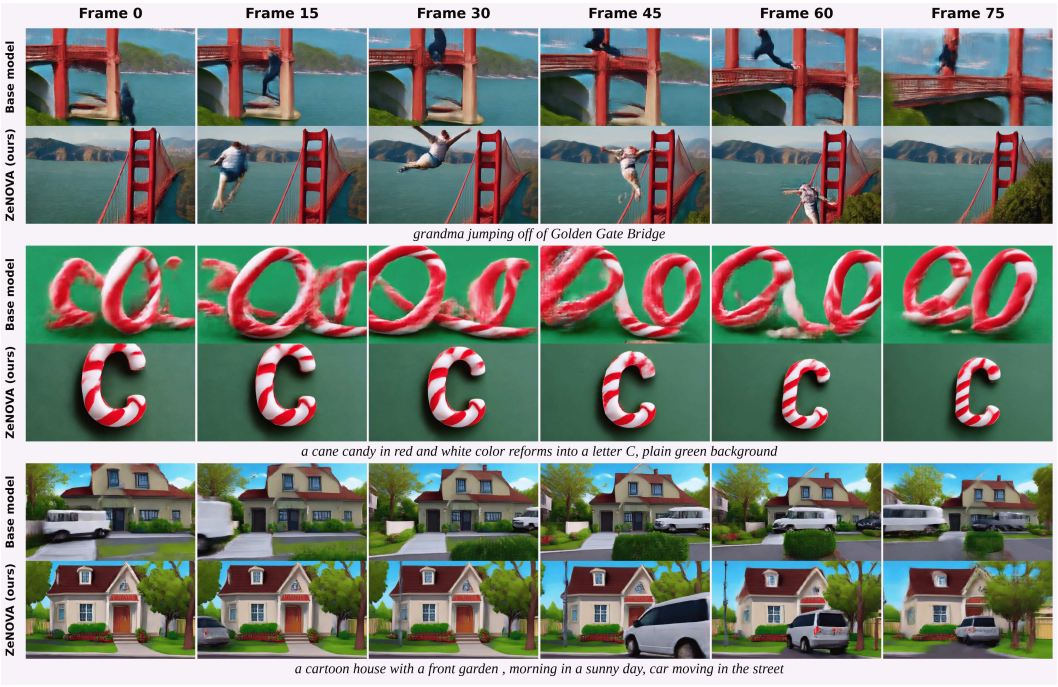

Figure 5: Qualitative video results from rCM with ZeNOVA, optimizing VideoAlign scores.

In terms of image quality, ZeNOVA achieved the highest Aesthetic score across various scenarios, while some unstable zeroth-order methods and first-order methods prone to reward hacking exhibited significant drops. While this is partly because some optimized rewards are already associated with image quality, but the results in other cases suggest that ZeNOVA properly preserves the prior distribution of the original generative model's input.

Text-to-Video Generation. Due to the increased data dimensionality introduced by the additional temporal axis, a growing interest for models that generate video frames or chunks in a few steps was emerged. We further applied ZeNOVA to a few-step distilled text-to-video flow-matching model rCM (Zheng et al., 2026b). Following Xue et al. (2025), we employed three VideoAlign (Liu et al., 2026) benchmark score for Visual Quality (VQ), Motion Quality (MQ), and Text Alignment (TA). Text prompts from VidProM (Wang & Yang, 2024) dataset were used. More details can be found in Appendix B.4.

Table 2 and Figure 5 show that ZeNOVA improves over uncontrolled generation in all four optimization settings. These gains demonstrate the scalability of the proposed initial-noise alignment method for a task with much more dimensionality and complexity, supporting both individual reward objectives and their combined optimization without modifying the pretrained generator.

Table 2: Quantitative performance on text-to-video models. **Bold**: Best performance.

	VideoAlign			
Method	VQ	MQ	TA	All
base model	3.4854	1.3684	2.1731	6.1997
ZeNOVA	**6.4464**	**3.3124**	**3.9088**	**10.8146**

5 Discussion

Effect of Particle Count per Iteration. Zeroth-order methods require reward evaluations for a fixed number(N) of particles per iteration to estimate each method's intended update direction of the initial noise; as N increases, this estimate approaches the theoretical ground truth and the performance converges. In practice, since the number of available reward evaluations is finite, the performance behavior with a small number of particles becomes a critical factor in the practicality of the algorithm.

Figure 6 compares ZeNOVA with Cartesian Langevin, exponential-map updates, and NES with ImageReward on HPDv2 prompts, holding the number of iterations fixed at $K = 25$. Reward

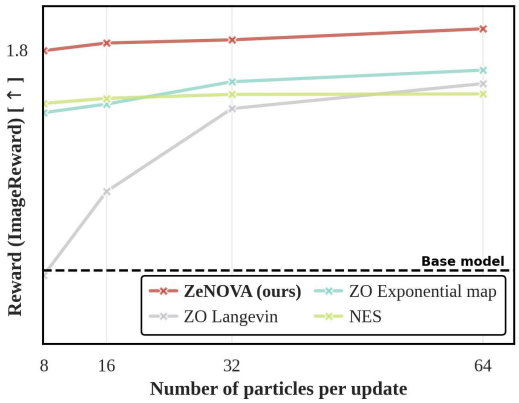

Figure 6: Effect of particle count per iteration for zeroth-order methods.

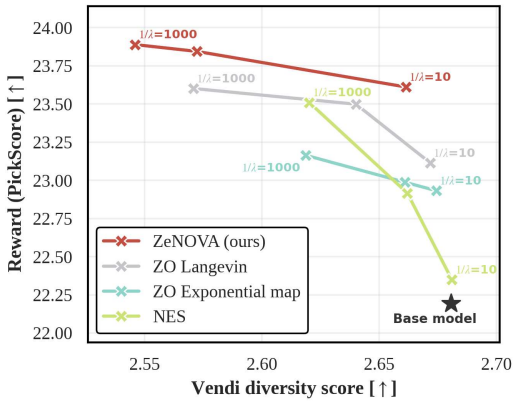

Figure 7: Pareto curve between PickScore and Vendi diversity at different guidance scales.

increases with N for every evaluated method, while ZeNOVA achieves the highest reward at every reported particle count and exhibits a small degradation as the particle budget decreases; reducing N from 64 to 8 lowers ZeNOVA's ImageReward by only 0.0743. Notably, ZeNOVA with only 8 particles outperforms all three baselines with 64 particles. These results demonstrate that ZeNOVA preserves strong alignment under limited reward-query budgets and are consistent with the stabilizing role of hyperspherical updates under noisy zeroth-order guidance.

Reward-Diversity Trade-off. ZeNOVA and the baselines have a guidance scale hyperparameter λ to control the trade-off between prior preservation and reward. Figure 7 plots the Pareto curve between PickScore reward and Vendi diversity (Friedman & Dieng, 2023) of 64 images per prompt, for different methods at three guidance scales. For all methods, increasing $1/\lambda$ improves the average reward while diversity decreases, indicating that stronger guidance yields reward gains by making the output distribution focus on a narrow high-reward region. Meanwhile, our approach achieves a dominant Pareto front located at the most upright of the plot, indicating higher diversity in its output for a fixed level of reward alignment.

Ablation Study. We've conducted an ablation study on more detailed design choices of the proposed ZeNOVA in Table 3 to analyze and emphasize their role. It is shown that all three components contributed to the method's performance and stability. We note that introducing the annealed soft value function yielded marginal or no performance improvements on ImageReward and PickScore, but led to a substantial improvement on OCR reward; the reward landscapes of the former two rewards are continuous with many high-reward modes, allowing sufficiently high rewards to be achieved using only local reward gradient information around the initial point. In the more challenging black-box reward of OCR, the value function allows the method to access a more global and meaningful reward gradient signal.

Table 3: Ablation study of the components of ZeNOVA.

