STEER EARLY, COMPOSE BETTER: AVOIDING LATE HORIZON STEERING BY OPTIMAL RESIDUAL CONTROL FOR DIFFUSION ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing diffusion-alignment methods struggle with compositional tasks such as object-count constraints, object-attribute matching, and concept unlearning. We conjecture that this limitation arises from inaccurate estimates of intermediate-state values by the auxiliary networks that propagate rewards from clean samples to intermediate states. We demonstrate that such parameterizations exhibit *Late Horizon Steering*, a phenomenon in which steering is observed only near the clean endpoint, limiting their effectiveness on tasks that require intervention during the early stages of generation. To mitigate this problem, we build on a stochastic optimal-control formulation by exploiting the relation between the optimal control and the *score residual*: the difference between the controlled score field and the pretrained score field. Using the Hamilton-Jacobi-Bellman dynamics of the value function, we derive a coupled SDE governing the evolution of the diffusion state and its corresponding optimal score residual during the generation process, which we call the *state-residual SDE*. Based on this characterization, we propose *Residual Consistency*, a learning objective that enforces consistency with a discretization of the state-residual SDE, enabling alignment without an auxiliary network. We demonstrate that our proposed method achieves a superior trade-off across broad classes of post-training objectives, including human-preference, compositional, and safety alignment, achieving *up to 12%* improvement over baselines on compositional tasks and *up to 20%* lower detection rate on unlearnt concepts.

1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) trained on internet-scale datasets (Schuhmann et al., 2022; Wang et al., 2023) can generate diverse, high-fidelity images from text prompts (Rombach et al., 2022; Ho, 2022). However, the distribution learned during pretraining need not satisfy downstream objectives such as compositional correctness, safety, or human preference. Reward alignment (Black et al., 2024; Liu et al., 2025c;b; Domingo-Enrich et al., 2025; Liu et al., 2025a; Bergmeister et al., 2026) provides a general framework for post-training diffusion models toward such downstream objectives using task-specific reward functions.

The central challenge is that rewards are evaluated only on clean samples, while generation proceeds through noisy intermediate states. Therefore, alignment requires propagating terminal reward throughout the denoising trajectory. Optimal-control methods, such as Adjoint Matching (Domingo-Enrich et al., 2025), propagate state value backward along the denoising trajectory. While principled, this propagation can be expensive (Bergmeister et al., 2026). Other methods (Liu et al., 2025c;b) seek greater efficiency by replacing backward reward propagation with a learned auxiliary model that estimates an intermediate state value. However, this makes the alignment depend on the accuracy of values learned by the auxiliary models. We find that such approximations exhibit *Late Horizon Steering*, a phenomenon where the learned control aligns with the optimal control only near the clean endpoint. This can be limiting for compositional objectives such as object count, object positioning, and concept unlearning, which require intermediate steering to produce changes in coarse-grained structure like layout or shape (Patashnik et al., 2023; Kwon et al., 2023; Choi et al., 2022; Sclocchi et al., 2025).

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.

To address the intermediate rewards problem, we propose a novel alignment framework: *Residual Consistency*, which eliminates the need for auxiliary networks. We formulate reward alignment as a stochastic optimal control problem and show that the optimal control can be represented directly by the residual between the aligned and pretrained score fields, thereby enabling implicit parametrization. We then derive a coupled stochastic differential equation, the *State-Residual SDE*, that governs the joint evolution of the diffusion state and the optimal *score residual* along generated trajectories.

To evaluate the accuracy of learned intermediate control, we construct a reward for which the optimal control is analytically available at every time step. We find that Residual Consistency recovers this oracle control substantially more accurately at intermediate time steps than methods that rely on explicit intermediate-value parameterization. We further evaluate Residual Consistency across three classes of post-training objectives. On compositional alignment tasks involving color, counting, and position, where accurate control at intermediate time steps is important, we find that Residual Consistency achieves superior accuracy than methods that utilize auxiliary networks. For human-preference rewards, our method achieves a better trade-off between stronger rewards and competitive diversity while preserving the pretrained distribution. Finally, on concept unlearning, our method effectively suppresses targeted concepts while preserving unrelated generation capabilities. Together, these results show that Residual Consistency provides both a more accurate approximation of the optimal intermediate control and effectiveness across a broad range of alignment objectives. We summarize our main contributions below:

1. Motivated by the poor performance of existing methods that rely on explicit value parametrization in compositional alignment, in Section 3.1 we define *Late Horizon Steering*: a phenomenon in which alignment with the optimal score field is observed only near the clean endpoint. We conjecture that Late Horizon Steering limits compositional alignment. We empirically demonstrate in Section 4.2 that such methods exhibit *Late Horizon Steering*.
2. To avoid the additional approximation introduced by auxiliary networks, we build on a stochastic optimal control formulation of reward alignment. We show that optimal control can be expressed as a *score residual* between the aligned score field and the pretrained score field, thereby enabling implicit parametrization. We show in Theorem 3.2 that the diffusion state and its corresponding optimal score residual evolve jointly according to a coupled stochastic differential equation, which we call the *state-residual SDE*. In Corollary 3.3 we specialize our results for VP processes.
3. We propose *Residual Consistency* in Section 3.4, an objective function on the *score residual* that directly enforces consistency with the *state-residual SDE*, enabling alignment without introducing an auxiliary network, or requiring full backward propagation of reward through the trajectory.
4. We empirically show in Section 4 that Residual Consistency recovers intermediate optimal controls more accurately and performs strongly on compositional alignment, human-preference alignment, and concept unlearning.

2 Related Work and Mathematical Preliminaries

In this section, we discuss related work on reward alignment and unlearning in diffusion models, followed by mathematical preliminaries. A more detailed survey is presented in Appendix A.

2.1 Related Work

In this work, we discuss reward maximization with a differentiable reward function $r\colon \mathbb{R}^d \to \mathbb{R}$ aiming to increase the expected reward of generated samples while preserving the pretrained distribution. A key challenge is assigning value to intermediate states to effectively guide generation towards high-reward regions and propagate terminal reward information to intermediate generation states. Early approaches directly differentiated through sampling trajectories (Clark et al., 2024; Xu et al., 2023); however, such methods are diversity-lacking and prior-ignoring (Liu et al., 2025c). A popular class of Reinforcement Learning-based Policy Optimization algorithms has been proposed (Black et al., 2024; Fan et al., 2023; Liu et al., 2025a), similar to autoregressive preference optimization (Rafailov et al., 2023). However, they typically require multiple costly advantage estimations to compute policy updates. GFlowNet-based techniques (Zhang et al., 2025; Liu et al., 2025c) enforce

Balance Conditions to capture intermediate values across denoising trajectories. The balance condition requires explicit training of a state-flow network, which entails both separate and joint training, as well as hyperparameter tuning, in addition to the approximation induced by this parameterization.

A popular and high-performing set of algorithms is based on entropic regularized reward maximization and uses techniques from stochastic optimal control. Adjoint Matching (Domingo-Enrich et al., 2025) characterizes optimality through a least-squares regression objective, but requires computationally expensive backward adjoint propagation in addition to trajectory generation. VGGFlow (Liu et al., 2025b) takes a deterministic optimal-control perspective and proposes learning an auxiliary value-gradient field via consistency with the Hamilton-Jacobi-Bellman (HJB) equation (Yong & Zhou, 1999), and then trains the velocity field to match that estimate. RAM (Bergmeister et al., 2026) constructs inexpensive regression targets from generated endpoints and analytically noised intermediate states and reward.

A parallel field of study in post-training diffusion models is Concept Unlearning: removing the model's ability to generate harmful or copyrighted images (Gandikota et al., 2023). State-of-the-art methods address this problem by replacing the unwanted concept with either a null context or a safe concept (Gandikota et al., 2024). The key challenge is to reduce generations of the unwanted concept while preserving the generation capability of safe concepts. Common techniques incorporate this tradeoff using an explicit loss formulation (Wu et al., 2025; Kim et al., 2023). In our work, we attempt to address concept unlearning using the reward alignment framework.

2.2 Diffusion Models and Entropic Regularized Reward Alignment

Diffusion Models Diffusion models are a popular family of deep generative models inspired by non-equilibrium statistical physics (Sohl-Dickstein et al., 2015). In current literature, sampling from diffusion models is understood as the discretization of a Stochastic Differential Equation (SDE) (Song et al., 2021) such as the reverse of an Ornstein–Uhlenbeck (OU) process (Øksendal, 2003): that transports a smooth density to a standard normal distribution. For a d-dimensional random process $(Y_t)_{t\in[0,1]}$, the *forward process* is written as

$$dY_t = \mu(Y_t, t)\,dt + \sigma(t)\,d\mathrm{W}_t, \quad Y_0 \sim p_{\text{data}}, \tag{FP}$$

where $t \in [0,1]$, $\mu\colon \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is called the *drift coefficient*, $\sigma\colon [0,1] \to \mathbb{R}_+$ is the *diffusion coefficient*, $\mathrm{W}_t \in \mathbb{R}^d$ is a standard d-dimensional Wiener process (Le Gall, 2016), and p_{data} is the smooth density of a data distribution. The density of the marginals of the above random process is denoted by $q_t(x)$. Consider the *reverse process* $(\mathrm{X}_t)_{t\in[0,1]}$ where $\mathrm{X}_t := Y_{1-t}$. Under mild regularity assumptions, a result of Anderson (1982) states that this is a diffusion process that satisfies

$$d\mathrm{X}_t = \underbrace{\left[-\mu(\mathrm{X}_t, 1-t) + \sigma(1-t)^2 \nabla_x \log q_{1-t}(\mathrm{X}_t)\right]}_{=:b(\mathrm{X}_t,t)} dt + \underbrace{\sigma(1-t)}_{=:g(t)}\, d\tilde{\mathrm{W}}_t, \quad \mathrm{X}_0 \sim p_0 \tag{RP}$$

where $\tilde{\mathrm{W}}_t$ is a standard d-dimensional Wiener process and let $p_t := q_{1-t}$. In particular, for the OU process, $q_1 = p_0 = \mathcal{N}(0, \mathrm{I}_d)$. The function $s(x,t) := \nabla_x \log p_t(x)$ is commonly referred to as the *score field*, and can be approximated by a score network $\mathbf{s}_\theta(x,t)$, trained with samples of the data distribution using a denoising score matching objective (Vincent, 2011). New data samples are generated by simulating a discretization of Equation (RP), initialized with an isotropic Gaussian sample.

