Supplementary Materials of Neighbor GRPO

1. Algorithm

We provide the pseudo-code for Neighbor GRPO in Algorithm 1. We suggest adopting a more efficient solver, such as DPM-Solver++, with fewer sampling timesteps T for the $\mathrm{RolloutSolver}$ to accelerate training. Compared with SDE-based GRPO methods, which call $\mathrm{TrainSolver}$ $G \times K$ times in each GRPO iteration, our approach requires $B \times K$ calls to $\mathrm{TrainSolver}$. Note that we usually have $B < G$, which allows us to further reduce the training cost with Neighbor GRPO.

2. Derivation of the SDE Formulation

For a Rectified Flow model, the forward process from data x_0 to noise ϵ is $x_t = (1-t)x_0 + t\epsilon$. The reverse process follows the ODE $\frac{dx_t}{dt} = v_\theta(x_t, t)$, where $v_\theta(x_t, t) \approx \epsilon - x_0$.

To introduce stochasticity, we construct an equivalent SDE that shares the same marginal probability density $p_t(x)$ for all $t \in [0, 1]$. A common choice is:

$$dx_t = \left[v_\theta(x_t, t) + \frac{\eta_t^2}{2t} \hat{\epsilon}_\theta(x_t, t) \right] dt + \eta_t dw_t, \qquad (1)$$

where dw_t is a standard Wiener process, η_t controls the noise intensity, and $\hat{\epsilon}_\theta(x_t, t) = x_t + (1-t)v_\theta(x_t, t)$ is the estimated initial noise.

For numerical simulation, we discretize the reverse SDE. With a backward step $\Delta t > 0$, the update rule is:

$$x_{t-\Delta t} = x_t - \left[v_\theta(x_t, t) + \frac{\eta_t^2}{2t} \hat{\epsilon}_\theta(x_t, t) \right] \Delta t + \eta_t \sqrt{\Delta t} \epsilon, \qquad (2)$$

where $\epsilon \sim \mathcal{N}(0, I)$. Let's define the single-step ODE update as $x_{t-\Delta t}^{(\mathrm{ODE})} = x_t - v_\theta(x_t, t)\Delta t$. Let $\sigma_t^2 = \eta_t^2 \Delta t$. The SDE update can be rewritten as:

$$x_{t-\Delta t}^{(\mathrm{SDE})} = x_t - v_\theta(x_t, t)\Delta t - \frac{\sigma_t^2}{2t\Delta t} \hat{\epsilon}_\theta(x_t, t)\Delta t + \sigma_t \epsilon \qquad (3)$$

$$= x_{t-\Delta t}^{(\mathrm{ODE})} - \frac{\sigma_t^2}{2t} \hat{\epsilon}_\theta(x_t, t) + \sigma_t \epsilon. \qquad (4)$$

This discretized update defines a Gaussian policy $\pi_\theta(x_{t-\Delta t} | x_t, c) = \mathcal{N}(\mu_\theta, \sigma_t^2 I)$, where the mean is $\mu_\theta(x_t, t) = x_{t-\Delta t}^{(\mathrm{ODE})} - \frac{\sigma_t^2}{2t} \hat{\epsilon}_\theta(x_t, t)$. The negative log-likelihood of this policy is:

$$-\log \pi_\theta(x_{t-\Delta t} \mid x_t, c) = \frac{1}{2\sigma_t^2} \|x_{t-\Delta t} - \mu_\theta(x_t, t)\|_2^2 + \text{const.} \qquad (5)$$