Variants	ImageReward	PickScore	OCR
ZeNOVA	**1.8078**	**23.4564**	**0.5114**
- MH jumping	1.7262	23.1867	0.4412
- Value Annealing	1.6984	23.2960	0.3756
- Hyperspherical update	1.3968	22.2734	0.3823

6 Conclusion

We introduced ZeNOVA, a stable and efficient initial-noise optimization framework for generative models through initial-noise optimization. Guided by a reward-tilted sampling objective, ZeNOVA combines annealed soft-value guidance to extract useful signals from irregular rewards, where several crucial challenges are adequately resolved to enable stability and global exploration. Experimental results on various scenarios demonstrate that ZeNOVA consistently achieves higher rewards than the zeroth-order baselines and retains strong performance with less computational cost.

AI USE STATEMENT

In this work, we used generative AI tools to [assist with translation]. We have not used generative AI tools to [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, interpret results], and [generate synthetic data sets, support qualitative and thematic data analysis] are not applicable to this work. Additionally, we used generative AI tools to [create or edit software code, edit a research paper to improve readability]. We have reviewed all AI-assisted work. [All authors reviewed and refined LLM-generated drafts and code snippets before using them]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work on image generative models, and we acknowledge that generative models can produce misleading, biased, or not-safe-for-work visual content, which may amplify misinformation, prejudice, or privacy violations.

REPRODUCIBILITY STATEMENT

For theoretical results, clear explanations of any assumptions and a complete proof of the claims are included in the Appendix. Any datasets used in the experiments are open access; a complete description of the data preparation are provided in the Appendix.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xQBRrtQM8u.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MXY0qsGgeO.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=8OTPepXzeh.

Wendell H Fleming and Raymond W Rishel. *Deterministic and stochastic optimal control*, volume 1. Springer Science & Business Media, 2012.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Wbr51vK331.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. TempFlow-GRPO: When timing matters for GRPO in flow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7mCo3R3Wyn.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023.

Yujia Huang, Adishree Ghatare, Yuanzhe Liu, Ziniu Hu, Qinsheng Zhang, Chandramouli Shama Sastry, Siddharth Gururani, Sageev Oore, and Yisong Yue. Symbolic music generation with non-differentiable rule guided diffusion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19772–19797. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24g.html.

Cristian Jensen, Luca Schaufelberger, Riccardo De Santi, Kjell Jorner, and Andreas Krause. Value matching: Scalable and gradient-free reward-guided flow adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 144926–144958, 2026.

H. Kappen. Path integrals and symmetry breaking for optimal control theory. *J Stat Mech- Theory E*, 2005, 05 2005. doi: 10.1088/1742-5468/2005/11/P11011.

Jaemin Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Free2guide: Training-free text-to-video alignment using image lvlm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17920–17929, October 2025.

Jeongsol Kim, Hongeun Kim, Jian Wang, and Jong Chul Ye. Gradient-free noise optimization for reward alignment in generative models. *arXiv preprint arXiv:2605.11347*, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=oCBKGw5HNf.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026.

Qingming Liu, Zhen Liu, Dinghuai Zhang, and Kui Jia. Nabla-R2D3: Effective and efficient 3D diffusion alignment with 2D rewards. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=Dk2qprCnu8.

Zhen Liu, Tim Z. Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed GFlownets. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=Aye5wL6TCn.

Yunhong Min, Daehyeon Choi, Kyeongmin Yeo, Jihyun Lee, and Minhyuk Sung. ORIGEN: Zero-shot 3d orientation grounding in text-to-image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=CKOyN2HShu.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pp. 38–50. Springer, 2003.

Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau, Xiaopeng Hong, and Michael P Pound. T2icount: Enhancing cross-modal understanding for zero-shot counting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25336–25345. IEEE, 2025.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. GitHub repository.

Qitan Shi, Cheng Jin, Ziyuan Liu, and Yuantao Gu. Manifold-constrained noise optimization for diverse diffusion sampling. *arXiv preprint arXiv:2607.23937*, 2026.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023b.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024a.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024b.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

TaeHoon Yoon, Yunhong Min, Kyeongmin Yeo, and Minhyuk Sung. ψ-sampler: Initial particle sampling for SMC-based inference-time reward alignment in score models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=slVqJAI5sT.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Dinghuai Zhang, Yizhe Zhang, Jiatao Gu, Ruixiang Zhang, Joshua M. Susskind, Navdeep Jaitly, and Shuangfei Zhai. Improving GFlownets for text-to-image diffusion alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=XDbY3qhM42.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026a.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026b.

A Theoretical Results

A.1 Proof of Proposition 1

Proposition 1 (Interpolation of target distribution)**.** *Let reward-tilted distribution $p^\star$ be defined as Eq. (4) and soft-value-tilted distribution $p^\star_\sigma$ as Eq. (10). Then $p^\star_\sigma$ is equal to $p^\star$ passed through a Gaussian kernel of Eq. (8):*

$$p^\star_\sigma(\boldsymbol{z}) \propto \int q_\sigma(\boldsymbol{z} \mid \boldsymbol{z}')\, p^\star(\boldsymbol{z}')\, d\boldsymbol{z}'. \tag{11}$$

Proof. Starting from the defition of the soft value function,

$$\begin{aligned} V_\sigma(\boldsymbol{z}) &:= \lambda \log \mathbb{E}_{\boldsymbol{z}' \sim q_\sigma(\cdot|\boldsymbol{z})} \left[e^{R(\boldsymbol{z}')/\lambda} \right] \\ \Leftrightarrow \exp\left(\frac{V_\sigma(\boldsymbol{z})}{\lambda}\right) &= \exp\Big(\log \mathbb{E}_{\boldsymbol{z}' \sim q_\sigma(\cdot|\boldsymbol{z})} \left[e^{R(\boldsymbol{z}')/\lambda} \right]\Big) \\ &= \mathbb{E}_{\boldsymbol{z}' \sim q_\sigma(\cdot|\boldsymbol{z})} \left[e^{R(\boldsymbol{z}')/\lambda} \right] \\ &= \int q_\sigma(\boldsymbol{z}' \mid \boldsymbol{z}) e^{R(\boldsymbol{z}')/\lambda}\, d\boldsymbol{z}' \end{aligned} \tag{19}$$

Substituting this into (10) gives

$$\begin{aligned} p^\star_\sigma(\boldsymbol{z}) &\propto p_0(\boldsymbol{z}) \exp\left(\frac{V_\sigma(\boldsymbol{z})}{\lambda}\right) \\ &= \int p_0(\boldsymbol{z}) q_\sigma(\boldsymbol{z}' \mid \boldsymbol{z}) e^{R(\boldsymbol{z}')/\lambda}\, d\boldsymbol{z}' \end{aligned} \tag{20}$$

Here, since OU kernel is reversible with respect to p_0, we have $p_0(\boldsymbol{z}) q_\sigma(\boldsymbol{z}' \mid \boldsymbol{z}) = p_0(\boldsymbol{z}') q_\sigma(\boldsymbol{z} \mid \boldsymbol{z}')$. Thus,

$$\begin{aligned} p^\star_\sigma(\boldsymbol{z}) &\propto \int p_0(\boldsymbol{z}') q_\sigma(\boldsymbol{z} \mid \boldsymbol{z}') e^{R(\boldsymbol{z}')/\lambda}\, d\boldsymbol{z}' \\ &\propto \int q_\sigma(\boldsymbol{z} \mid \boldsymbol{z}') p^\star(\boldsymbol{z}')\, d\boldsymbol{z}'. \end{aligned} \tag{21}$$

□

A.2 Proof of Proposition 2

Proposition 2 (Metropolis-Hastings algorithm for reward-tilted distribtuion)**.** *The Metropolis–Hastings chain with proposal $\boldsymbol{z}' \sim p_0$ and acceptance probability $\alpha(\boldsymbol{z}, \boldsymbol{z}') = \min\{1, \exp((R(\boldsymbol{z}') - R(\boldsymbol{z}))/\lambda)\}$ converges to the reward-tilted distribution $p^\star(\boldsymbol{z}) \propto p_0(\boldsymbol{z}) e^{R(\boldsymbol{z})/\lambda}$.*