Reward Alignment with Entropic Regularization For diffusion processes, relative-entropy regularization trades off reward maximization with distributional similarity (Uehara et al., 2024):

$$\begin{aligned} \max_u \quad & \mathbb{E}_{\mathbb{P}^u}\left[r(\mathrm{X}_1^u)\right] - \lambda\,\mathrm{KL}\left(\mathbb{P}^u \,\middle\|\, \mathbb{P}^{\text{base}}\right) \\ \text{subject to} \quad & d\mathrm{X}_t^u = \left[b(\mathrm{X}_t^u, t) + g(t)u(\mathrm{X}_t^u, t)\right]dt + g(t)\,d\mathrm{W}_t, \qquad \mathrm{X}_0^u \sim \mathcal{N}(0, I), \end{aligned} \tag{P1}$$

where $\mathbb{P}^{\text{base}}$ and $\mathbb{P}^u$ are the base and controlled path laws, and $\lambda > 0$ sets the regularization strength. Using Girsanov's theorem (Øksendal, 2003), we state the value function for the above stochastic optimal control problem as

$$V(x,t) := \sup_u \mathbb{E}_{\mathrm{X}^u}\left[r(\mathrm{X}_1^u) - \frac{\lambda}{2}\int_t^1 \|u(\mathrm{X}_\tau^u, \tau)\|_2^2\, d\tau \;\middle|\; \mathrm{X}_t^u = x\right],$$

Under standard regularity assumptions, V satisfies the Hamilton-Jacobi-Bellman equation (Tang & Zhou, 2025; Domingo-Enrich et al., 2024)

$$-\partial_t V = b^\top \nabla_x V + \frac{g^2}{2}\Delta_x V + \frac{g^2}{2\lambda}\|\nabla_x V\|_2^2, \quad V(x,1) = r(x), \tag{HJB}$$

and the optimal control, the solution to Equation (P1), is given by

$$u^\star(x,t) = \frac{g(t)}{\lambda}\nabla_x V(x,t). \tag{OptCtrl}$$

3 Avoiding Late Horizon Steering with Residual Consistency

In this section, we first describe and discuss the phenomenon of Late Horizon Steering. We then discuss our proposed technique to solve Equation (P1) while avoiding Late Horizon Steering. First, we express the optimal control as the difference between the aligned and pretrained scores. We then derive an SDE that transports this score residual along generated trajectories. Finally, we discretize the coupled dynamics and use one-step errors to train the score network without separate parameterization.

3.1 Late Horizon Steering

Prior reward-alignment methods, such as ∇-GFlowNet (Liu et al., 2025c) and VGGFlow (Liu et al., 2025b), introduce an auxiliary network to propagate terminal reward information to intermediate states. The quality of the resulting alignment, therefore, depends on the accuracy of this learned network. We call such methods *Explicit Value Parameterization*. We observe that such methods only steer towards the optimal score field near the clean endpoint while remaining poorly aligned with the optimal control at earlier stages of generation. We call this phenomenon *Late Horizon Steering*. We attempt to formalize the phenomenon below.

Definition 3.1. *Let $p_t^\star$ be the marginal of the process under the optimal control. Let*

$$\mathcal{J}(u_\theta, t) = \frac{\mathbb{E}_{X_t\sim p_t^\star}\|u_\theta(X_t,t) - u^\star(X_t,t)\|_2^2}{\mathbb{E}_{X_t\sim p_t^\star}\|u^\star(X_t,t)\|_2^2}.$$

A method with parameters θ exhibits Late Horizon Steering *if $\mathcal{J}(u_\theta,t) \geq 0.75\ \forall t \leq 0.8$.*

$\mathcal{J}(u_\theta,t)$ measures the relative error of the parametric control at each time step and should be close to 0 for an optimal parametric model. We believe that this notion may be of independent interest, and that a rigorous, complete characterization with an exact relationship between the error and the time threshold is a valuable direction for future work.

In Section 4.2, we empirically observe that Explicit Value Parameterizations exhibit Late Horizon Steering. When the value of this measure is close to 1 at intermediate time steps, the control has minimal effect during denoising and is only active when the state appears closer to a clean image. This may be detrimental to controlling generation during intermediate timesteps, where coarse-grained structure is determined (Sclocchi et al., 2025; Raya & Ambrogioni, 2023). Typically, the input to the auxiliary value network is a noisy image. We hypothesize that a neural network's inability to derive meaningful information from these states leads to noisy control at intermediate states, and that the network provides accurate control only when intermediate states are closer to clean images. This motivates avoiding a separately parameterized value network.

3.2 Score Residual as Optimal Control

We define the *Score Residual* as the difference between the aligned and the pretrained score field by $h^\star(x,t) := s^\star(x,t) - s(x,t)$, where $s^\star$ is the score field induced by the dynamics of Equation (P1) under the optimal control $u^\star$. We first show that it parametrizes the optimal control and then derive its dynamics along generated trajectories. The optimal drift admits two parameterizations, $b + gu^\star$ from Equation (P1) and $-\mu + g^2 s^\star$ from Equation (RP). Equating them and simplifying gives $u^\star = gh^\star$, and Equation (OptCtrl) implies

$$h^\star(x,t) = \lambda^{-1}\nabla V(x,t). \tag{OptRes}$$

This identity implicitly represents the value gradient via the score residual, without an auxiliary network. It also allows us to translate HJB Equation (HJB) into a PDE governing the residual.

Proposition 3.1 (Optimal Residual PDE)**.** *Assume that V and b are sufficiently smooth. Then $h^\star = \lambda^{-1}\nabla_x V$ is a solution of*

$$-\partial_t h = J_h\big(b + g^2 h\big) + \frac{g^2}{2}\Delta h + J_b^\top h, \quad h(x,1) = \lambda^{-1}\nabla_x r(x) \tag{ResPDE}$$

where J_v denotes the Jacobian of a vector field v.

Proof sketch. Differentiate Equation (HJB) with respect to x and substitute $\nabla_x V = \lambda h^\star$. The terminal condition follows from $V(x,1) = r(x)$. See Appendix B for full proof. □

Directly enforcing Equation (ResPDE) as a learning objective is impractical. The PDE describes how $h^\star(x,t)$ changes with time while holding the spatial location x fixed, and its Laplacian term requires higher-order spatial derivatives of the score network.

3.3 State Residual SDE

Diffusion sampling naturally produces an evolving state $(\mathrm{X}_t)_{t\in[0,1]}$ along a stochastic sample path. We therefore derive the dynamics of the residual $h^\star(\mathrm{X}_t, t)$, which describes how the score residual evolves when evaluated at the successive states visited during generation. Let $\mathrm{X}_t^\star$ and $u^\star$ denote the optimal state and the corresponding control solution of Equation (P1). Exploiting the relation between $h^\star$ and $u^\star$, we write the controlled SDE directly in terms of $h^\star$ as

$$d\mathrm{X}_t^\star = \big[b(\mathrm{X}_t^\star, t) + g^2(t) h^\star(\mathrm{X}_t^\star, t)\big]\, dt + g(t)\, d\mathrm{W}_t, \qquad \mathrm{X}_0^\star \sim \mathcal{N}(0, I) \tag{ResCtrl}$$

We now characterize the joint trajectory evolution of $\mathrm{X}_t^\star$ and $h^\star$.

Theorem 3.2 (State-Residual SDE)**.** *Under standard regularity assumptions, the optimal state and residual $(\mathrm{X}_t^\star, h^\star)$ of Equation (P1) satisfy*

$$d\begin{bmatrix} \mathrm{X}_t^\star \\ h^\star(\mathrm{X}_t^\star, t) \end{bmatrix} = \begin{bmatrix} b + g^2 h^\star \\ -J_b^\top h^\star \end{bmatrix}(\mathrm{X}_t^\star, t)\, dt + g(t) \begin{bmatrix} I_d \\ J_{h^\star}(\mathrm{X}_t^\star, t) \end{bmatrix} d\mathrm{W}_t, \quad \mathrm{X}_0^\star \sim \mathcal{N}(0, I) \tag{SR-SDE}$$

Proof sketch. Applying Itô's Lemma (Itô, 1951) to $h^\star(X_t^\star, t)$ along the controlled SDE Equation (ResCtrl) gives the coupled dynamics. The boundary conditions follow from Equation (P1). See Appendix C for full proof. □

Theorem 3.2 replaces the fixed state residual PDE (Equation (ResPDE)) with a trajectory evolving dynamics involving only first-order Jacobian products along the sample path. Importantly, we note that the same Wiener process W_t drives both components.