Algorithm 1 Neighbor GRPO

Require: prompts C, parameters θ; group size G; anchors per batch B; number of rollout steps T; number of train steps K; time schedule $\mathcal{T} = \{t_1, t_2, \ldots, t_T\}$
Ensure: $t_T = 1, t_1 = 0$
for iteration $m = 1$ to M **do**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Sample prompt $c \in C$
  Draw shared initial noise $\epsilon^* \sim \mathcal{N}(0, I)$
  for $i = 1$ to G **do**
    Draw $\delta^{(i)} \sim \mathcal{N}(0, I)$
    $x_1^{(i)} \leftarrow \sqrt{1-\sigma^2}\epsilon^* + \sigma\delta^{(i)}$ ▷ Perturb noise.
    for $k = T$ to 1 **do**
      $x_{t_{k-1}}^{(i)} \leftarrow \mathrm{RolloutSolver.step}\big(x_{t_k}^{(i)}, c; \theta_{\text{old}}\big)$ ▷ Stop gradient.
    end for
  end for
  Compute advantages $\{A_i\}$
  Sample B anchor indices $S \subset \{1, \ldots, G\}$
  Sample K steps $\mathcal{K} \subset \{1, 2, \ldots, T-1\}$
  for $j \in S$ **do**
    for $k \in \mathcal{K}$ **do**
      $x_{t_{k+1}}^{(\theta)} \leftarrow x_{t_{k+1}}^{(j)}$
      $x_{t_k}^{(\theta)} \leftarrow \mathrm{TrainSolver.step}\big(x_{t_{k+1}}^{(\theta)}, c; \theta\big)$
      for $i = 1$ to G **do**
        $\pi_\theta^{(i)} \leftarrow \mathrm{softmax}_i(-\|x_{t_k}^{(i)} - x_{t_k}^{(\theta)}\|_2^2)$
        $\pi_{\theta_{\text{old}}}^{(i)} \leftarrow \mathrm{softmax}_i\big(-\|x_{t_k}^{(i)} - x_{t_k}^{(j)}\|_2^2\big)$
        $\rho^{(i|j)} \leftarrow \pi_\theta^{(i)} / \pi_{\theta_{\text{old}}}^{(i)}$
      end for
      Compute $J = J(A, \rho; \theta)$
      Accumulate gradients $\nabla_\theta J$
    end for
    Update θ using accumulated gradients
  end for
end for

This provides the basis for the distance-based contrastive learning interpretation of SDE-based GRPO.

3. Derivation of the Drift Residual

The objective of GRPO is to maximize the expected log-likelihood of samples from the old policy $\pi_{\theta_{\text{old}}}$ under the new policy π_θ. This is equivalent to minimizing the negative log-likelihood:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_{t-\Delta t}\sim\pi_{\theta_{\text{old}}}}[-\log \pi_\theta(x_{t-\Delta t})]. \tag{6}$$

From the SDE formulation in the previous section, the policy π_θ is an isotropic Gaussian distribution $\pi_\theta(x_{t-\Delta t}|x_t) = \mathcal{N}(\mu_\theta, \sigma_t^2 I)$. The mean μ_θ and a sample from the old policy $x_{t-\Delta t}^{(\text{SDE,old})}$ are given by:

$$\mu_\theta = x_{t-\Delta t}^{(\text{ODE},\theta)} - \frac{\sigma_t^2}{2t}\hat{\epsilon}_\theta(x_t, t), \tag{7}$$

$$x_{t-\Delta t}^{(\text{SDE,old})} = x_{t-\Delta t}^{(\text{ODE,old})} - \frac{\sigma_t^2}{2t}\hat{\epsilon}_{\theta_{\text{old}}}(x_t, t) + \sigma_t\epsilon, \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, I)$, and the superscripts denote the policy parameters used for the ODE step and noise prediction.

The negative log-likelihood term, omitting constants, is a mean squared error (MSE):

$$-\log \pi_\theta(x_{t-\Delta t}^{(\text{SDE,old})}) = \frac{1}{2\sigma_t^2}\left\|x_{t-\Delta t}^{(\text{SDE,old})} - \mu_\theta\right\|_2^2. \tag{9}$$

Substituting the expressions for the mean and the sample, we get the term inside the norm:

$$\begin{aligned} &x_{t-\Delta t}^{(\text{SDE,old})} - \mu_\theta \\ &= \left(x_{t-\Delta t}^{(\text{ODE,old})} - \frac{\sigma_t^2}{2t}\hat{\epsilon}_{\theta_{\text{old}}} + \sigma_t\epsilon\right) - \left(x_{t-\Delta t}^{(\text{ODE},\theta)} - \frac{\sigma_t^2}{2t}\hat{\epsilon}_\theta\right) \end{aligned} \tag{10}$$

$$= \left(x_{t-\Delta t}^{(\text{ODE,old})} + \sigma_t\epsilon\right) - x_{t-\Delta t}^{(\text{ODE},\theta)} - \left(\frac{\sigma_t^2}{2t}\hat{\epsilon}_{\theta_{\text{old}}} - \frac{\sigma_t^2}{2t}\hat{\epsilon}_\theta\right). \tag{11}$$

The main paper presents this MSE term in a simplified form:

$$\frac{1}{2\sigma_t^2}\left\|\tilde{x}_{t-\Delta t}^{(\text{ODE})} - x_{t-\Delta t}^{(\text{ODE})} + o_t(x_t)\right\|_2^2, \tag{12}$$

where $\tilde{x}_{t-\Delta t}^{(\text{ODE})} = x_{t-\Delta t}^{(\text{ODE,old})} + \sigma_t\epsilon$ is the perturbed ODE sample from the old policy, and $x_{t-\Delta t}^{(\text{ODE})}$ represents the ODE step from the new policy.

