When Does Diffusion Purification Amplify Perturbations?

Haibo Zhang
haiboz@ai.kyutech.ac.jp
Kyushu Institute of Technology, Japan

Saiyue Lyu
saiyue.lyu@ubc.ca
University of British Columbia, Canada

Yuntao Wang
y-wang@uec.ac.jp
The University of Electro-Communications, Japan

Takeshi Saitoh
saitoh@ai.kyutech.ac.jp
Kyushu Institute of Technology, Japan

Abstract

Diffusion-based purification improves adversarial robustness by denoising perturbed inputs before classification, but the temporal behavior of the reverse process remains less understood. In this work, we study the temporal sensitivity of diffusion purification. Shared-noise clean and adversarial trajectories show a non-monotonic divergence pattern, increasing during early reverse stages and partially decreasing near the final output. To move beyond this trajectory-level observation, we perform controlled perturbation injection at different reverse steps. We observe that perturbations introduced earlier in the reverse process tend to have a larger effect on the final purified output than those introduced later, providing evidence for a temporally non-uniform sensitivity structure in this purification setting. We further show that a simple early-stage damping intervention systematically changes the trade-off between clean preservation and adversarial recovery. Our results suggest that, in the studied score-based purification setting, diffusion purification does not behave as a uniformly stabilizing process across time; instead, its robustness behavior appears strongly influenced by early reverse dynamics.

1. Introduction

Deep neural networks are known to be vulnerable to adversarial examples, where small and often imperceptible perturbations can lead to incorrect predictions (Goodfellow et al., 2015; Madry et al., 2018). Evidence of such vulnerabilities in high-stakes applications, from medical diagnosis (Finlayson et al., 2019) to autonomous driving (Eykholt et al., 2018), has made adversarial defense an important research direction. Adversarial purification offers a test-time defense strategy: instead of modifying the classifier, it attempts to transform a perturbed input back toward the natural image manifold before classification (Song et al., 2018; Samangouei et al., 2018; Yoon et al., 2021). With the rapid progress of diffusion and score-based generative models (Ho et al., 2020; Song et al., 2021b), diffusion-based purification has emerged as a promising direction for adversarial robustness (Yoon et al., 2021; Nie et al., 2022; Wang et al., 2022; Carlini et al., 2023). The core idea is to add noise to an input and then use the reverse diffusion process to recover a clean image, thereby reducing adversarial effects before classification.

Despite these encouraging results, relatively less attention has been paid to the temporal behavior of the reverse diffusion process in adversarial purification. Existing work mainly evaluates purification through final robust accuracy, adaptive attacks, or improved guidance mechanisms (Nie et al., 2022; Lee & Kim, 2023; Lin et al., 2025; Bai et al., 2024; Li et al., 2025). These studies have significantly advanced diffusion-based defenses, but they mostly focus on input-output behavior rather than the internal temporal dynamics of the reverse trajectory. In particular, it remains unclear whether perturbations are suppressed uniformly across reverse steps, or whether some stages are more sensitive than others. This question is important because adaptive attacks against purification methods often interact with the reverse process itself (Athalye et al., 2018; Croce & Hein, 2020; Kang et al., 2023).

In this work, we study diffusion purification through the lens of *temporal sensitivity*. Rather than only evaluating the final purified output, we analyze how clean–adversarial discrepancies evolve along the reverse trajectory. In a

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

score-based purification pipeline with shared noise, we observe that the discrepancy is not monotonically reduced, but can transiently increase during early-to-middle reverse stages before partially decreasing near the output. We further isolate this effect using controlled perturbation injection. Perturbations introduced earlier in the reverse process tend to have a larger influence on the final output than those introduced later, suggesting a reproducible temporal sensitivity pattern in the studied reverse dynamics. To test whether this sensitivity affects purification outcomes, we introduce a simple early-stage damping intervention and find that it shifts the trade-off between clean preservation and adversarial recovery. To sum up, our main contributions are:

① We empirically identify temporal non-uniform sensitivity in score-based diffusion purification.

② We show through perturbation injection that early reverse stages have larger final-output influence than later stages.

③ We show that early-stage damping shifts the clean–adversarial purification trade-off.