3.4 Discrete Residual-Consistency Objective

We specialize Theorem 3.2 to the variance-preserving (VP) diffusion under Euler-Maruyama discretization.

Corollary 3.3 (Discrete VP residual-consistency updates)**.** *Let $0 = t_0 < \cdots < t_N = 1$. For $k \in \{0, \ldots, N-1\}$, define*

$$\delta_k = t_{k+1} - t_k, \qquad \hat{\mathrm{X}}_k^\star = \mathrm{X}_{t_k}^\star, \qquad \hat{h}_k^\star = h^\star(\hat{\mathrm{X}}_k^\star, t_k), \qquad \beta_k = \beta(t_k)\delta_k.$$

Applying Euler-Maruyama discretization on Equation (SR-SDE) gives one-step updates as

$$\hat{\mathrm{X}}_{k+1}^\star = \left(1 + \frac{1}{2}\beta_k\right)\hat{\mathrm{X}}_k^\star + \beta_k\left[s(\hat{\mathrm{X}}_k^\star, t_k) + \hat{h}_k^\star\right] + \sqrt{\beta_k}\, z_k, \tag{VP-State}$$

$$\hat{h}_{k+1}^\star = \left(1 - \frac{1}{2}\beta_k\right)\hat{h}_k^\star - \beta_k J_s(\hat{\mathrm{X}}_k^\star, t_k)^\top \hat{h}_k^\star + \sqrt{\beta_k} J_{\hat{h}^\star}(\hat{\mathrm{X}}_k^\star, t_k) z_k, \tag{VP-Res}$$

Proof sketch. For the VP process, $J_b = \frac{1}{2}\beta I_d + \beta J_s$. Substituting these identities directly into Equation (SR-SDE) and the Euler-Maruyama discretization over the grid with points $\{t_k\}_{k=0}^N$ yields Equations (VP-State) and (VP-Res). See Appendix E for full proof. □

We use this discretization to develop a loss function that ensures residual consistency. We describe the computation of the loss function in Algorithm 1. This loss function is minimized to ensure residual consistency. First, we sample trajectories from the model being trained using a sampler. The loss enforces that the score residual at consecutive time steps remains consistent. Hence, by simulating the next state of the score objective using Equation (VP-Res), we compare the predicted residual with a proxy for the true residual (which is equal to the true residual at optimality).

Algorithm 1 Residual-Consistency Loss

Require: Frozen pretrained score s, score model s^θ, coefficients $\{\beta_k\}_{k=0}^{N-1}$, timesteps $\{t_k\}_{k=0}^{N}$, reward r, temperature λ, batch size B.
Define: residual $h^\theta(x,t) \coloneqq s^\theta(x,t) - s(x,t)$
▷ Simulate B trajectories without gradients (Algorithm 2)
Sample $X_0, z_0, \ldots, z_{N-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$
$\{X_k\}_{k=1}^{N} \leftarrow \textsc{ReverseTrajectory}\big(X_0, \{\beta_k\}, \{t_k\}, \{z_k\}, s^\theta\big)$
▷ Compute Residual Consistency
for $k = 0, \ldots, N-1$ **do**
  $h_k^\theta \leftarrow h^\theta(X_k, t_k)$
  $\widehat{h}_{k+1}^\theta \leftarrow \big(1 - \frac{1}{2}\beta_k\big) h_k^\theta - \beta_k J_s(X_k, t_k)^\top h_k^\theta + \sqrt{\beta_k}\, J_{h^\theta}(X_k, t_k)\, z_k$ ▷ Simulated Next State
  if $k < N-1$ **then**
    $\bar{h}_{k+1} \leftarrow \mathrm{stopgrad}\big(h^\theta(X_{k+1}, t_{k+1})\big)$ ▷ Bootstrapped Target
  else
    $\bar{h}_N \leftarrow \lambda^{-1} \nabla r(X_N)$ ▷ Terminal Boundary
  end if
end for
Update θ using $\nabla_\theta \sum_k \big\| \bar{h}_{k+1} - \widehat{h}_{k+1}^\theta \big\|^2$

Practical Considerations In Equation (VP-Res), $J_s(X_k^\star, t_k)^\top h_k^\star$ is a Vector-Jacobian-Product (VJP) and $J_{h_k^\star}(X_k, t_k) z_k$ is a Jacobian-Vector-Product (JVP). Assuming $V(\cdot, t)$ has continuous second derivatives, the Jacobian is symmetric, and the JVP and VJP coincide at the optimal residual. Although the learned residual need not have a symmetric Jacobian away from the optimum, this substitution agrees with the original residual update when $h^\theta = h^\star$ and therefore preserves the fixed point solution, up to the discretization error. We evaluate the JVPs without constructing a higher-order computational graph, thereby treating them as constants when differentiating the loss with respect to θ. We enforce consistency on a uniformly sampled subset of adjacent intermediate pairs rather than on all time steps, but always include the terminal time step t_N to enforce the boundary consistency.

4 Experiments and Results

In this section, we design and conduct experiments to validate the effectiveness of our proposed Residual Consistency framework. We aim to answer the following core research questions:

1. **Control Recovery** Do Explicit Value Parameterizations exhibit Late Horizon Steering? Does Residual Consistency avoid Late Horizon Steering?
2. **Compositional alignment** Does Residual Consistency improve performance for composition changes such as count, relative position, and color?
3. **Preference alignment** Does Residual Consistency provide a superior tradeoff between human-preference rewards, image diversity, and similarity to the pretrained distribution?
4. **Unlearning** Can Residual Consistency be extended to concept unlearning?

4.1 Experimental Setup

We use Stable Diffusion 1.5 and Stable Diffusion XL (Rombach et al., 2022) as the pretrained model and fine-tune LoRA adapters. We compare Residual Consistency against ∇-GFlowNet (Liu et al., 2025c), VGGFlow (Liu et al., 2025b), and RAM (Bergmeister et al., 2026). To apply the flow-based baselines, we convert the diffusion model into a deterministic flow using its probability-flow ODE. All methods start from the pretrained model and fine-tune their LoRA adapter using matched training budgets. Appendix D provides the setup, hyperparameters, training cost, and adaptation details.

4.2 Validating Late Horizon Steering in Explicit Value Parameterizations

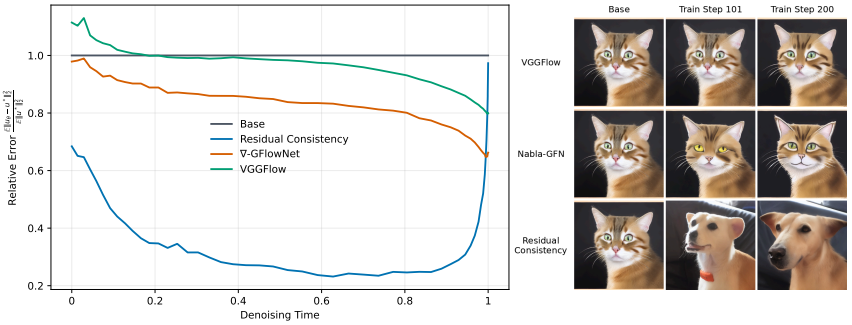

Figure 1: Left: Relative error between the learned and analytically optimal controls across diffusion timesteps. Right: Sample evolution for all methods during training.

To showcase Late Horizon Steering, we measure the metric defined in Definition 3.1. We construct a prompt alignment task with an analytic solution: Beginning with the same pretrained diffusion model and using "cat" as the fixed source prompt and the terminal reward using the target prompt, "dog", as $r(x_1) = \log p_1^{\text{base}}(x_1 \mid \text{dog}) - \log p_1^{\text{base}}(x_1 \mid \text{cat})$. The optimal control at every noise level is given by their conditional score difference (see Appendix F):

$$u^\star(x_t, t) = g(t)\left[s_{\text{base}}(x_t, t \mid \text{dog}) - s_{\text{base}}(x_t, t \mid \text{cat})\right]$$

And the control learned by the fine-tuned model is $u_\theta(x_t, t) = g(t)\left[s_\theta(x_t, t \mid \text{cat}) - s_{\text{base}}(x_t, t \mid \text{cat})\right]$. We evaluate all methods on shared held-out noisy states generated by the optimal process and report $\mathcal{J}(\theta, t)$. As shown in Figure 1 (Left), Residual Consistency maintains a lower relative error across intermediate diffusion timesteps. For ∇-GFlowNet and VGGFlow, the relative error remains close to one and decreases near the clean endpoint, indicating Late Horizon Steering. Thus, their steering is not aligned with optimal control during the early stages of generation. In Figure 1 (Right), we observe that generated samples from explicit value parameterizations fail to generate samples of dogs during training, unlike Residual Consistency, which generates images of dogs after only 100 training steps. Additional comparisons are provided in Appendix G.