Proof. To construct a Markov chain for sampling from an unnormalized density $\widetilde{p}(\boldsymbol{z})$, Metropolis-Hastings algorithm draws a candidate $\boldsymbol{z}' \sim q(\cdot \mid \boldsymbol{z})$ and accepts it with probability

$$\alpha_{\mathrm{MH}}(\boldsymbol{z}, \boldsymbol{z}') = \min\left\{1, \frac{\widetilde{p}(\boldsymbol{z}') q(\boldsymbol{z} \mid \boldsymbol{z}')}{\widetilde{p}(\boldsymbol{z}) q(\boldsymbol{z}' \mid \boldsymbol{z})}\right\}, \tag{22}$$

retaining the current state if the proposal is rejected.

$$p^\star(\boldsymbol{z}) \propto p_0(\boldsymbol{z}) \exp\left(\frac{R(\boldsymbol{z})}{\lambda}\right), \tag{23}$$

To apply Metropolis-Hastings to an unnormalized reward-tilted distribution $\widetilde{p}(\boldsymbol{z}) = p_0(\boldsymbol{z}) e^{R(\boldsymbol{z})/\lambda}$, on can choose the independent Gaussian prior proposal and reward-based acceptance probability as

$$q(\boldsymbol{z}' \mid \boldsymbol{z}) = p_0(\boldsymbol{z}') = \mathcal{N}(\boldsymbol{z}'; 0, \mathbf{I}). \tag{24}$$

By this selection, Metropolis-Hastings acceptance ratio for $\alpha(\boldsymbol{z}, \boldsymbol{z}')$ becomes

$$\frac{\widetilde{p}(\boldsymbol{z}') q(\boldsymbol{z} \mid \boldsymbol{z}')}{\widetilde{p}(\boldsymbol{z}) q(\boldsymbol{z}' \mid \boldsymbol{z})} = \frac{p_0(\boldsymbol{z}') e^{R(\boldsymbol{z}')/\lambda} p_0(\boldsymbol{z})}{p_0(\boldsymbol{z}) e^{R(\boldsymbol{z})/\lambda} p_0(\boldsymbol{z}')} = \exp\left(\frac{R(\boldsymbol{z}') - R(\boldsymbol{z})}{\lambda}\right), \tag{25}$$

which reconstructs the claimed algorithm to generate samples from $p^\star(\boldsymbol{z}) \propto p_0(\boldsymbol{z}) e^{R(\boldsymbol{z})/\lambda}$. □

A.3 Derivation of Zeroth-order Value Gradient Estimator

Stochastic Optimal Control theory describes the optimal control term injected into a given reference process to maximize an entropy-regularized objective and relates it to the soft value function. Specifically, in our OU reference process over a timestep $t = 0$ to $t = h$,

$$d\boldsymbol{z}_s = -\boldsymbol{z}_s\, ds + \sqrt{2}\, d\boldsymbol{w}_s, \qquad \boldsymbol{z}_0 = \boldsymbol{z}. \tag{26}$$

the optimal control $\boldsymbol{u}^\star(\boldsymbol{z}, 0)$ is defined to maximize the following objectives under the controlled process.

$$\max \mathbb{E}_u \left[R(\boldsymbol{z}_h^u) - \frac{\lambda}{2} \int_0^h \|\boldsymbol{u}_s\|^2 ds \,\middle|\, \boldsymbol{z}_0^u = \boldsymbol{z} \right] \tag{27}$$

$$\text{s.t. } d\boldsymbol{z}_s^u = -\boldsymbol{z}_s^u ds + \sqrt{2}\big(\boldsymbol{u}_s\, ds + d\boldsymbol{w}_s\big), \boldsymbol{z}_0^u = \boldsymbol{z} \tag{28}$$

With $\sigma^2 := 1 - e^{-2h}$, Hamilton-Jacobi-Bellman equation (Fleming & Rishel, 2012) states that the optimal control at $t = 0$ $\boldsymbol{u}^\star(\boldsymbol{z}, 0)$ satisfies

$$\boldsymbol{u}^\star(\boldsymbol{z}, 0) = \frac{\sqrt{2}}{\lambda} \nabla_{\boldsymbol{z}} V_\sigma(\boldsymbol{z}). \tag{29}$$

Furthermore, the path-integral control identity (Kappen, 2005) gives

$$\boldsymbol{u}^\star(\boldsymbol{z}, 0) = \lim_{dt \to 0} \frac{\mathbb{E}_{\text{ref}}[e^{R(\boldsymbol{z}_h)/\lambda} d\boldsymbol{w}_0 \mid \boldsymbol{z}_0 = \boldsymbol{z}]}{dt\, \mathbb{E}_{\text{ref}}[e^{R(\boldsymbol{z}_h)/\lambda} \mid \boldsymbol{z}_0 = \boldsymbol{z}]}. \tag{30}$$

Here, $d\boldsymbol{w}_0$ and dt denote the injected Brownian motion of a single trajectory at $t = 0$ and its corresponding timestep increment. All expectations in this identity are under the uncontrolled reference process. The terminal reward is evaluated at $\boldsymbol{z}_h$.

Now, we reformulate Eq. (30) into terms that are more easily tractable. First, we can get the exact OU solution for $\boldsymbol{z}_h$ as

$$\boldsymbol{z}_h = \sqrt{1 - \sigma^2}\boldsymbol{z} + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{31}$$

Our goal is to rewrite the numerator of Eq. (30), which contains both the terminal reward and the initial Brownian increment $d\boldsymbol{w}_0$, using only the terminal Gaussian sample $\boldsymbol{z}_h$. Here, note that the reward weight depends only on $\boldsymbol{z}_h$. Therefore, applying the tower property to the numerator of Eq. (30) yields

$$\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda} d\boldsymbol{w}_0] = \mathbb{E}\Big[\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda} d\boldsymbol{w}_0 \mid \boldsymbol{z}_h]\Big] \tag{32}$$

$$= \mathbb{E}\Big[e^{R(\boldsymbol{z}_h)/\lambda} \mathbb{E}[d\boldsymbol{w}_0 \mid \boldsymbol{z}_h]\Big] \tag{33}$$

For $\mathbb{E}[d\boldsymbol{w}_0 \mid \boldsymbol{z}_h]$, we use the conditional mean formula jointly Gaussian variables:

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X] + \mathrm{Cov}(X, Y)\, \mathrm{Var}(Y)^{-1}(Y - \mathbb{E}[Y]). \tag{34}$$

Applying this formula with $X = d\boldsymbol{w}_0$ and $Y = \boldsymbol{z}_h$, and using $\mathbb{E}[d\boldsymbol{w}_0] = 0$ and $\mathbb{E}[\boldsymbol{z}_h] = \sqrt{1 - \sigma^2}\boldsymbol{z}$, gives

$$\mathbb{E}[d\boldsymbol{w}_0 \mid \boldsymbol{z}_h] = 0 + \mathrm{Cov}(d\boldsymbol{w}_0, \boldsymbol{z}_h)\, \mathrm{Var}(\boldsymbol{z}_h)^{-1}(\boldsymbol{z}_h - \sqrt{1 - \sigma^2}\boldsymbol{z}) \tag{35}$$