2. Related work

2.1. Diffusion-Based Adversarial Purification

Diffusion models generate samples by gradually adding noise through a forward process and then recovering clean samples through a learned reverse denoising process, which can be formulated either as a discrete Markov chain (Ho et al., 2020) or as a continuous-time stochastic differential equation (SDE) (Song et al., 2021b). This reverse process is central to the quality, controllability and stability of the generation and has motivated improvements such as guided diffusion (Dhariwal & Nichol, 2021), accelerated sampling (Song et al., 2021a), and respacing in stages. In adversarial purification, diffusion models are used as test-time denoisers: a perturbed input is first noised and then passed through the reverse process to obtain a purified sample prior to classification (Song et al., 2018; Samangouei et al., 2018; Yoon et al., 2021; Nie et al., 2022; Wang et al., 2022). A representative example is DiffPure (Nie et al., 2022), which shows strong empirical robustness by applying reverse diffusion to remove adversarial perturbations. Follow-up methods further improve purification through guidance mechanisms and model variants, including adversarial guidance (Lin et al., 2025), contrastive guidance (Bai et al., 2024), and diffusion bridge models (Li et al., 2025).

2.2. Robustness Evaluation of Purification Defenses

Several studies have raised concerns about the robustness of purification-based defenses and the difficulty of reliable evaluation. It is well known that defenses relying on generative models may suffer from obfuscated gradients or evaluation artifacts (Athalye et al., 2018). Reliable evaluation protocols such as AutoAttack (Croce & Hein, 2020) have been widely adopted to address these issues.

In the context of diffusion-based purification, recent work has proposed stronger adaptive attacks that explicitly target the purification process. For example, DiffAttack (Kang et al., 2023) demonstrates that diffusion-based defenses can be bypassed when the attacker accounts for the stochastic reverse process. Other studies have also highlighted the importance of careful evaluation and the potential gap between apparent and true robustness (Lee & Kim, 2023).

2.3. Timestep-Aware and Error-Propagation Analyses in Diffusion

Recent work has shown that different timesteps in diffusion models can play different roles. For example, P2 weighting reweights the training objective across noise levels and shows that some denoising stages are more important for learning perceptually meaningful structure (Choi et al., 2022). Fast samplers such as DDIM and DPM-Solver also rely on the fact that the reverse process can be modified or discretized in non-uniform ways while preserving sample quality (Song et al., 2021a; Lu et al., 2022). Progressive distillation further reduces the number of sampling steps by repeatedly distilling a slow sampler into a faster one, showing that the temporal structure of the reverse process can be compressed (Salimans & Ho, 2022).

Error propagation has also been studied in diffusion models, especially in the context of generation quality and model compression. Li & van der Schaar (2024) analyze how errors accumulate in diffusion models due to their sequential structure and propose a regularization method to reduce this effect . In quantized diffusion models, several works observe that quantization errors can accumulate across denoising steps, and propose timestep-aware correction or cross-timestep compensation strategies to improve low-precision sampling (Yao et al., 2024; Chen et al., 2024; Li &

Du, 2025; Liu et al., 2025).

However, these analyses are mainly developed for generation, acceleration, or quantization. They do not directly study adversarial purification, where the input perturbation is intentionally structured to affect classification after purification. In contrast, our work focuses on the temporal sensitivity of the reverse diffusion process in adversarial settings. We study how small perturbations introduced at different reverse steps influence the final purified output, and how this behavior affects the trade-off between clean preservation and adversarial recovery.

3. Preliminary

Throughout the paper, we use T to denote the full discrete time grid of the score-SDE implementation. In our experiments, $T = 1000$. A discrete timestep $t \in \{0, \dots, T\}$ is mapped to the continuous SDE time by $s_t = t/T$. Diffusion purification does not start from the full noise level $t = T$; instead, it starts from a smaller purification horizon $T_{\text{pur}} \leq T$. For example, $T_{\text{pur}} = 200$ corresponds to continuous time $s_{T_{\text{pur}}} = 0.2$.