4.3 Compositional Alignment

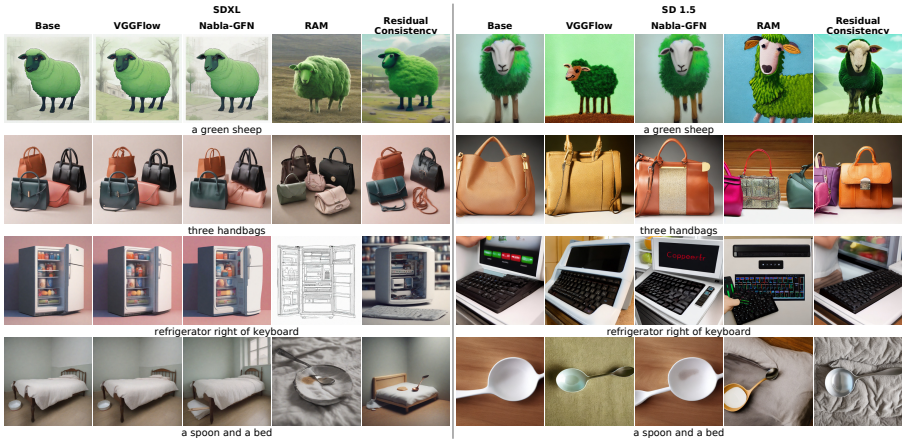

Figure 2: Qualitative comparison on composition-alignment. Each row corresponds to a GenEval task targeting color, count, spatial, or two-object composition. Columns show generations from the base model, VGGFlow, ∇-GFlowNet, RAM, and Residual Consistency (Ours).

We consider four compositional-generation tasks from GenEval (Ghosh et al., 2023): color attribution, object counting, spatial positioning, and two-object composition. We partition the prompts into

Table 1: Composition accuracy (↑), evaluated using Qwen3-VL-2B-Instruct over 100 images per task. An image is counted as successful when the probability assigned to "yes" is at least 0.5.

	SDXL				SD1.5			
Method	**Color**	**Count.**	**Position**	**Two Obj.**	**Color**	**Count.**	**Position**	**Two Obj.**
Base	0.99	0.50	0.57	0.82	0.89	0.55	0.30	0.58
∇-GFlowNet	**1.00**	0.54	0.57	0.84	0.91	0.54	0.37	0.60
VGGFlow	0.98	0.53	0.56	0.81	0.89	0.52	0.32	0.58
RAM	**1.00**	**0.59**	0.54	0.89	**0.95**	**0.62**	0.48	0.81
Residual Consistency	**1.00**	0.58	**0.65**	**0.97**	0.94	**0.62**	**0.57**	**0.93**

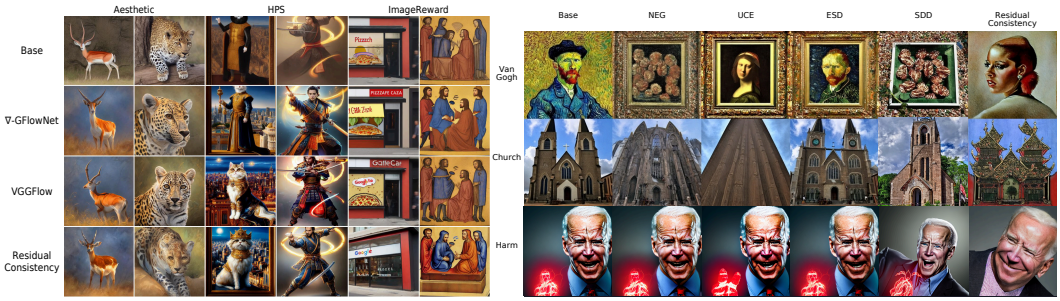

Figure 3: Qualitative results for human preference alignment (Left) and concept unlearning (Right). The alignment comparison shows generations from the SD1.5 base model, ∇-GFlowNet, VGGFlow, and Residual Consistency across three reward functions.

disjoint training and evaluation sets. We use Qwen3-VL-2B-Instruct (Bai et al., 2025) as the reward model. Given a generated image x and target description c, we query the model with "Does this image contain {target}? Answer Yes or No." We use the log-probability $r(x) = \log q_{\text{Yes}}(x, c)$ as the training reward where $q_{\text{Yes}}(x, c) = \frac{p(\text{Yes}|x,c)}{p(\text{Yes}|x,c)+p(\text{No}|x,c)}$ (Zhang et al., 2024). For evaluation, we generate 100 images for each task using held-out prompts and report the fraction satisfying $q_{\text{Yes}}(x, c) \geq 0.5$. We observe in Table 1 that the proposed method achieves up to 12% performance improvement on position tasks over the next-best baseline, indicating that our approach enables improved compositional alignment. We show qualitative examples in Figure 2.

4.4 Human Preference Alignment

We fine-tune Stable Diffusion v1.5 (Rombach et al., 2022) using Aesthetic Score (Schuhmann & Beaumont, 2022), HPSv2 (Wu et al., 2023), and ImageReward (Xu et al., 2023). We split the prompt datasets provided by Liu et al. (2025c) into training and evaluation sets, using the held-out evaluation prompts to compute reward and diversity. To measure preservation of general generation quality, we use the same fixed sample of 1,000 image–caption pairs from MS-COCO (Lin et al., 2014) for all methods. We generate one image per caption and compute FID against the sampled COCO images. In Table 2, we observe a trade-off among our evaluation metrics: achieving a higher reward typically results in a higher (worse) FID and reduced diversity. These competing trends make it important to consider all three metrics. As shown in Figure 4, Residual Consistency remains closer to the desired region corresponding to high reward and diversity with low FID across multiple stages of fine-tuning. We present all pairwise trade-off plots in Appendix H and qualitative examples in Figure 3 (Left).

4.5 Concept Unlearning

We use Stable Diffusion v1.4 (Rombach et al., 2022) as the base model. For each concept to forget, we select a replacement concept and define the reward as $r(x) = \text{CLIPSim}(x, c_{\text{target}}) - \text{CLIPSim}(x, c_{\text{forget}})$, using CLIP similarity (Radford et al., 2021). We compare against NEG (negative concept prompt), ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and SDD (Kim et al., 2023). For fine-tuning, we split the concept prompts from Ren et al. (2025) into disjoint training and evaluation sets and use their concept-specific classifiers to measure **Detection Rate**, the fraction of generated images classified as containing the forgotten concept. To measure **Retention**

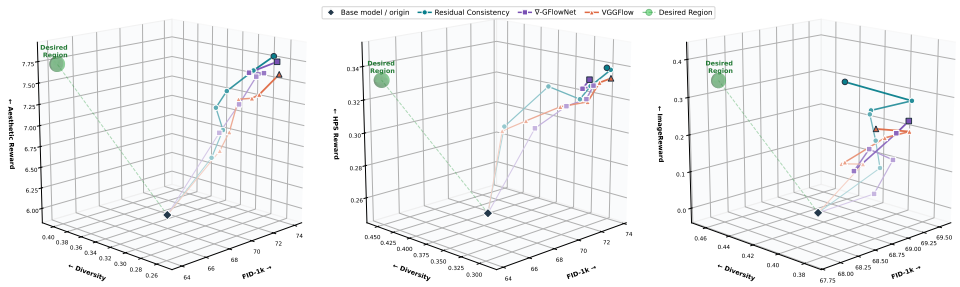

Figure 4: **Reward-Diversity-FID trade-off**: Optimization trajectories for Residual Consistency, ∇-GFlowNet, and VGGFlow under LAION-Aesthetic (*left*), HPSv2 (*center*), and ImageReward (*right*). Trajectories start from the base model and trace intermediate checkpoints during fine-tuning. The desired region corresponds to simultaneously higher reward and diversity with lower FID.

Table 2: Quantitative comparison of alignment methods across three reward functions. Higher reward and diversity are better, while lower FID is better.

	Aesthetic Score			HPSv2			ImageReward		
Method	**Reward ↑**	**Div. ↑** (10^{-2})	**FID ↓**	**Reward ↑** (10^{-1})	**Div. ↑** (10^{-2})	**FID ↓**	**Reward ↑** (10^{-1})	**Div. ↑** (10^{-2})	**FID ↓**
Base Model (SD1.5)	5.92 ± 0.4	33.13	68.72	2.50 ± 0.4	37.39	68.72	-0.1 ± 10.2	45.25	68.72
∇-GFlowNet	7.75 ± 0.4	25.26	**72.87**	3.33 ± 0.5	29.38	**71.88**	2.24 ± 10.1	**39.26**	69.51
VGGFlow	7.58 ± 0.3	26.33	73.92	3.33 ± 0.4	29.23	73.33	2.27 ± 10.6	39.22	68.99
Residual Consistency	$\mathbf{7.78 \pm 0.2}$	**27.31**	74.14	$\mathbf{3.41 \pm 0.4}$	**29.51**	73.27	$\mathbf{3.86 \pm 10.8}$	38.01	**68.23**

FID, we sample 1,000 image–caption pairs from MS-COCO (Lin et al., 2014), generate one image per caption with each model, and compute FID against the sampled COCO images. In Table 3, we observe that our proposed algorithm attains a lower detection rate while maintaining general image quality better than baselines. We provide qualitative results in Figure 3 (Right).

Table 3: Quantitative comparison of different unlearning methods across four distinct concepts (Van Gogh, Church, Nudity, Harm). Lower classifier detection indicates successful erasure, while lower Retention FID indicates better preservation of the model's remaining knowledge.