By comparing the two forms, we can identify the drift residual term $o_t(x_t)$:

$$o_t(x_t) = \frac{\sigma_t^2}{2t}\hat{\epsilon}_\theta(x_t, t) - \frac{\sigma_t^2}{2t}\hat{\epsilon}_{\theta_{\text{old}}}(x_t, t). \tag{13}$$

This term represents the drift correction between the old and new policies. As the policy update is constrained within a trust region, $\theta \approx \theta_{\text{old}}$, and this residual term approaches zero.

4. Marginal Distribution Preservation

Consider the standard flow matching framework with marginal distribution $p_\theta(x_t)$ at timestep $t \in [0, 1]$, induced by the deterministic ODE flow $\Phi_{1\to t}(\cdot;\theta)$ starting from initial noise x_1:

$$p_\theta(x_t) = \mathbb{E}_{x_1\sim\mathcal{N}(0;I)}\left[\delta(x_t - \Phi_{1\to t}(x_1;\theta))\right], \tag{14}$$

where $\delta(\cdot)$ is the Dirac delta function. We define a leaping probability $\pi(x_t|x_t') = \mathcal{N}(x_t|x_t', \sigma_\pi I)$, denoting the probability that any reachable latent sample $x_t' = \Phi_{1\to t}(x_1';\theta)$ (derived from an initial noise x_1') leaps to x_t. Following this distribution, the marginal distribution of x_t leaped from any x_t' is

$$\pi(x_t) = \int p_\theta(x_t')\pi(x_t|x_t')dx_t'. \tag{15}$$

Given a sufficiently small σ_π (or as $\sigma_\pi \to 0$), the support of $\pi(x_t|x_t')$ collapses to a single point where $x_t' = x_t$. Consequently, $\pi(x_t) \to p_\theta(x_t)$.

Now consider a Monte-Carlo approximation of $\pi(x_t)$:

$$\hat{\pi}(x_t) = \frac{1}{G}\sum_{i=1}^{G}\pi(x_t|x_t^{(i)}), \quad x_t^{(i)} \sim p_\theta(x_t). \tag{16}$$

Let $\hat{\pi}(x_t^{(i)}) = \frac{1}{G}$. We have:

$$\hat{\pi}(x_t^{(i)}|x_t) = \frac{\hat{\pi}(x_t^{(i)})\hat{\pi}(x_t \mid x_t^{(i)})}{\hat{\pi}(x_t)} \tag{17}$$

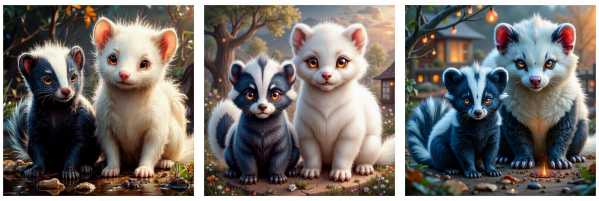
$$= \frac{\pi(x_t|x_t^{(i)})}{\sum_{j=1}^{G}\pi(x_t|x_t^{(j)})} \tag{18}$$

$$= \frac{\exp(-\|x_t - x_t^{(i)}\|_2^2/2\sigma_\pi^2)}{\sum_j \exp(-\|x_t - x_t^{(j)}\|_2^2/2\sigma_\pi^2)}. \tag{19}$$

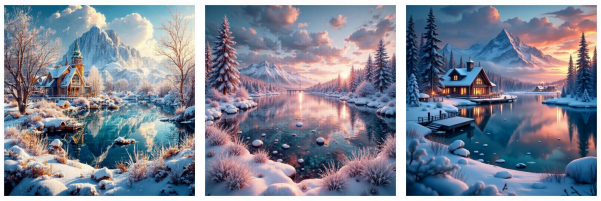
This formulation results in a softmax distribution with a temperature of $2\sigma_\pi^2$. In practice, we omit this temperature parameter because the Euclidean distance $\|x_t - x_t^{(i)}\|_2^2$ is empirically large enough when $i \neq j$. Replacing x_t with $x_t^{(\theta)}$ yields the surrogate leaping policy presented in the main paper. Maximizing this posterior probability $\hat{\pi}(x_t^{(i)}|x_t^{(\theta)})$ encourages the generated sample to align closely with the support of the empirical marginal distribution defined by $\{x_t^{(i)}\}_{i=1}^G$. This serves as a tractable surrogate objective for marginal distribution preservation.