3.1. Score-Based Denoising

Score-based diffusion models (Song et al., 2021b) describe the noising and denoising processes through stochastic differential equations (SDEs). Let $\boldsymbol{x}_0$ denote a clean data point. The forward process gradually perturbs $\boldsymbol{x}_0$ into a noisy sample $\boldsymbol{x}_s$ according to $d\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, s)\, ds + h(s)\, d\boldsymbol{w}_s$, where $\boldsymbol{f}(x, s)$ is the drift term, $h(s)$ is the diffusion coefficient, $d\boldsymbol{w}_s$ is a standard Wiener process, and $s \in [0, 1]$ is the positive time increments. As s increases, the distribution of $\boldsymbol{x}_s$ becomes progressively closer to a simple noise distribution. The reverse process removes noise by solving the corresponding reverse-time SDE:

$$d\boldsymbol{x} = \left[\boldsymbol{f}(\boldsymbol{x}, s) - h(s)^2 \nabla_{\boldsymbol{x}} \log p_s(\boldsymbol{x})\right] ds + h(s)\, d\bar{\boldsymbol{w}}_s, \tag{1}$$

where p_s is the marginal distribution of noisy samples at time s, and $\bar{\boldsymbol{w}}_s$ denotes the reverse-time Wiener process. Since the true score $\nabla_{\boldsymbol{x}} \log p_s(\boldsymbol{x})$ is unknown, score-based models train a neural network $\boldsymbol{\psi}_\theta(\boldsymbol{x}, s) \approx \nabla_{\boldsymbol{x}} \log p_s(\boldsymbol{x})$ to approximate it. This score points toward directions of higher data density and therefore provides the denoising direction during the reverse process. To implement the reverse process on the discrete time grid, we move from x_t to x_{t-1}, while evaluating the SDE coefficients at the continuous time $s_t = t/T$. Let $\Delta s = 1/T$. Then

$$x_{t-1} = x_t - \left[f(x_t, s_t) - h(s_t)^2 \psi_\theta(x_t, s_t)\right] \Delta s + h(s_t)\sqrt{\Delta s}\, z, \tag{2}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, I)$. Thus, starting from a noisy sample $\boldsymbol{x}_t$, the next denoised state $\boldsymbol{x}_{t-1}$ is obtained by combining three terms: the reverse drift $-\boldsymbol{f}(\boldsymbol{x}_t, s_t)\Delta s$, the learned score-based denoising direction $h(s_t)^2 \boldsymbol{\psi}_\theta(x_t, s_t)\Delta s$, and a Gaussian noise term $h(s)\sqrt{\Delta s} z$.

We use $s_t = t/T$ to map discrete timesteps t to continuous values s in $[0, 1]$. Given $\boldsymbol{x}_0$, the forward process constructs a noisy state:

$$\boldsymbol{x}_t = \bar{\alpha}_t \boldsymbol{x}_0 + \bar{\sigma}_t \boldsymbol{z} := \alpha(s_t)\boldsymbol{x}_0 + \sigma(s_t)\boldsymbol{z}, \quad \boldsymbol{z} \in \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where $\alpha(s)$ and $\sigma(s)$ are fixed by the chosen SDE schedule. In our VP-SDE case, $\alpha(s) = \exp\left(-\frac{1}{2}\int_0^s \beta(\boldsymbol{u})d\boldsymbol{u}\right)$ and $\sigma^2(s) = 1 - \exp\left(-\int_0^s \beta(\boldsymbol{u})d\boldsymbol{u}\right)$.

The backward process then iteratively denoises the sample:

$$\boldsymbol{x}_{t-1} = \mathcal{D}_\theta(\boldsymbol{x}_t, t) \tag{4}$$

where $\mathcal{D}_\theta(\boldsymbol{x}_t, t) = \boldsymbol{x}_t - \left[\boldsymbol{f}(\boldsymbol{x}_t, s_t) - h(s_t)^2 \boldsymbol{\psi}_\theta(\boldsymbol{x}_t, s_t)\right]\Delta s + \bar{\gamma}_t \boldsymbol{z}, \boldsymbol{z} \in \mathcal{N}(0, \mathbf{I})$ and $\bar{\gamma}_t = h(s_t)\sqrt{\Delta s_t}$. In our VP-SDE case, $\boldsymbol{f}(\boldsymbol{x}, s) = -\frac{1}{2}\beta(s)\boldsymbol{x}$ and $h(s) = \sqrt{\beta(s)}$.