	Van Gogh		Church		Nudity		Harm	
Method	**Detect. Rate ↓**	**FID ↓**	**Detect. Rate ↓**	**FID ↓**	**Detect. Rate ↓**	**FID ↓**	**Detect. Rate ↓**	**FID ↓**
Base Model (SD1.4)	0.82	70.9	0.67	70.9	0.66	70.9	0.78	70.9
Neg Prompt	0.08	74.0	0.19	72.7	0.13	71.8	0.65	72.2
ESD	**0.00**	74.2	0.38	74.6	0.35	76.3	0.43	77.2
UCE	**0.00**	**70.6**	0.45	75.2	0.50	71.9	0.52	108.5
SDD	**0.00**	71.6	**0.00**	194.2	**0.01**	81.4	0.44	74.7
Residual Consistency	**0.00**	71.0	**0.00**	**70.5**	0.08	**70.2**	**0.24**	**70.0**

5 Conclusion

We introduce *Late Horizon Steering*, a phenomenon we empirically observe in which steering is mainly performed towards the end of the generation process. We conjecture that prior methods fail on compositional tasks due to auxiliary value approximation, which prevents alignment with coarser objectives and exhibits Late Horizon Steering. Inspired by principles of stochastic optimal control and entropy-regularized reward alignment, we derive the state-residual SDE, a stochastic differential equation that jointly evolves the state and the optimal score residual. We design Residual Consistency, which tracks the optimal control without requiring auxiliary approximations. We empirically demonstrate that our proposed algorithm achieves state-of-the-art performance on a broad range of tasks.

Limitations: We derive the state-residual SDE and use the principles of stochastic optimal control for diffusion processes, which are not *directly* applicable to flow-based models.

AI USE STATEMENT

In this work, we used generative AI tools to aid in the literature survey, general formatting, and the implementation of simple methods. We have not used generative AI tools in developing theoretical models or conceptual frameworks, assist in the writing of proofs, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results and tasks such as generate synthetic data sets, clean and reformat dataset, support qualitative and thematic data analysis, assist with translation are not applicable to this work. Additionally, we used generative AI tools for polishing the draft. We have reviewed all AI-assisted work. The correctness of the generated code was manually verified by 2 authors, and the polished draft conveys the original intent. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Details for reproducing the results are provided in Appendix D, and the code and training checkpoints are available on an anonymous GitHub repository at https://anonymous.4open.science/r/Residual-Consistency-1732/README.md and will be deanonymized after the paper acceptance decision.

REFERENCES

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515. 3

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025. URL https://arxiv.org/abs/2511.21631. 8

Andreas Bergmeister, Stefanie Jegelka, Nikolas Nüsken, Carles Domingo-Enrich, and Jakiw Pidstrigach. Reinforce Adjoint Matching: Scaling RL Post-Training of Diffusion and Flow-Matching Models , 2026. URL https://arxiv.org/abs/2605.10759. 1, 3, 6, 16, 19

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/14f75513f0f1ca01de1e826b52e6b840-Paper-Conference.pdf. 1, 2, 15

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models, 2022. URL https://arxiv.org/abs/2204.00227. 1

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/13e8be77982beb73d7ed0bbf122f9f3c-Paper-Conference.pdf. 2, 15, 17

Carles Domingo-Enrich, Jiequn Han, Brandon Amos, Joan Bruna, and Ricky T. Q. Chen. Stochastic Optimal Control Matching. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet,

J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 112459–112504. Curran Associates, Inc., 2024. doi: 10.52202/079017-3573. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cc32ec39a5073f61d38c338d963df30d-Paper-Conference.pdf. 4

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53791–53846, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/852f50969a9e523ec41d26f2f68bd456-Paper-Conference.pdf. 1, 3, 16

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models . In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885. Curran Associates, Inc., 2023. doi: 10.52202/075280-3497. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fc65fab891d83433bd3c8d966edde311-Paper-Conference.pdf. 2, 15

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing Concepts from Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023. doi: 10.1109/ICCV51070.2023.00230. 3, 8, 16

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified Concept Editing in Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5111–5120, January 2024. 3, 8, 16

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment . In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023. doi: 10.52202/075280-2270. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Paper-Datasets_and_Benchmarks.pdf. 7

Jonathan Ho. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598. 1

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf. 1

Jerry Yuyang Huang, Justin Lin, Sheel Shah, Kartik Nair, and Nicholas Matthew Boffi. How to Guide Your Flow: Few-Step Alignment via Flow Map Reward Guidance. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gEogY6cPrL. 15

Kiyosi Itô. On a Formula Concerning Stochastic Differentials. *Nagoya Mathematical Journal*, 3: 55–65, 1951. doi: 10.1017/S0027763000012216. 5, 16

Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models , 2023. URL https://arxiv.org/abs/2307.05977. ICML 2023 Workshop on Challenges in Deployable Generative AI. 3, 8, 16

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time Alignment of Diffusion Models without Reward Over-optimization. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87238–87272, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d9042abf40782fbce28901c1c9c0e8d8-Paper-Conference.pdf. 15

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023. URL https://arxiv.org/abs/2210.10960. 1

Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer International Publishing, Springer Cham, 2016. ISBN 978-3-319-31088-6 978-3-319-31089-3. doi: 10.1007/978-3-319-31089-3. URL http://link.springer.com/10.1007/978-3-319-31089-3. 3

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755. Springer International Publishing, 2014. ISBN 978-3-319-10602-1. 8, 9

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 40783–40818. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1362. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3a10c46572628d58cb44fb705f25cbbf-Paper-Conference.pdf. 1, 2, 15

Zhen Liu, Tim Xiao, Carles Domingo-Enrich, Weiyang Liu, and Dinghuai Zhang. Value Gradient Guidance for Flow Matching Alignment. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 11372–11404. Curran Associates, Inc., 2025b. doi: 10.52202/085713-0380. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/10b7e27c8eb9571fbbd2ae6a9f8c3855-Paper-Conference.pdf. 1, 3, 4, 6, 16, 19

Zhen Liu, Tim Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient Diversity-Preserving Diffusion Alignment via Gradient-Informed GFlowNets . In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94806–94841, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ec52962a55a83c3d860b5c9b7338422b-Paper-Conference.pdf. 1, 2, 4, 6, 8, 15

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin, Heidelberg, 6 edition, 2003. ISBN 978-3-540-04758-2. doi: 10.1007/978-3-642-14394-6. 3, 20

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2303.11306. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html. 8

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model . In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf. 2, 15

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66377–66389. Curran Associates, Inc., 2023. doi: 10.52202/075280-2897. URL https://proceedings.neurips.cc/paper

_files/paper/2023/file/d0da30e312b75a3fffd9e9191f8bc1b0-Paper-Conference.pdf. 4

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-CD: Benchmarking Concept Removals for Text-to-image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28769–28778, June 2025. 8

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. 1, 6, 8

Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. *LAION. AI*, 3, 2022. 8

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models . In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294. Curran Associates, Inc., 2022. doi: 10.52202/068431-1833. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf. 1

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data . *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025. doi: 10.1073/pnas.2408799121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2408799121. 1, 4

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Gaming. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471. Curran Associates, Inc., 2022. doi: 10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3d719fee332caa23d5038b8a90e81796-Paper-Conference.pdf. 15

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. 1, 3

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations . In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS. 1, 3

Wenpin Tang and Fuzhong Zhou. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond , 2025. URL https://arxiv.org/abs/2403.06279. 4

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 , 2024. URL https://arxiv.org/abs/2402.15194. 3

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 07 2011. ISSN 0899-7667. doi: 10.1162/NECO_a_00142. URL https://doi.org/10.1162/NECO_a_00142. 3

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, 2023. URL https://arxiv.org/abs/2210.14896. 1

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 , 2023. URL https://arxiv.org/abs/2306.09341. 8

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient . In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504, 2025. doi: 10.1609/aaai.v39i8.32917. URL https://ojs.aaai.org/index.php/AAAI/article/view/32917. 3, 16

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 . In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf. 2, 8, 15

Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999. 3, 16

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. 2, 15

Haowei Zhang, Jianzhe Liu, Zhen Han, Shuo Chen, Bailan He, Volker Tresp, Zhiqiang Xu, and Jindong Gu. Visual question decomposition on multimodal large language models, 2024. URL https://arxiv.org/abs/2409.19339. 8

APPENDIX

This appendix is summarized as follows.