Here, we have a stochastic-integral representation of the terminal state of the OU process as $\boldsymbol{z}_h = e^{-h}\boldsymbol{z} + \sqrt{2}\int_0^h e^{-(h-s)} d\boldsymbol{w}_s$. Both $d\boldsymbol{w}_0$ and $\boldsymbol{z}_h$ are thus linear functions of Gaussian noise and are jointly Gaussian. The initial noise contributes to $\boldsymbol{z}_h$ through the interval $[0, dt]$, while later increments are independent of it. Their cross-covariance and terminal covariance are consequently

$$\begin{aligned} \mathrm{Cov}(d\boldsymbol{w}_0, \boldsymbol{z}_h) &= \sqrt{2} \int_0^{dt} e^{-(h-s)} ds \mathbf{I} = \sqrt{2}\sqrt{1 - \sigma^2}(e^{dt} - 1)\mathbf{I}, \\ \mathrm{Var}(\boldsymbol{z}_h) &= \sigma^2 \mathbf{I}. \end{aligned} \tag{36}$$

Now we can complete the formula:

$$\mathbb{E}[d\boldsymbol{w}_0 \mid \boldsymbol{z}_h] = \frac{\sqrt{2}\sqrt{1-\sigma^2}(e^{dt}-1)}{\sigma^2}(\boldsymbol{z}_h - \sqrt{1-\sigma^2}\boldsymbol{z}) \tag{37}$$

$$\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda} d\boldsymbol{w}_0] = \frac{\sqrt{2}\sqrt{1-\sigma^2}(e^{dt}-1)}{\sigma^2}\mathbb{E}\left[e^{R(\boldsymbol{z}_h)/\lambda}(\boldsymbol{z}_h - \sqrt{1-\sigma^2}\boldsymbol{z})\right]. \tag{38}$$

Substituting it into Eq. (30) gives

$$\begin{aligned}\boldsymbol{u}^\star(\boldsymbol{z}, 0) &= \lim_{dt\to 0} \frac{\sqrt{2}\sqrt{1-\sigma^2}}{\sigma^2}\frac{e^{dt}-1}{dt}\frac{\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda}(\boldsymbol{z}_h - \sqrt{1-\sigma^2}\boldsymbol{z})]}{\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda}]} \\ &= \frac{\sqrt{2}\sqrt{1-\sigma^2}}{\sigma^2}\frac{\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda}(\boldsymbol{z}_h - \sqrt{1-\sigma^2}\boldsymbol{z})]}{\mathbb{E}[e^{R(\boldsymbol{z}_h)/\lambda}]}\end{aligned} \tag{39}$$

Finally, we substitute $\boldsymbol{z}_h = \sqrt{1-\sigma^2}\boldsymbol{z} + \sigma\boldsymbol{\epsilon}$:

$$\boldsymbol{u}^\star(\boldsymbol{z}, 0) = \frac{\sqrt{2}\sqrt{1-\sigma^2}}{\sigma}\frac{\mathbb{E}_{\boldsymbol{\epsilon}}[e^{R(\sqrt{1-\sigma^2}\boldsymbol{z}+\sigma\boldsymbol{\epsilon})/\lambda}\boldsymbol{\epsilon}]}{\mathbb{E}_{\boldsymbol{\epsilon}}[e^{R(\sqrt{1-\sigma^2}\boldsymbol{z}+\sigma\boldsymbol{\epsilon})/\lambda}]}. \tag{40}$$

The resulting expression depends only on a Gaussian perturbation and its terminal reward, so no full OU trajectory or initial Brownian increment needs to be sampled. Multiplying by $\lambda/\sqrt{2}$ recovers (12).

A.4 Derivation of Hyperspherical Langevin Dynamics

We decompose a sample $\boldsymbol{z}$ from a standard Gaussian as its norm ρ and its unit direction vector $\boldsymbol{u}$.

$$\boldsymbol{z} = \rho\boldsymbol{u}, \qquad \rho := \|\boldsymbol{z}\|_2 > 0, \qquad \boldsymbol{u} \in \mathbb{S}^{d-1} \tag{41}$$

When we convert the soft-value-tilted distribtuion with ρ and $\boldsymbol{u}$ as well, Eq. (10) becomes

$$\widetilde{p}^\star_\sigma(\rho, \boldsymbol{u}) \propto \rho^{d-1}\exp\left(-\frac{\rho^2}{2} + \frac{V_\sigma(\rho\boldsymbol{u})}{\lambda}\right). \tag{42}$$

This expression exposes two qualitatively different directions: the angular component moves along the Gaussian shell, whereas the radial component controls how far the sample lies from the origin. Rather than allowing a noisy d-dimensional update to perturb both components implicitly, we construct Langevin dynamics for them explicitly.

Angular Langevin on the sphere. With a fixed radial ρ, the score function of the angular conditional distribution induced by (42) is

$$\nabla_{\mathbb{S}} \log \widetilde{p}^\star_\sigma(\rho, \boldsymbol{u}) = \mathbf{P}^\perp_{\boldsymbol{u}}\left(\frac{\rho}{\lambda}\boldsymbol{g}_\sigma(\rho\boldsymbol{u})\right), \tag{43}$$

where $\boldsymbol{g}_\sigma(\boldsymbol{z}) := \nabla_{\boldsymbol{z}} V_\sigma(\boldsymbol{z})$ and $\mathbf{P}^\perp_{\boldsymbol{u}} := \mathbf{I} - \boldsymbol{u}\boldsymbol{u}^\top$ denotes the projection onto the tangent space of $\boldsymbol{u}$. Since $\boldsymbol{u}$ must lie on a unit sphere, we use intrinsic spherical Langevin dynamics

$$d\boldsymbol{u}_s = \nabla_{\mathbb{S}} \log \widetilde{p}^\star_\sigma(\rho_s, \boldsymbol{u}_s)\, ds + \sqrt{2}\, d\boldsymbol{w}^{\mathbb{S}}_s, \tag{44}$$

where $\boldsymbol{w}^{\mathbb{S}}_s$ denotes Brownian motion also lying on the unit sphere. A practical Euler step implementation of this with step size η_A can be a tangent Gaussian followed by the spherical exponential map (Lee, 2018):

$$\boldsymbol{v}_k = \eta_A \frac{\rho_k}{\lambda}\mathbf{P}^\perp_{\boldsymbol{u}_k}\widehat{\boldsymbol{g}}_{\sigma_k}(\boldsymbol{z}_k) + \sqrt{2\eta_A}\,\mathbf{P}^\perp_{\boldsymbol{u}_k}\boldsymbol{\xi}_k, \tag{45}$$

$$\boldsymbol{u}_{k+1} = \mathrm{Exp}_{\boldsymbol{u}_k}(\boldsymbol{v}_k), \qquad \boldsymbol{\xi}_k \sim \mathcal{N}(0, \mathbf{I}), \tag{46}$$

where $\mathrm{Exp}_{\boldsymbol{u}}(\boldsymbol{v})$ implements the rotation of $\boldsymbol{u}$ to the origin toward the direction of $\boldsymbol{v}$ as

$$\mathrm{Exp}_{\boldsymbol{u}}(\boldsymbol{v}) = \cos(\|\boldsymbol{v}\|_2)\boldsymbol{u} + \sin(\|\boldsymbol{v}\|_2)\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}. \tag{47}$$

with the continuous extension $\mathrm{Exp}_{\boldsymbol{u}}(\mathbf{0}) = \boldsymbol{u}$. By construction, both the estimated reward drift and the Langevin noise remain tangent to the sphere.