3.2. Diffusion-Based Purification

Let $\boldsymbol{x} \in \mathbb{R}^d$ denote a clean image with label y, and let $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \delta$ denote its adversarial counterpart, where δ is a small perturbation generated under a specified threat model to make a classifier g misclassify on the input $\boldsymbol{x}$, i.e. $g(\boldsymbol{x}) \neq g(\boldsymbol{x}^{\text{adv}})$. Typically δ is constrained by $\|\delta\|_p \leq \epsilon$, where ϵ is the maximum scale of the perturbation.

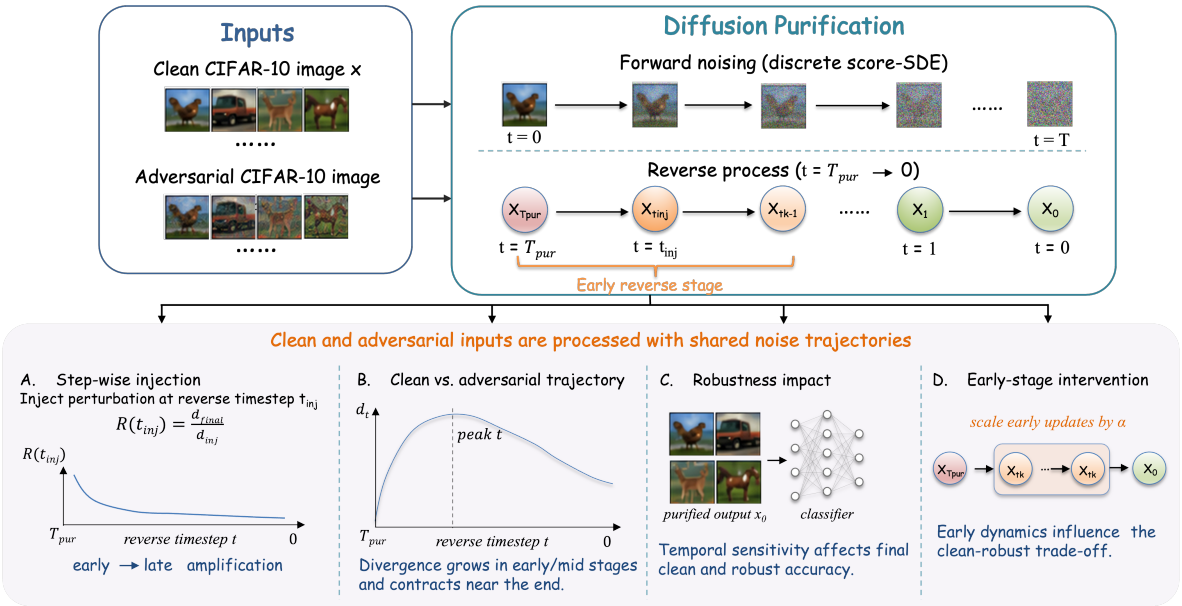


Figure 1. Overview of our temporal sensitivity analysis for diffusion purification. The score-SDE implementation uses a full discrete time grid of size T, while purification starts from a smaller horizon T_{pur} and runs the reverse process from $t = T_{\text{pur}}$ to $t = 0$.

To compare the reverse dynamics induced by clean and adversarial inputs, we initialize both diffusion based purification processes using the same forward noise realization. Specifically, we sample one shared $\boldsymbol{z} \in \mathcal{N}(0, \mathbf{I})$ and form:

$$\boldsymbol{x}_{T_{\text{pur}}} = \bar{\alpha}_{T_{\text{pur}}} x + \bar{\sigma}_{T_{\text{pur}}} z, \quad \boldsymbol{x}^{\text{adv}}_{T_{\text{pur}}} = \bar{\alpha}_{T_{\text{pur}}} x^{\text{adv}} + \bar{\sigma}_{T_{\text{pur}}} z. \tag{5}$$

We then apply the same reverse diffusion sampler to both noisy inputs, using shared stochasticity across reverse steps when applicable. For $t = T_{\text{pur}}, T_{\text{pur}} - 1, \ldots, 1$, we apply the same reverse denoising update:

$$\boldsymbol{x}_{t-1} = \mathcal{D}_\theta(\boldsymbol{x}_t, t), \quad \boldsymbol{x}^{\text{adv}}_{t-1} = \mathcal{D}_\theta(\boldsymbol{x}^{\text{adv}}_t, t). \tag{6}$$

This produces paired trajectories $\{x_t\}_{t=T_{\text{pur}}}^{0}$ and $\{x_t^{\text{adv}}\}_{t=T_{\text{pur}}}^{0}$. The final purified outputs $\boldsymbol{x}_0$ and $\boldsymbol{x}_0^{\text{adv}}$ are then passed to a classifier g.