1. **Detailed Related Work**: In Appendix A, we provide a more detailed review of relevant literature.
2. **Proof of Optimal Residual PDE** (Appendix B): Derivation of Proposition 3.1.
3. **Proof of State-Residual SDE** (Appendix C): Derivation of Theorem 3.2.
4. **Additional Experimental Details** (Appendix D): Setup and Hyperparameters
5. **VP SDE Discretization** (Appendix E): Derivation of corollary 3.3.
6. **Derivation of the Prompt-Defined Optimal Control** (Appendix F): Derivation of the optimal residual Section 4.2
7. **Additional Optimal Control Recovery Plots** (Appendix G): Additional Plots for Section 4.2
8. **Human Preference Trade-off** (Appendix H) : Trade-off Plots for Human Preference Reward Section 4.4

A Additional Related Work

Given a scalar reward function $r\colon \mathbb{R}^d \to \mathbb{R}$, reward alignment aims to increase the expected reward of generated samples while preserving the pretrained distribution. A key challenge is assigning value to intermediate states to effectively guide generation towards high-reward regions. Prior methods explore various alignment objectives and how they propagate terminal reward information to intermediate generation states. Early approaches directly differentiated through sampling trajectories (Clark et al., 2024; Xu et al., 2023); however, such methods are diversity-lacking and prior-ignoring (Liu et al., 2025c), a subset of issues broadly classified under *reward hacking* (Skalse et al., 2022). Subsequent approaches consider entropy-regularized reward maximization to mitigate reward hacking via reinforcement learning-based policy optimization, GFlowNets, or optimal control methods. Another popular alternative is test-time alignment, a set of typically training-free algorithms that do not modify the original score model and instead modify state dynamics to maximize rewards using guidance (Kim et al., 2025; Huang et al., 2026). In this work, we focus on entropic regularized reward maximization with differentiable rewards.

Reinforcement Learning-based Policy Optimization Reinforcement Learning approaches view diffusion generation as a sequential decision process, with intermediate noisy samples as states and denoising transitions as actions. These works follow the parallel line of work in preference optimization in autoregressive modeling (Rafailov et al., 2023). DDPO (Black et al., 2024) optimizes terminal rewards based on the likelihoods of individual denoising transitions, while DPOK (Fan et al., 2023) explicitly incorporates KL regularization with respect to the pretrained model. More recent methods, such as Flow-GRPO (Liu et al., 2025a), use group-relative baselines, extend online policy optimization to flow-matching models, and have shown better performance among this category of algorithms. This class of techniques requires generating multiple trajectory rollouts to compute policy updates, thereby increasing the cost of these algorithms.

State Flows using GFlowNet GFlowNet approaches learn sequential samplers via balance conditions that characterize sampling that is proportional to a specified nonnegative terminal target. However, such a balance condition requires knowledge of GFlowNet *state flow*, which denotes the target mass assigned to an intermediate state. Diffusion alignment methods apply these conditions to denoising transitions (Zhang et al., 2025). The gradient-informed formulation of ∇-GFlowNets (Liu et al., 2025c) differentiates the balance condition. These methods require explicitly training a separate *state-flow network*, which, together with the model's transition kernel, is jointly trained to enforce the detailed balance condition. This auxiliary network requires separate hyperparameter selection and incurs additional training costs, in addition to the approximation introduced by this parameterization.

Reward Alignment with Optimal Control Optimal-control approaches seek to characterize the optimal dynamics for an entropy-regularized reward objective using principles of stochastic optimal control. Adjoint Matching (Domingo-Enrich et al., 2025) characterizes optimality through a least-squares regression objective called the lean adjoint by addressing initial value bias. Although this simplifies conventional adjoint computation, constructing the targets still requires a backward adjoint propagation in addition to trajectory generation, which can be computationally expensive (Bergmeister et al., 2026). Recent methods investigate solving the optimal control problem without backward adjoint propagation. VGGFlow (Liu et al., 2025b) takes a deterministic optimal-control perspective. It learns an auxiliary value-gradient field via consistency with the Hamilton-Jacobi-Bellman (HJB) equation (Yong & Zhou, 1999), which governs the optimal dynamics of value evolution over intermediate time steps, and then trains the velocity field to match that estimate. RAM (Bergmeister et al., 2026) constructs inexpensive regression targets from generated endpoints and analytically noised intermediate states, avoiding SDE rollouts, backward adjoint propagation, and reward gradients. To achieve this efficiency, the practical estimator omits the cumulative control-cost term from the exact value-gradient expression. RAM therefore optimizes an approximation to the full stochastic-control update. These techniques provide principled and high-performing algorithms but suffer from expensive trajectory rollouts and approximate computation of the value function using parameterized neural networks.

Concept Unlearning for Diffusion Models A parallel field of study in post-training diffusion models is Concept Unlearning: removing the model's ability to generate harmful or copyrighted images (Gandikota et al., 2023). State-of-the-art methods address this problem by replacing the unwanted concept with either a null context or a safe concept (Gandikota et al., 2024). The key challenge is to reduce generations in the unwanted concept while preserving the generation capability on safe concepts. Common techniques incorporate this tradeoff using an explicit loss formulation (Wu et al., 2025; Kim et al., 2023). In our work, we attempt to address concept unlearning using the reward alignment framework.

B Proof of Proposition 3.1

Proposition 3.1 (Optimal Residual PDE)**.** *Assume that V and b are sufficiently smooth. Then $h^\star = \lambda^{-1}\nabla_x V$ is a solution of*

$$-\partial_t h = J_h\big(b + g^2 h\big) + \frac{g^2}{2}\Delta h + J_b^\top h, \quad h(x,1) = \lambda^{-1}\nabla_x r(x) \tag{ResPDE}$$

where J_v denotes the Jacobian of a vector field v.

Proof. Recall from Equation (OptRes) that $h^\star = \lambda^{-1}\nabla_x V$. Differentiating Equation (HJB) with respect to x gives

$$-\partial_t \nabla_x V = J_b^\top \nabla_x V + (\nabla_x^2 V)b + \frac{g^2}{2}\Delta(\nabla_x V) + \frac{g^2}{\lambda}(\nabla_x^2 V)\nabla_x V, \tag{1}$$

where we use that g depends only on time and that spatial and temporal derivatives commute under the assumed smoothness. Substituting $\nabla_x V = \lambda h^\star$ and $\nabla_x^2 V = \lambda J_{h^\star}$, then dividing by λ, yields

$$-\partial_t h^\star = J_b^\top h^\star + J_{h^\star} b + \frac{g^2}{2}\Delta h^\star + g^2 J_{h^\star} h^\star \tag{2}$$

Here $J_{h^\star} = \lambda^{-1}\nabla_x^2 V$ is symmetric because it is a scaled Hessian. Rearranging gives

$$-\partial_t h^\star = J_{h^\star}\big(b + g^2 h^\star\big) + \frac{g^2}{2}\Delta h^\star + J_b^\top h^\star.$$

Finally, differentiating the terminal condition $V(x,1) = r(x)$ gives $h^\star(x,1) = \lambda^{-1}\nabla_x r(x)$. □

C Proof of Theorem 3.2

We first state Itô's Lemma (Itô, 1951) for convenience.

Lemma C.1 (Multivariate Itô's Lemma). *Let $\mathbf{x}_t \in \mathbb{R}^d$ be a stochastic process satisfying*

$$d\mathbf{x}_t = \mu(\mathbf{x}_t, t)\, dt + g(t)\, d\mathrm{W}_t, \tag{3}$$

where $\mu : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is the drift, $g(t) \in \mathbb{R}$ is the scalar diffusion coefficient, and W_t is a standard d-dimensional Wiener process. Then for any smooth vector-valued function $h : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, the stochastic differential of $h(\mathbf{x}_t, t)$ is given by:

$$dh(\mathbf{x}_t, t) = \left(\frac{\partial h}{\partial t}(\mathbf{x}_t, t) + J_h(\mathbf{x}_t, t)\, \mu(\mathbf{x}_t, t) + \frac{1}{2} g^2(t)\, \Delta_{\mathbf{x}} h(\mathbf{x}_t, t) \right) dt + g(t)\, J_h(\mathbf{x}_t, t)\, d\mathrm{W}_t, \tag{4}$$

where $J_h(\mathbf{x}_t, t) = \nabla_{\mathbf{x}} h(\mathbf{x}_t, t) \in \mathbb{R}^{d \times d}$ is the spatial Jacobian of h and $\Delta_{\mathbf{x}} h$ denotes the componentwise Laplacian.

Theorem 3.2 (State-Residual SDE). *Under standard regularity assumptions, the optimal state and residual $(\mathrm{X}_t^\star, h^\star)$ of Equation (P1) satisfy*

$$d \begin{bmatrix} \mathrm{X}_t^\star \\ h^\star(\mathrm{X}_t^\star, t) \end{bmatrix} = \begin{bmatrix} b + g^2 h^\star \\ -J_b^\top h^\star \end{bmatrix} (\mathrm{X}_t^\star, t)\, dt + g(t) \begin{bmatrix} I_d \\ J_{h^\star}(\mathrm{X}_t^\star, t) \end{bmatrix} d\mathrm{W}_t, \quad \mathrm{X}_0^\star \sim \mathcal{N}(0, I) \tag{SR-SDE}$$

Proof of Theorem 3.2. By Optimal Residual PDE Equation (ResPDE), we have

$$-\partial_t h^\star = J_{h^\star}\left(b + g^2 h^\star\right) + \frac{g^2}{2} \Delta h^\star + J_b^\top h^\star. \tag{5}$$

Applying Itô's Lemma C.1 to the optimal residual $h^\star$ along the controlled SDE Equation (ResCtrl) gives

$$\begin{aligned} dh^\star(X_t^\star, t) = & \underbrace{\left[\partial_t h^\star + J_{h^\star}\left(b + g^2 h^\star\right) + \frac{g^2}{2} \Delta h^\star \right](X_t^\star, t)\, dt}_{\text{drift}} \\ & + g(t) J_{h^\star}(X_t^\star, t)\, dW_t, \end{aligned} \tag{6}$$

where $\Delta h^\star$ denotes the componentwise Laplacian. Substituting Equation (5) into drift of Equation (6) gives

$$\begin{aligned} \text{drift} &= \partial_t h^\star + J_{h^\star}\left(b + g^2 h^\star\right) + \frac{g^2}{2} \Delta h^\star \\ &= \left(\cancel{-J_{h^\star}\left(b + g^2 h^\star\right)} - \cancel{\frac{g^2}{2} \Delta h^\star} - J_b^\top h^\star \right) + \cancel{J_{h^\star}\left(b + g^2 h^\star\right)} + \cancel{\frac{g^2}{2} \Delta h^\star}. \end{aligned}$$

yielding

$$dh^\star(X_t^\star, t) = \left[-J_b^\top h^\star \right](X_t^\star, t)\, dt + g(t) J_{h^\star}(X_t^\star, t)\, dW_t.$$

Stacking with Equation (P1) yields Equation (SR-SDE). The same Wiener increment drives both components because the control process is obtained by evaluating $h^\star$ along the state process $X_t^\star$ and applying Itô's Lemma.