Radial Langevin. For soft-value-tilted distribution, with a fixed angular state $\boldsymbol{u}$, the score function of the radial conditional induced by (42) is

$$\frac{d}{d\rho} \log \widetilde{p}_{\sigma}^{\star}(\rho \mid \boldsymbol{u}) = \frac{d-1}{\rho} - \rho + \frac{1}{\lambda} \boldsymbol{u}^{\top} \boldsymbol{g}_{\sigma}(\rho \boldsymbol{u}). \tag{48}$$

Note that the first two terms are equal to the score function of the standard Gaussian's radial distribution $\rho \sim \chi_d$. We therefore directly apply one-dimensional Langevin dynamics to the radius,

$$d\rho_s = \left(\frac{d-1}{\rho_s} - \rho_s + \frac{1}{\lambda} \boldsymbol{u}_s^{\top} \boldsymbol{g}_{\sigma}(\rho_s \boldsymbol{u}_s) \right) ds + \sqrt{2}\, dw_s. \tag{49}$$

A practical Euler step with step size η_R is

$$\rho_{k+1} = \rho_k + \eta_R \left(\frac{d-1}{\rho_k} - \rho_k + \frac{1}{\lambda} \boldsymbol{u}_k^{\top} \boldsymbol{g}_{\sigma_k}(\boldsymbol{z}_k) \right) + \sqrt{2\eta_R}\, \xi_k, \qquad \xi_k \sim \mathcal{N}(0,1). \tag{50}$$

The first two terms provide the exact Gaussian radial bias, while the scalar projection $\boldsymbol{u}^{\top} \widehat{\boldsymbol{g}}_{\sigma}$ captures the reward-induced radial preference. Thus the radius remains stochastic and reward-adaptive, unlike several empirical regularizations like explicit norm constraints.

B Implementation Details

B.1 Detailed Algorithm of ZeNOVA

Algorithm 1 describes a detailed algorithm of the proposed ZeNOVA with Annealed Soft Value, Hyperspherical Langevin, and MH Jumping. Note that the reciprocal of the temperature $1/\lambda$ can be considered as an effective reward guidance scale, making the method more focused on reward maximization than prior preservation.

B.2 Implementation of Zeroth-Order Baselines

ZO Langevin. We substitute the zeroth-order gradient estimator $\widehat{\boldsymbol{g}}(\boldsymbol{z}_k)$ in Eq. (7) into Eq. (5) and apply an Euler–Maruyama step of size η:

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \eta \left(-\boldsymbol{z}_k + \frac{1}{\lambda} \widehat{\boldsymbol{g}}(\boldsymbol{z}_k) \right) + \sqrt{2\eta}\, \boldsymbol{\xi}_k, \qquad \boldsymbol{\xi}_k \sim \mathcal{N}(0, \mathbf{I}). \tag{51}$$

ZO Exponential Map. Let $\rho_k = \|\boldsymbol{z}_k\|_2$ and $\boldsymbol{u}_k = \boldsymbol{z}_k / \rho_k$. We project the same estimator onto the tangent space and update the direction through the spherical exponential map with step size η:

$$\boldsymbol{g}_k^{\tan} = (\mathbf{I} - \boldsymbol{u}_k \boldsymbol{u}_k^{\top}) \widehat{\boldsymbol{g}}(\boldsymbol{z}_k), \qquad \boldsymbol{u}_{k+1} = \mathrm{Exp}_{\boldsymbol{u}_k} \left(\frac{\eta \rho_k}{\lambda} \boldsymbol{g}_k^{\tan} \right), \qquad \boldsymbol{z}_{k+1} = \rho_k \boldsymbol{u}_{k+1}. \tag{52}$$

Natural Evolution Strategies. NES (Wierstra et al., 2014) maximizes the expected reward $J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim \pi_{\boldsymbol{\theta}}}[R(\boldsymbol{z})]$ by updating the parameters of a search distribution along its natural gradient:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \mathbf{F}(\boldsymbol{\theta}_k)^{-1} \widehat{\nabla}_{\boldsymbol{\theta}} J, \qquad \widehat{\nabla}_{\boldsymbol{\theta}} J = \frac{1}{N} \sum_{n=1}^{N} (R(\boldsymbol{z}^{(n)}) - \bar{R}) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{z}^{(n)}), \tag{53}$$

where $\mathbf{F}$ is the Fisher information matrix. Among the possible search distributions, we use an isotropic Gaussian $\pi_{\boldsymbol{z}_k} = \mathcal{N}(\boldsymbol{z}_k, \sigma^2 \mathbf{I})$ with fixed $\sigma =$ `grad_sigma` and optimize only its mean. For $\boldsymbol{z}_k^{(n)} = \boldsymbol{z}_k + \sigma \boldsymbol{\epsilon}^{(n)}$, the resulting update is

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \eta \sigma^2 \widehat{\boldsymbol{g}}_k, \qquad \widehat{\boldsymbol{g}}_k = \frac{1}{N\sigma} \sum_{n=1}^{N} \left(R(\boldsymbol{z}_k + \sigma \boldsymbol{\epsilon}^{(n)}) - \bar{R} \right) \boldsymbol{\epsilon}^{(n)}, \qquad \boldsymbol{\epsilon}^{(n)} \sim \mathcal{N}(0, \mathbf{I}). \tag{54}$$

Algorithm 1 ZeNOVA with Annealed Soft Value, Hyperspherical Langevin, and MH Jumping

Require: Generator G_θ, reward r, temperature λ, horizons $1 > \sigma_1 > \cdots > \sigma_K > 0$, number of particles N, number of MH jumping N_{MH}, number of iteration K, step sizes η_A, η_R

$\boldsymbol{z}_0 \sim \mathcal{N}(0, \mathbf{I})$; $\rho_0 \leftarrow \|\boldsymbol{z}_0\|_2$; $\boldsymbol{u}_0 \leftarrow \boldsymbol{z}_0/\rho_0$
for $k = 0, \ldots, K-1$ **do**
  for $j = 1, \ldots, N_{\text{MH}}$ **do** ▷ independent MH jumping
    $\boldsymbol{z}' \sim p_0$; $a \sim \text{Unif}(0,1)$
    if $a < \min\{1, \exp((R(\boldsymbol{z}') - R(\boldsymbol{z}))/\lambda)\}$ **then** $\boldsymbol{z}_k \leftarrow \boldsymbol{z}'$
  $\rho_k \leftarrow \|\boldsymbol{z}_k\|_2$; $\boldsymbol{u}_k \leftarrow \boldsymbol{z}_k/\rho_k$
  for $n = 1, \ldots, N$ **do** ▷ parallel reward queries
    $\boldsymbol{\epsilon}_k^{(n)} \sim \mathcal{N}(0, \mathbf{I})$
    $R_k^{(n)} \leftarrow r\Big(G_\theta(\sqrt{1-\sigma_k^2}\boldsymbol{z}_k + \sigma_k \boldsymbol{\epsilon}_k^{(n)})\Big)$
  $w_k^{(n)} \leftarrow \text{softmax}(R_k^{(n)}/\lambda)$
  $\widehat{\boldsymbol{g}}_k \leftarrow \dfrac{\lambda\sqrt{1-\sigma_k^2}}{\sigma_k} \sum_{n=1}^N w_k^{(n)} \boldsymbol{\epsilon}_k^{(n)}$ ▷ soft value gradient
  $g_k^{\text{rad}} \leftarrow \boldsymbol{u}_k^\top \widehat{\boldsymbol{g}}_k$
  $\boldsymbol{g}_k^{\tan} \leftarrow \widehat{\boldsymbol{g}}_k - g_k^{\text{rad}} \boldsymbol{u}_k$
  $\boldsymbol{\xi}_k \sim \mathcal{N}(0, \mathbf{I})$; $\boldsymbol{\xi}_k^{\tan} \leftarrow \boldsymbol{\xi}_k - (\boldsymbol{u}_k^\top \boldsymbol{\xi}_k)\boldsymbol{u}_k$
  $\boldsymbol{u}_{k+1} \leftarrow \text{Exp}_{\boldsymbol{u}_k}(\eta_A \dfrac{\rho_k}{\lambda} \boldsymbol{g}_k^{\tan} + \sqrt{2\eta_A}\boldsymbol{\xi}_k^{\tan})$ ▷ spherical update
  $\rho_{k+1} \leftarrow \rho_k + \eta_R \left[\dfrac{d-1}{\rho_k} - \rho_k + \dfrac{1}{\lambda} g_k^{\text{rad}}\right] + \sqrt{2\eta_R}\xi_k$ **where** $\xi_k \sim \mathcal{N}(0,1)$ ▷ radial update
  $\boldsymbol{z}_{k+1} \leftarrow \rho_{k+1}\boldsymbol{u}_{k+1}$
return $G_\theta(\boldsymbol{z}_K)$

B.3 Text-to-Image Generation

Models. We utilized SDXL-turbo (Sauer et al., 2024) and DMD2 (Yin et al., 2024a) without Classifier-Free Guidance scale for our text-to-image generative models. All experiments for zeroth-order methods are conducted using NVIDIA RTX 4090 GPUs with VRAM 24GB. For first-order methods which require backpropagation through reward model and generator, NVIDIA A100 with VRAM 40GB was used.