5. More Qualitative Results

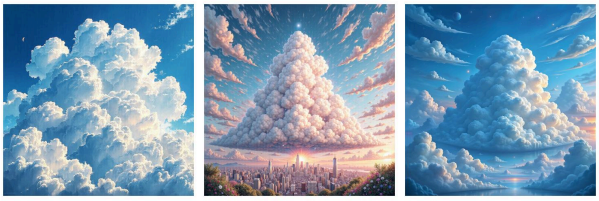
We present more visualization results comparing different methods in Figure 1.



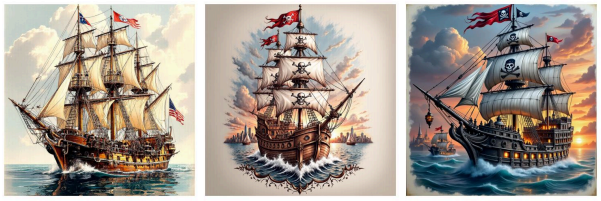
Two skunks - one small and dark blue with yellow eyes and a larger albino one with red eyes.



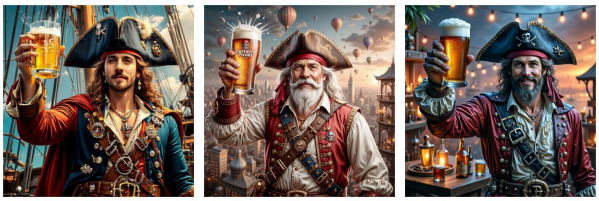
A snowy lake in Sweden captured in a vibrant, cinematic style with intense detail and raytracing technology showcased on Artstation.



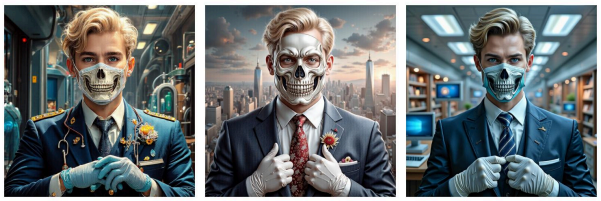
Anime-style vector illustration of a cloudy sky with a unique perspective.



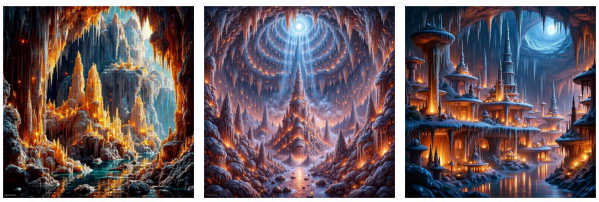
Realism tattoo design sketch of a pirate ship.



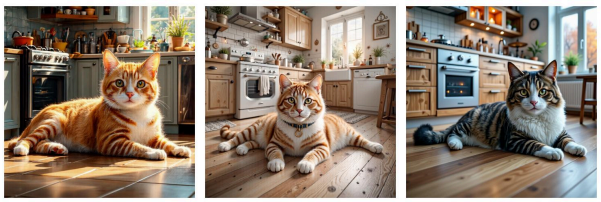
The image shows a pirate holding up a beer in celebration.



A blond person wearing a suit, medical gloves and a skull face mask is shown in a frontal portrait by Kim Kyoung Hwan.

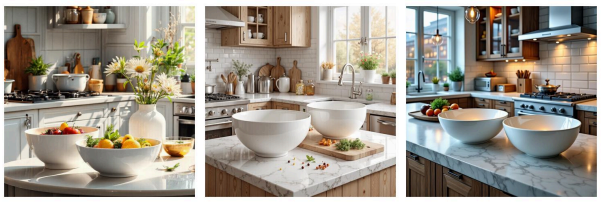


The image is of a massive cave interior filled with glowing stalactites and stalagmites.



A cat laying on the floor of a kitchen.

A portrait of a woman with a paper bag over her head.



A kitchen counter top with a white bowl sitting next to another white bowl.

Figure 1. Visualization of different approaches (DanceGRPO - MixGRPO - NeighborGRPO).