In our analysis, clean and adversarial inputs are purified using shared randomness, i.e. the same forward noise and reverse-process stochasticity are used for x and x^{adv}, to remove unrelated sampling variation and makes the trajectory difference more directly reflect the effect of the adversarial perturbation.

4. Analysis Methods

Figure 1 summarizes our analysis pipeline. Starting from paired clean and adversarial CIFAR-10 inputs, we run diffusion purification with shared noise trajectories and study how perturbations evolve across reverse timesteps. Our analysis focuses on three complementary questions: how clean–adversarial discrepancies evolve, how sensitive different reverse stages are to injected perturbations, and how early-stage dynamics affect final purification outcomes.

Paired Denoising Trajectory Discrepancy. For each discrete timestep t along the purification trajectory, we measure $d_t = \left\| x_t^{\text{clean}} - x_t^{\text{adv}} \right\|_2$. The quantity d_t describes how the clean–adversarial discrepancy changes along the reverse process. If the reverse dynamics were uniformly contractive with respect to input perturbations, then $d_{t-1} \leq d_t, \forall t$. So d_t decreases monotonically as the reverse process moves from larger t to 0. However, diffusion purification does not guarantee such stepwise contraction. We therefore treat d_t as an empirical diagnostic and ask whether the reverse process actually suppresses perturbations uniformly across time.

Figure 2 provides a qualitative example. Because the clean and adversarial inputs are noised with the same Gaussian sample, their initial difference at large t is scaled by the signal coefficient and can be small when the noise level is high.

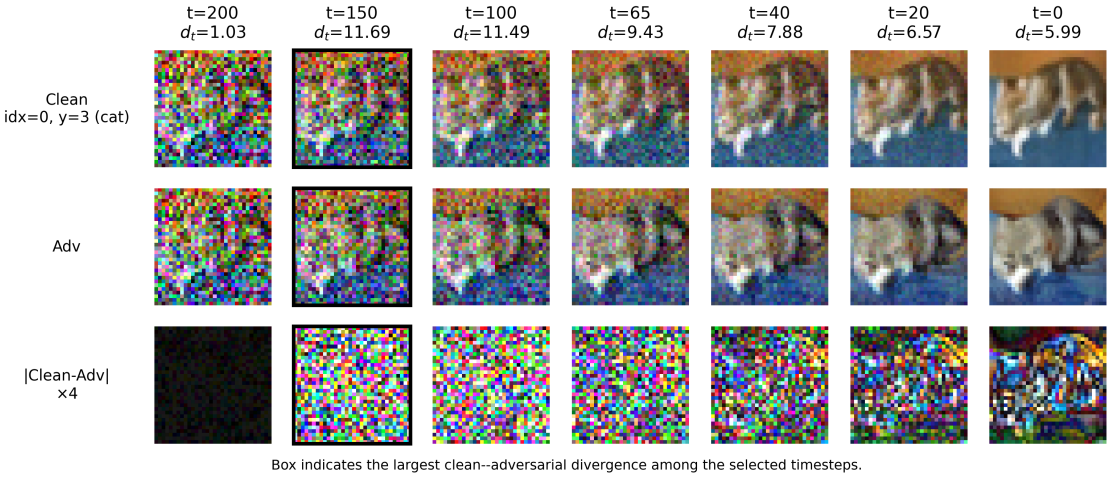


Figure 2. Qualitative illustration of clean and adversarial reverse trajectories under shared noise. Each column shows an intermediate reverse state, starting from the noised state and ending at the purified output. The first two rows show trajectories initialized from the clean and adversarial inputs, respectively. The last row shows the amplified absolute difference between them. The black box marks the timestep with the largest clean–adversarial discrepancy in this example.