Reward function and text prompts. ImageReward (Xu et al., 2024) and PickScore (Kirstain et al., 2023) return the human preference of a given image related to the text prompt. GenEval (Ghosh et al., 2023) evaluates the alignment between a given image and text prompt by aggregating detection result for text attributes. For counting reward, we employ an external model T2ICount (Qian et al., 2025) that estimates the count of given query objects in a given image. Here, to implement a challenging black-box reward scenario, we rounded the continuous T2ICount output to a nearest whole number before used by any initial noise optimization methods. OCR reward is computed using PaddleOCR (Cui et al., 2025) according to a normalized Levenshtein distance between the recognized and target text.

For counting reward task, we imported a quantity-aware generation prompt set from Yoon et al. (2026) which contains text prompts of various objects attached with a certain query number. For other tasks, we used HPDv2 (Wu et al., 2023) prompts for ImageReward, Pick-a-Pic (Kirstain et al., 2023) prompts for PickScore and OCR, GenEval test prompts for GenEval, where we randomly selected 200 prompts.

Hyperparameters. For the results in Table 1, we use $N = 16$ particles per update for the particle-based zeroth-order optimization methods. The number of iteration was set to $K = 25$ for ImageReward and Counting, and $K = 50$ for PickScore, GenEval, and OCR. These settings correspond to $N \times K = 400$ and 800 number of reward function evaluations, respectively. First-order methods require one function evaluation and its backpropagation per iteration, so we allocated $N \times K$ iterations for each task. Note that ZeNOVA utilized N_{MH} number of reward evaluations for MH jumping, so

only $N - N_{\text{MH}}$ number of particles were used for value gradient estimation unlike other zeroth-order methods. We used $N_{\text{MH}} = 4$ for all tasks.

We set the guidance scale to $1/\lambda = 100.0$ for the Counting reward and $1/\lambda = 1000.0$ for the rest of the reward function, unless specified. Step size for ZeNOVA was set to $\eta_A = 0.1$ and $\eta_R = 0.03$. For horizon scheduling, we fix $\sigma_1 = 0.9$ and $\sigma_K = 0.1$, then use geometric interpolation to get the intermediate horizons so that the consecutive σ have the same ratio.

B.4 Text-to-Video Generation

Model. We use rCM (Zheng et al., 2026b), a few-step distilled text-to-video flow-matching model as our reference video generator. All experiments are conducted on a single NVIDIA RTX PRO 6000 Blackwell.

Reward function and text prompts. Following Xue et al. (2025), we use VideoAlign (Liu et al., 2026) to assess generated videos along three dimensions: Visual Quality (VQ), Motion Quality (MQ), and Text Alignment (TA). We randomly select 30 text prompts from VidProM (Wang & Yang, 2024). We conduct four separate optimization settings, using VQ, MQ, TA, or their sum (*i.e.*, All) as the reward objective.

Hyperparameters. We use $N = 32$ particles per update and $K = 15$ optimization iterations in all four reward settings, corresponding to $N \times K = 480$ particle reward evaluations. The other hyperparameters are set as the same as in the text-to-image generation.

C Additional Results

Figures 8-13 show qualitative examples with various black-box rewards from different text-to-image generative models(SDXL-Turbo, DMD2) and the text-to-video generative model(rCM) with ZeNOVA.

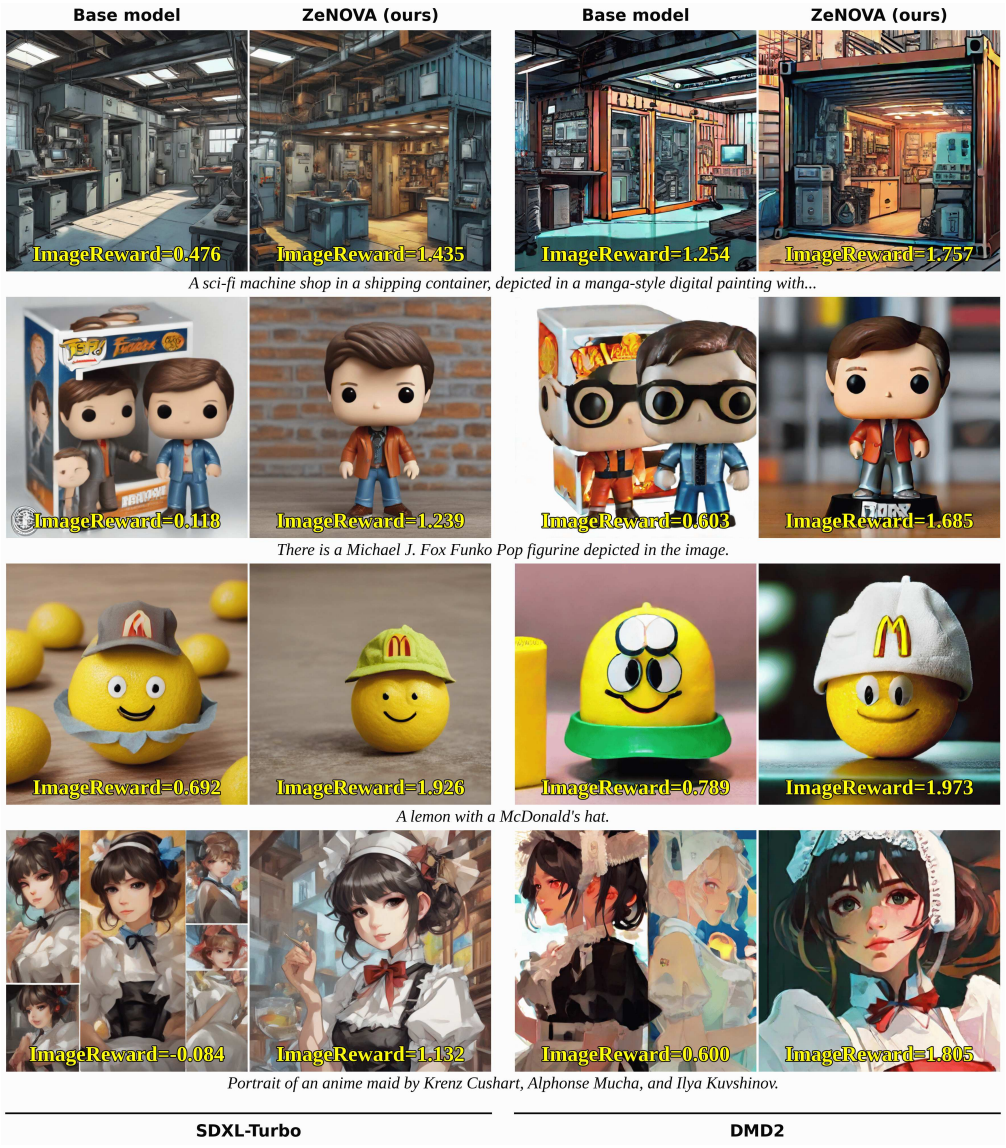

Figure 8: Images generated by SDXL-Turbo and DMD2 with ZeNOVA, using ImageReward as the reward signal.