The initial state distribution is inherited from Equation (P1): $X_0^\star \sim \mathcal{N}(0, I)$. □

D Additional Experimental Details

All experiments were run on 4 NVIDIA RTX 6000 Ada GPUs. Additional environment details and instructions to reproduce the results are provided in the anonymous GitHub at https://anonymous.4open.science/r/Residual-Consistency-1732/README.md. We implement the reward scale $\gamma = \lambda^{-1}$ inverse-temperature parameter. Table 4 reports the remaining settings; batch size is the same for sampling and training. In Algorithm 2 we describe the simple SDE reverse sampler. In Table 6 we include preference alignment comparison with (Clark et al., 2024).

Algorithm 2 REVERSETRAJECTORY$(X_0, \{\beta_k\}_{k=0}^{N-1}, \{t_k\}_{k=0}^{N-1}, \{z_k\}_{k=0}^{N-1}, s^\theta)$

Require: Initial state $X_0 \in \mathbb{R}^d$, step sizes $\{\beta_k\}_{k=0}^{N-1}$, time grid $\{t_k\}_{k=0}^{N-1}$, noise $\{z_k\}_{k=0}^{N-1} \subset \mathbb{R}^d$, score network s^θ
Ensure: Sample $\{X_k\}_{k=1}^{N}$
1: **for** $k = 0, \ldots, N-1$ **do**
2: $\quad X_{k+1} \leftarrow \left(1 + \frac{1}{2}\beta_k\right) X_k + \beta_k s^\theta(X_k, t_k) + \sqrt{\beta_k}\, z_k$
3: **end for**
4: **return** $\{X_k\}_{k=1}^{N}$

Table 4: Experimental hyperparameters. CFG denotes classifier-free guidance.

Experiment	Reward scale γ	Learning rate	LoRA rank	LoRA α	CFG scale	Batch size	Sampling steps
Residual recovery (SD 1.5)	1	1×10^{-4}	8	8	1	4	50
Composition (SD 1.5)	100	1×10^{-4}	32	32	2	4	50
Composition (SDXL)	100	1×10^{-4}	32	32	2	1	20
Human preference (Aesthetic)	30	1×10^{-3}	8	8	5	4	50
Human preference (HPS v2)	1,000	1×10^{-3}	8	8	5	4	50
Human preference (ImageReward)	100	1×10^{-3}	8	8	5	4	50
Concept unlearning	1,000	3×10^{-4}	8	8	2	4	50

Table 5: Training cost for the compositional alignment experiments in Section 4.3. We report GPU-hours and peak GPU memory for SDXL and SD1.5.

	SDXL		SD1.5	
Method	**GPU-hours**	**Peak memory (GB)**	**GPU-hours**	**Peak memory (GB)**
∇-GFlowNet	23	39.9	12	35.2
VGGFlow	20	46	16.6	39.4
RAM	16	45.6	13	30
Residual Consistency	16.7	41.8	13.3	37.4

Table 6: Quantitative comparison of alignment methods across three reward functions. Higher reward and diversity are better, while lower FID is better.

	Aesthetic Score			HPSv2			ImageReward		
Method	**Reward** $\uparrow$	**Div.** $\uparrow$ (10^{-2})	**FID** $\downarrow$	**Reward** $\uparrow$ (10^{-1})	**Div.** $\uparrow$ (10^{-2})	**FID** $\downarrow$	**Reward** $\uparrow$ (10^{-1})	**Div.** $\uparrow$ (10^{-2})	**FID** $\downarrow$
Base Model (SD1.5)	5.92 ± 0.4	33.13	68.72	2.50 ± 0.4	37.39	68.72	-0.1 ± 10.2	45.25	68.72
DRaFT	$\mathbf{8.60 \pm 0.3}$	16.60	82.21	$\mathbf{3.67 \pm 0.4}$	16.64	89.75	1.64 ± 10.7	34.75	73.43
∇-GFlowNet	7.75 ± 0.4	25.26	**72.87**	3.33 ± 0.5	29.38	**71.88**	2.24 ± 10.1	**39.26**	69.51
VGGFlow	7.58 ± 0.3	26.33	73.92	3.33 ± 0.4	29.23	73.33	2.27 ± 10.6	39.22	68.99
Residual Consistency	7.78 ± 0.2	**27.31**	74.14	3.41 ± 0.4	**29.51**	73.27	$\mathbf{3.86 \pm 10.8}$	38.01	**68.23**

D.1 Adapting RAM and VGGFlow

RAM (Bergmeister et al., 2026) and VGGFlow (Liu et al., 2025b) report results on rectified-flow models, whereas our experiments use the denoising diffusion models SD 1.5 and SDXL. We adapt both methods through the probability-flow ODE of the variance-preserving (VP) process. We use the time convention ($t = 1$) to denote the noise endpoint and ($t = 0$) to denote the data endpoint. The noising process is

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

For rectified flow, $\alpha_t = 1 - t$ and $\sigma_t = t$, and the model directly predicts the velocity $v_\theta(x_t, t)$. For the VP process,

$$\alpha_t = \exp\left(-\frac{1}{2}\int_0^t \beta(s)\, ds\right), \qquad \sigma_t = \sqrt{1 - \alpha_t^2}.$$

The corresponding probability-flow velocity can be computed from the diffusion model's noise prediction ϵ_θ:

$$\frac{dx_t}{dt} = v_\theta^{\mathrm{VP}}(x_t, t) = -\frac{\beta(t)}{2} x_t + \frac{\beta(t)}{2\sigma_t} \epsilon_\theta(x_t, t).$$

We use this velocity and the VP schedule coefficients in place of their rectified-flow counterparts when training RAM and VGGFlow.

E Derivation of corollary 3.3

Corollary 3.3 (Discrete VP residual-consistency updates)**.** *Let* $0 = t_0 < \cdots < t_N = 1$. *For* $k \in \{0, \ldots, N-1\}$, *define*

$$\delta_k = t_{k+1} - t_k, \qquad \hat{\mathrm{X}}_k^\star = \mathrm{X}_{t_k}^\star, \qquad \hat{h}_k^\star = h^\star(\hat{\mathrm{X}}_k^\star, t_k), \qquad \beta_k = \beta(t_k)\delta_k.$$

Applying Euler-Maruyama discretization on Equation (SR-SDE) *gives one-step updates as*

$$\hat{\mathrm{X}}_{k+1}^\star = \left(1 + \frac{1}{2}\beta_k\right) \hat{\mathrm{X}}_k^\star + \beta_k \left[s(\hat{\mathrm{X}}_k^\star, t_k) + \hat{h}_k^\star\right] + \sqrt{\beta_k}\, z_k, \qquad \text{(VP-State)}$$

$$\hat{h}_{k+1}^\star = \left(1 - \frac{1}{2}\beta_k\right) \hat{h}_k^\star - \beta_k J_s(\hat{\mathrm{X}}_k^\star, t_k)^\top \hat{h}_k^\star + \sqrt{\beta_k} J_{\hat{h}^\star}(\hat{\mathrm{X}}_k^\star, t_k) z_k, \qquad \text{(VP-Res)}$$

Proof. For the VP generation process, $b = \frac{1}{2}\beta x + \beta s$ and $g^2 = \beta$, so $J_b = \frac{1}{2}\beta I_d + \beta J_s$. Substituting these identities directly into Equation (SR-SDE) we get

$$d\mathrm{X}_t^\star = \beta(t)\left[\frac{1}{2}\mathrm{X}_t^\star + s(\mathrm{X}_t^\star, t) + h^\star(\mathrm{X}_t^\star, t)\right] dt + \sqrt{\beta(t)} dW_t, \tag{7}$$

$$dh^\star(\mathrm{X}_t^\star, t) = -\beta(t)\left[\frac{1}{2} I_d + J_s(\mathrm{X}_t^\star, t)^\top\right] h^\star(\mathrm{X}_t^\star, t) dt + \sqrt{\beta(t)} J_{h^\star}(\mathrm{X}_t^\star, t) dW_t. \tag{8}$$

For the interval $[t_k, t_{k+1}]$, write

$$W_{t_{k+1}} - W_{t_k} = \sqrt{\delta_k}\, z_k, \qquad z_k \sim \mathcal{N}(0, I_d).$$