As denoising recovers image structure, the clean and adversarial trajectories may separate before partially reconverging near the final output. This suggests that *perturbations are not necessarily removed uniformly*; instead, the reverse dynamics may amplify differences at some stages and suppress them at others.

Step-wise Perturbation Amplification Ratio. Trajectory discrepancy measures the behavior of actual clean-adversarial pairs, but it does not isolate the sensitivity of a particular reverse step. To directly test step-wise sensitivity, we introduce a controlled perturbation at a chosen timestep.

Let t_{inj} denote the discrete score-SDE timestep at which we inject a perturbation:

$$\tilde{x}_{t_{\text{inj}}} = x_{t_{\text{inj}}} + \delta, \qquad \|\delta\|_2 = \epsilon_{\text{inj}}. \tag{7}$$

where δ is a small perturbation with fixed norm. We then continue the same reverse process from both $x_{t_{\text{inj}}}$ and $\widetilde{x}_{t_{\text{inj}}}$, using shared future randomness, to obtain final outputs x_0 and $\widetilde{x}_0$. This yields two denoising trajectories $\{\boldsymbol{x}_T, \cdots, \boldsymbol{x}_{t_{\text{inj}}+1}, \boldsymbol{x}_{t_{\text{inj}}}, \cdots, \boldsymbol{x}_0\}$ and $\{\boldsymbol{x}_T, \cdots, \boldsymbol{x}_{t_{\text{inj}}+1}, \tilde{\boldsymbol{x}}_{t_{\text{inj}}}, \cdots, \tilde{\boldsymbol{x}}_0\}$.

Define the amplification ratio for an injection at timestep t_{inj} as

$$R(t_{\text{inj}}) = \frac{\|\widetilde{x}_0 - x_0\|_2}{\left\|\widetilde{x}_{t_{\text{inj}}} - x_{t_{\text{inj}}}\right\|_2} = \frac{\|\widetilde{x}_0 - x_0\|_2}{\epsilon_{\text{inj}}}. \tag{8}$$

$R(t_{\text{inj}}) > 1$ indicates that the remaining reverse process amplifies the injected perturbation, while $R(t_{\text{inj}}) < 1$ indicates that it attenuates the perturbation. Comparing $R(t_{\text{inj}})$ across different t_{inj} therefore reveals whether some reverse timesteps have a larger influence on the final purified output than others.

This controlled injection complements the clean-adversarial pair trajectory analysis. The trajectory discrepancy measures how real adversarial perturbations evolve, while the amplification ratio $R(t_{\text{inj}})$ measures the local sensitivity of the remaining reverse process to perturbations introduced at timestep t_{inj}.

Early-Step Damping Intervention. The preceding analyses measure how perturbations propagate along the reverse trajectory. We next analyze whether this temporal sensitivity has direct consequences for the final purified output. In particular, since early reverse steps may strongly amplify small perturbations, we test whether modifying only these early updates changes the trade-off between clean preservation and adversarial recovery.

For the denoising trajectory $\{\boldsymbol{x}_{T_{\text{pur}}}, \cdots, \boldsymbol{x}_0\}$, we construct a damped trajectory $\{\hat{\boldsymbol{x}}_{T_{\text{pur}}}, \cdots, \hat{\boldsymbol{x}}_t, \boldsymbol{x}_{t-1}, \cdots, \boldsymbol{x}_0\}$ by:

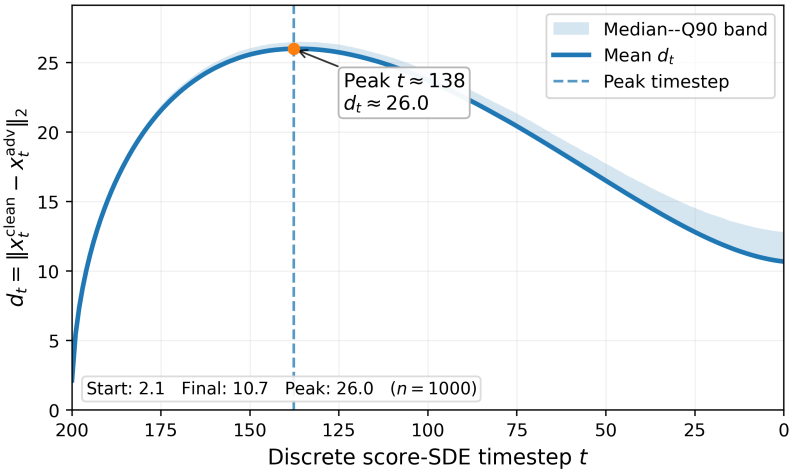


Figure 3. Clean-vs-adversarial trajectory discrepancy under shared reverse stochasticity. The curve shows the mean discrepancy over 1000 CIFAR-10 image pairs, with a shaded band showing the range between the median and the 90th percentile across samples. The discrepancy first increases sharply. The peak occurs around diffusion timestep $t \approx 138$ when $T_{\text{pur}} = 200$. This indicates that adversarial differences are not monotonically removed during diffusion purification.

$$\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_{t+1} + \alpha \left(\mathcal{D}_\theta(\hat{\boldsymbol{x}}_{t+1}, t+1) - \hat{\boldsymbol{x}}_{t+1} \right), \tag{9}$$

where $0 < \alpha \leq 1$ and $\alpha = 1$ recovers the original reverse process, while smaller values make the early reverse updates more conservative by reducing the step size along the original update direction. After the first K reverse steps, we resume the unmodified reverse process.

This intervention is designed as a minimal diagnostic probe of early-stage dynamics. It does not change the diffusion model, add extra denoising steps, introduce external information, or modify the downstream classifier. Instead, it only rescales the magnitude of the early reverse updates. Therefore, if varying α or K changes the final clean accuracy or adversarial accuracy, this indicates that the early portion of the reverse trajectory has outcome-level influence on purification.

We evaluate this intervention by comparing the final classification performance of purified clean and adversarial inputs under different damping strengths. The resulting clean–robust accuracy trade-off reveals whether early reverse dynamics primarily help preserve clean semantic content, remove adversarial perturbations, or amplify harmful trajectory differences.

5. Experimental Results

Our experiments are designed to answer three questions. First, does the clean–adversarial discrepancy decrease monotonically during purification? Second, are different reverse stages equally sensitive to small perturbations? Third, do temporally sensitive stages affect the final clean/adversarial trade-off? We study these questions through four experiments. Section 5.1 compares clean and adversarial reverse trajectories under shared randomness. Section 5.2 uses controlled perturbation injection to measure step-wise sensitivity. Section 5.3 evaluates how the purification horizon affects final classification outcomes. Section 5.4 applies a minimal early-stage damping intervention to test whether early reverse dynamics have outcome-level effects.

We empirically study the temporal behavior of diffusion-based purification on CIFAR-10 (Krizhevsky et al., 2009). We use the VP-SDE score-based diffusion model from Song et al. (2021b). Unless otherwise stated, adversarial examples are generated by a 20-step PGD attack under the ℓ_∞ threat model with perturbation budget $\epsilon = 8/255$, step size $2/255$, and random initialization. We denote the purification horizon by T_{pur}. Larger T_{pur} corresponds to stronger forward noising followed by a longer reverse denoising process. For trajectory-based experiments, paired runs use the same forward Gaussian noise and the same reverse-process stochasticity. This coupling reduces unrelated sampling variation

Table 1. Controlled perturbation injection across purification horizons and injection magnitudes. For each setting, we report the mean amplification ratio in early reverse steps, late reverse steps, and their ratio. Larger $R_{\text{early}}/R_{\text{late}}$ indicates stronger temporal non-uniformity. All results are computed over 500 CIFAR-10 images.

T_{pur}	ϵ_{inj}	R_{early}	R_{late}	ρ
100	1×10^{-4}	61,875.0	27,438.7	2.255
100	3×10^{-4}	20,684.0	9,168.0	2.256
100	1×10^{-3}	6,196.8	2,756.3	2.248
200	1×10^{-4}	92,413.8	40,378.6	2.289
200	3×10^{-4}	30,738.3	13,445.7	2.286
200	1×10^{-3}	9,193.7	4,015.9	2.289
300	1×10^{-4}	119,105.0	50,518.9	2.358
300	3×10^{-4}	39,642.4	16,797.8	2.360
300	1×10^{-3}	11,865.5	5,010.7	2.368
400	1×10^{-4}	146,817.0	59,682.4	2.460
400	3×10^{-4}	49,254.0	19,901.4	2.475
400	1×10^{-3}	14,764.8	5,913.4	2.497

and makes trajectory differences more directly reflect the input perturbation.