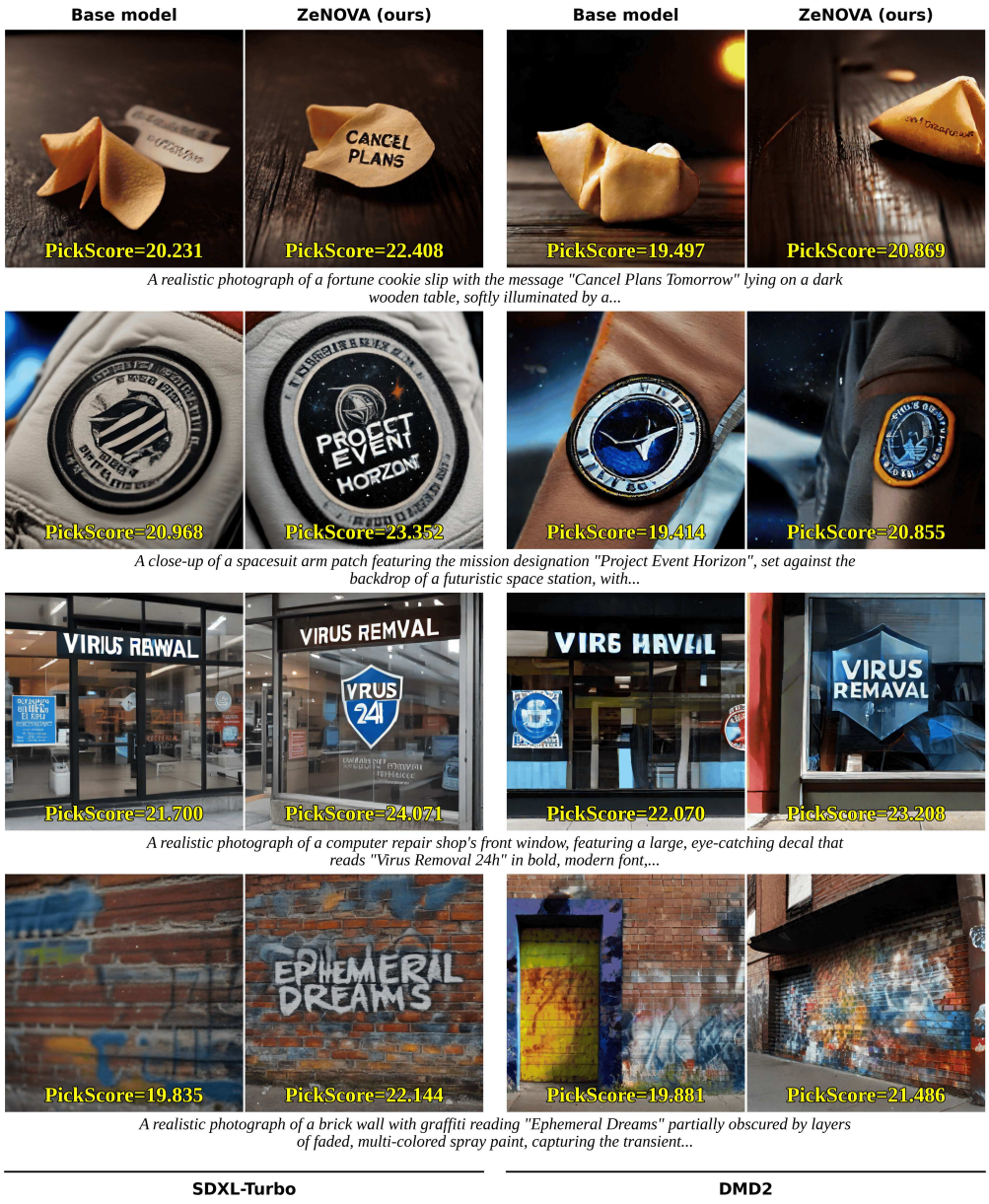

Figure 9: Images generated by SDXL-Turbo and DMD2 with ZeNOVA, using PickScore as the reward signal.

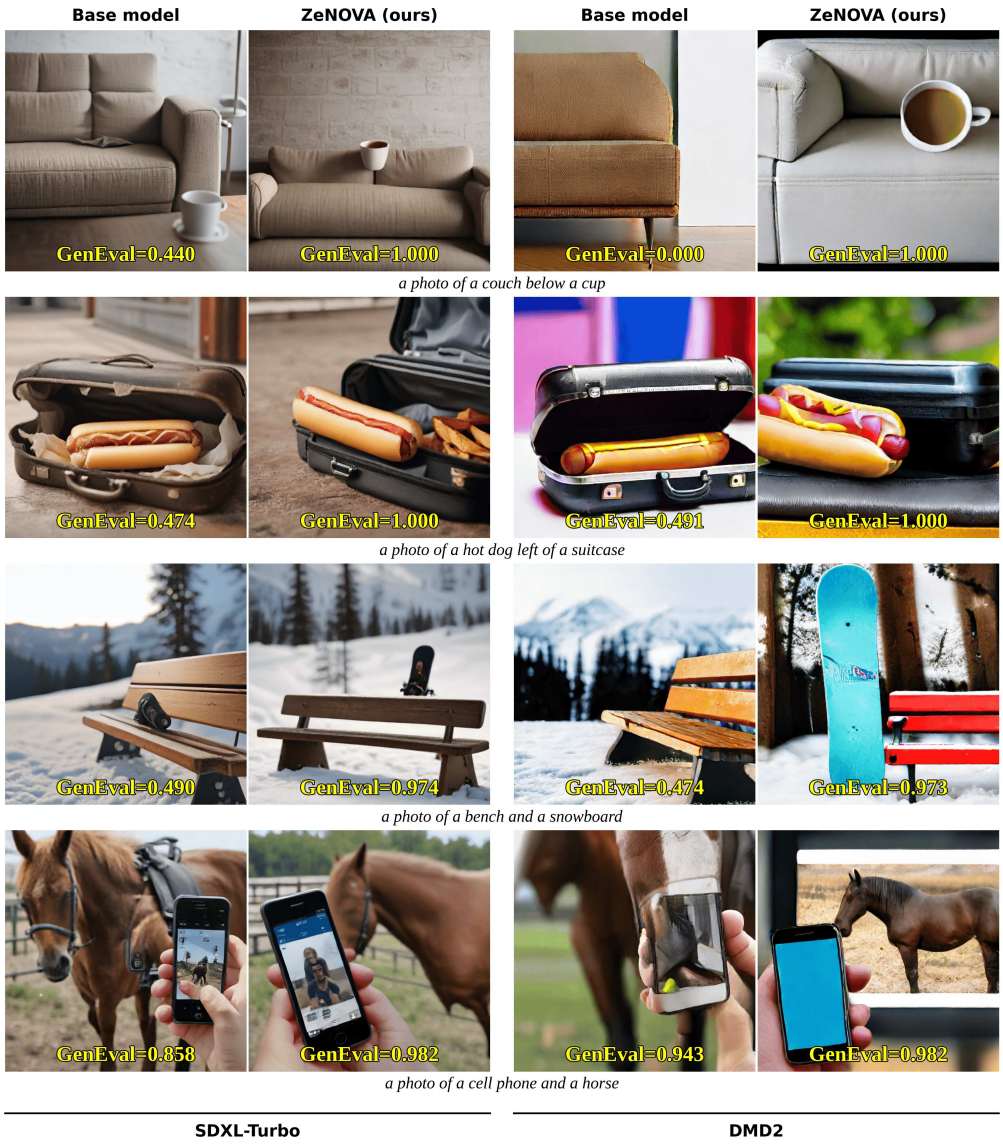

Figure 10: Images generated by SDXL-Turbo and DMD2 with ZeNOVA, using GenEval as the reward signal.



Figure 11: Images generated by SDXL-Turbo and DMD2 with ZeNOVA, using the counting reward as the reward signal.