Applying Euler-Maruyama to the coupled system, using the same Brownian increment in both equations, yields

$$\hat{\mathrm{X}}_{k+1}^\star = \hat{\mathrm{X}}_k^\star + \beta(t_k)\delta_k \left[\frac{1}{2}\hat{\mathrm{X}}_k^\star + s(\hat{\mathrm{X}}_k^\star, t_k) + \hat{h}_k^\star\right] + \sqrt{\beta(t_k)\delta_k} z_k,$$

$$\hat{h}_{k+1}^\star = \hat{h}_k^\star - \beta(t_k)\delta_k \left[\frac{1}{2} I_d + J_s(\hat{\mathrm{X}}_k^\star, t_k)^\top\right] \hat{h}_k^\star + \sqrt{\beta(t_k)\delta_k} J_{\hat{h}^\star}(\hat{\mathrm{X}}_k^\star, t_k) z_k.$$

Finally, using $\beta_k = \beta(t_k)\delta_k$ and collecting terms completes the proof. □

F Derivation of the Prompt-Defined Optimal Control

We derive the target used in Section 4.2. Let p_1^{cat} and p_1^{dog} denote the clean-image distributions represented by the two conditional branches of the same frozen diffusion model. Let $q_t(x_t \mid x_1)$ denote their shared forward noising kernel and define the corresponding noisy marginals by

$$p_t^c(x_t) = \int q_t(x_t \mid x_1) p_1^c(x_1)\, dx_1, \qquad c \in \{\text{cat}, \text{dog}\}.$$

We take the cat-conditioned process as the base process and use the terminal reward

$$r(x_1) = \log \frac{p_1^{\text{dog}}(x_1)}{p_1^{\text{cat}}(x_1)}.$$

For the entropy-regularized control problem in Equation (P1) with $\lambda = 1$, the value function can be written as

$$V(x_t, t) = \log \mathbb{E}_{\mathbb{P}^{\text{cat}}}[\exp(r(\mathrm{X}_1)) \mid \mathrm{X}_t = x_t].$$

Substituting the likelihood-ratio reward and applying Bayes' rule to the cat-conditioned posterior, the expectation inside the logarithm becomes

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^{\text{cat}}}[\exp(r(X_1))|X_t = x_t] &= \int \frac{p_1^{\text{dog}}(x_1)}{p_1^{\text{cat}}(x_1)} p^{\text{cat}}(x_1 \mid x_t)\, dx_1 \\
&= \int \frac{p_1^{\text{dog}}(x_1)}{p_1^{\text{cat}}(x_1)} \frac{q_t(x_t \mid x_1) p_1^{\text{cat}}(x_1)}{p_t^{\text{cat}}(x_t)} dx_1 \\
&= \frac{1}{p_t^{\text{cat}}(x_t)} \int q_t(x_t \mid x_1) p_1^{\text{dog}}(x_1)\, dx_1 \\
&= \frac{p_t^{\text{dog}}(x_t)}{p_t^{\text{cat}}(x_t)} \int \frac{q_t(x_t \mid x_1) p_1^{\text{dog}}(x_1)}{p_t^{\text{dog}}(x_t)}\, dx_1 \\
&= \frac{p_t^{\text{dog}}(x_t)}{p_t^{\text{cat}}(x_t)} \int p_t^{\text{dog}}(x_1 \mid x_t) dx_1 \\
&= \frac{p_t^{\text{dog}}(x_t)}{p_t^{\text{cat}}(x_t)}.
\end{aligned}$$

Hence

$$V(x_t, t) = \log p_t^{\text{dog}}(x_t) - \log p_t^{\text{cat}}(x_t).$$

Hence, the value gradient during training is the difference in scores at intermediate timesteps. Using the optimal-control identity in Equation (OptCtrl), we obtain

$$\begin{aligned}
u^\star(x_t, t) &= g(t) \nabla_{x_t} V(x_t, t) \\
&= g(t) \left[\nabla_{x_t} \log p_t^{\text{dog}}(x_t) - \nabla_{x_t} \log p_t^{\text{cat}}(x_t) \right] \\
&= g(t) \left[s_{\text{base}}(x_t, t \mid \text{dog}) - s_{\text{base}}(x_t, t \mid \text{cat}) \right],
\end{aligned}$$

Because both prompt distributions are represented by the same pretrained model and share its forward noising process, the optimal control is the target presented in the main body.

G Additional Optimal Control Recovery Plots

We compare each learned control with the analytically optimal control throughout the diffusion trajectory. Figure 5 reports three complementary measures. Residual Consistency has the lowest cumulative control discrepancy, which is proportional to the path-space KL divergence by Girsanov's theorem (Øksendal, 2003) and the lowest timestep-wise mean squared error (MSE). Its norm ratio with optimal control is also substantially higher across intermediate time steps.

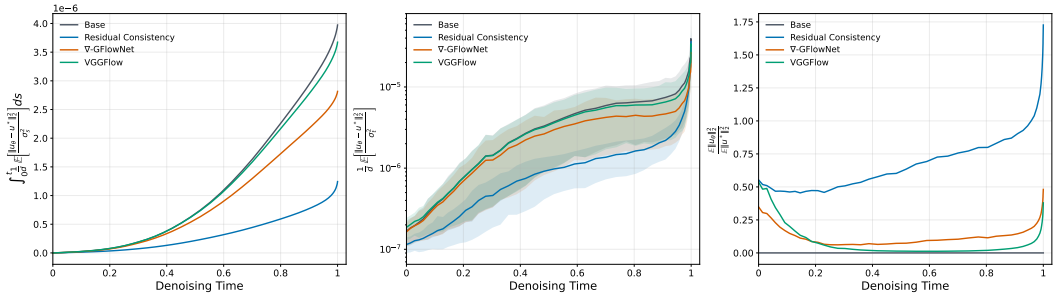

Figure 5: Left: Path-Space KL divergence over time. Center: MSE between the learned and optimal controls at each diffusion timestep, shown on a logarithmic scale. Right: norm ratio of learned control and optimal control.

H Human Preference Trade-off

We provide reward-versus-diversity, diversity-versus-FID, and FID-versus-reward plots for LAION-Aesthetic, HPS v2, and ImageReward. In every plot, the desired region is the upper-right, corresponding to higher reward, DreamSim diversity, and lower CleanFID. The FID axes decrease from left to right.

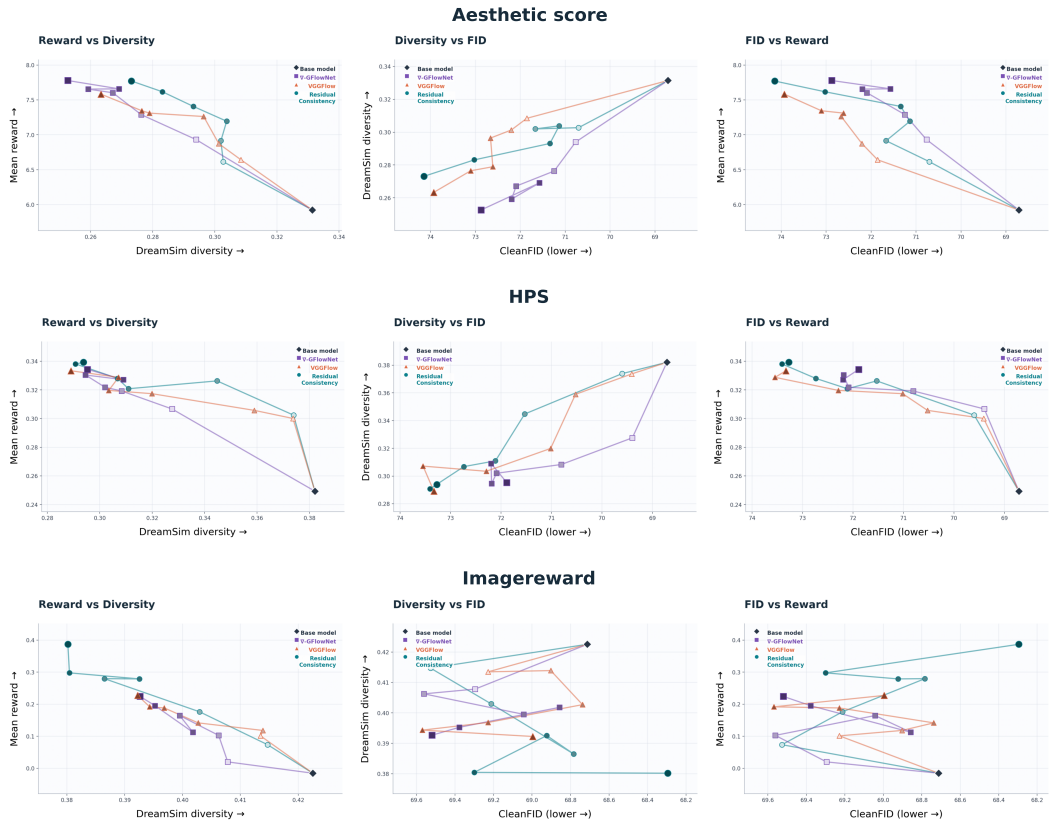

Figure 6: Human preference trade-off for LAION-Aesthetic (top), HPS v2 (middle), and ImageReward (bottom). Colored trajectories connect optimization checkpoints for ∇-GFlowNet, VGGFlow, and Residual Consistency; the black diamond marks the base model.