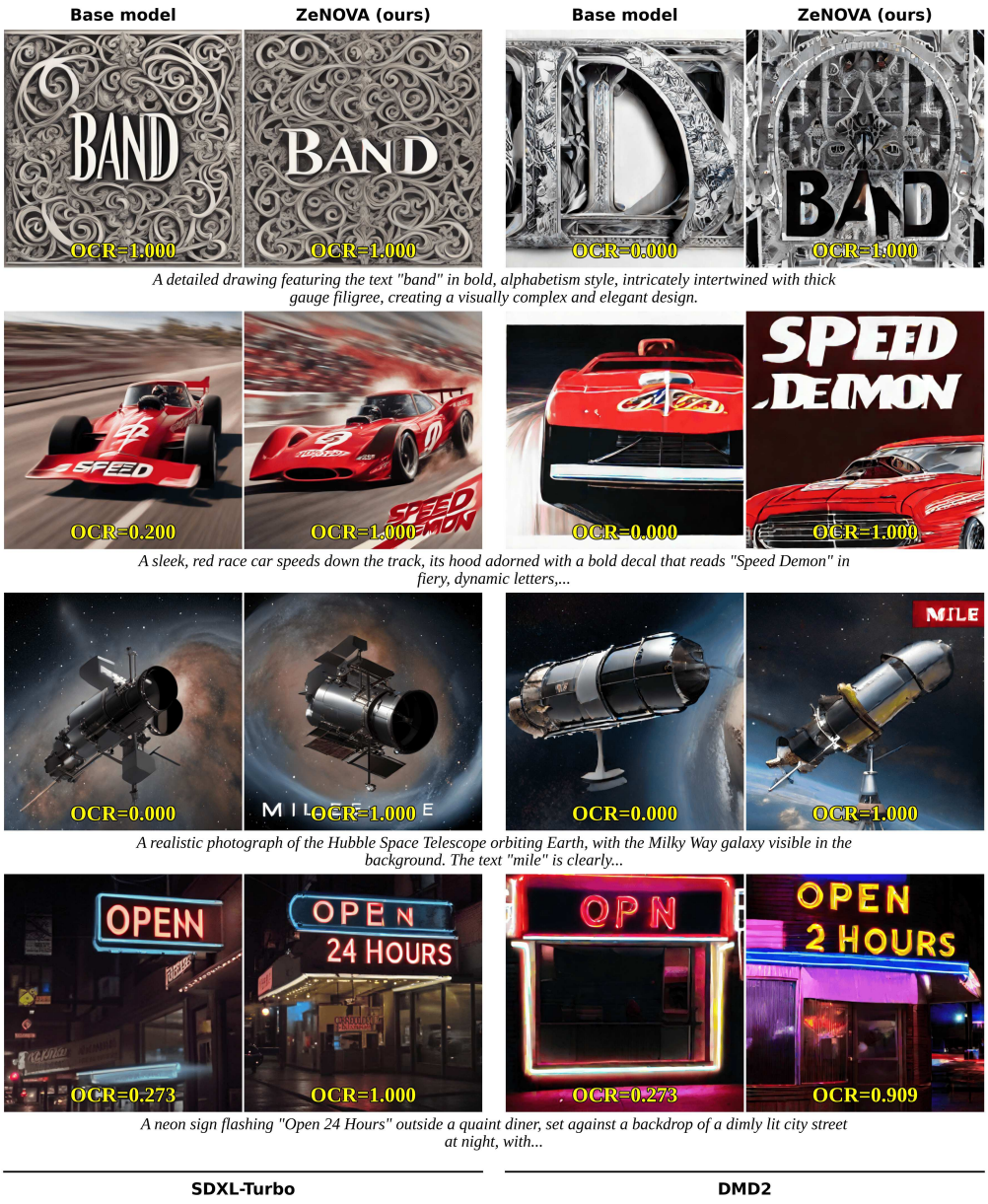

Figure 12: Images generated by SDXL-Turbo and DMD2 with ZeNOVA, using the OCR reward as the reward signal.

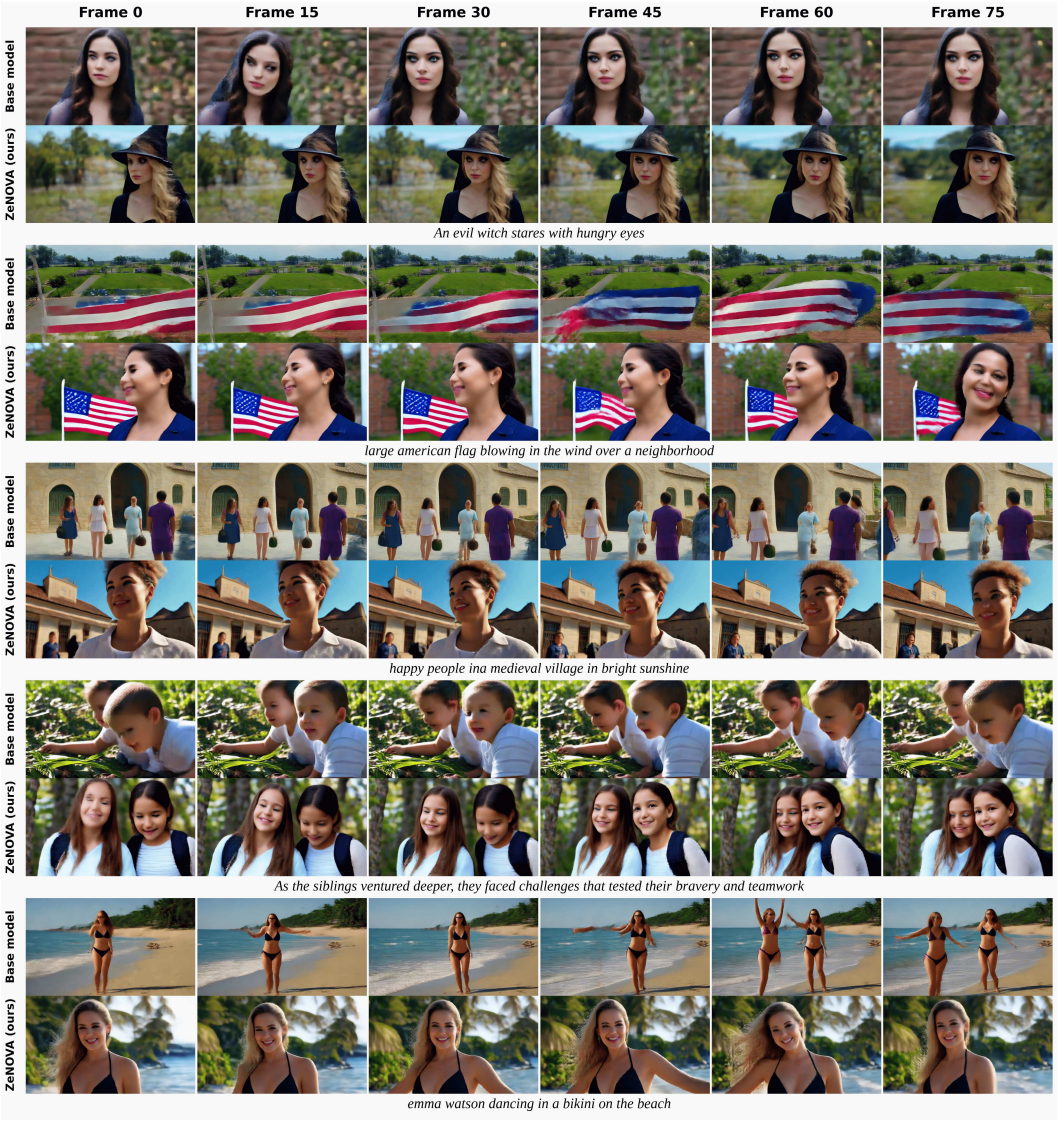

Figure 13: Vidoes generated by rCM with ZeNOVA, using a sum of three VideoAlign rewards as the reward signal.