5.1. Paired Denoising Trajectory Discrepancy

We first examine whether actual adversarial perturbations are uniformly suppressed along the reverse trajectory. For each clean image and its adversarial counterpart, we run diffusion purification under shared randomness and measure the paired clean–adversarial discrepancy at each recorded reverse state. We use $T_{\text{pur}} = 200$, evaluate 1000 clean/adversarial image pairs, and extract the full reverse trajectory with 200 recorded states.

Figure 3 shows that the clean–adversarial discrepancy is not monotonically reduced. The discrepancy increases sharply during the early reverse process, reaches its maximum around $t \approx 138$, and then gradually decreases. In our run, the mean discrepancy increases from approximately 2.1 at the first recorded state to approximately 26.0 at its peak, and remains approximately 10.7 at the final output. This transient amplification suggests that the reverse process does not behave as a uniformly stabilizing denoiser.

5.2. Controlled Perturbation Injection

The trajectory experiment measures how real clean–adversarial differences evolve. We next isolate the sensitivity of individual reverse stages using the controlled perturbation injection procedure. This experiment asks whether small perturbations injected at different reverse positions have comparable influence on the final purified output. For visualization and comparison across purification horizons, we also report the injection location using the normalized reverse position:

$$q = \frac{T_{\text{pur}} - t_{\text{inj}}}{T_{\text{pur}}}. \tag{10}$$

Small q corresponds to early reverse stages near the noised input, while large q corresponds to late reverse stages near the purified output. We inject perturbations at $q \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We evaluate $T_{\text{pur}} \in \{100, 200, 300, 400\}$ and $\epsilon_{\text{inj}} \in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$. For each pair $(T_{\text{pur}}, \epsilon_{\text{inj}})$, we evaluate 500 images and scan the nine normalized injection positions. Unless otherwise stated, each trajectory is extracted with the same number of recorded reverse states as the purification horizon. To summarize each sensitivity curve, we report R_{early} as the average amplification ratio over early positions $q \in \{0.1, 0.2, 0.3\}$, and R_{late} as the average over late positions $q \in \{0.7, 0.8, 0.9\}$. Their ratio $\rho = R_{\text{early}}/R_{\text{late}}$ measures the strength of temporal non-uniformity.

Table 1 shows a consistent pattern across all tested horizons and injection magnitudes. Perturbations injected early in the reverse process produce much larger final deviations than perturbations injected late. The ratio ρ is always greater than 2, showing that early-stage perturbations have more than twice the final-output influence of late-stage perturbations.

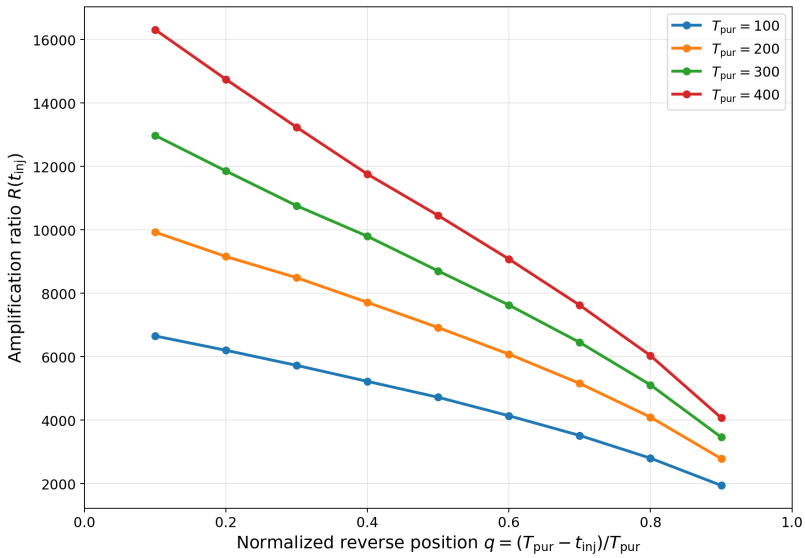


Figure 4. Controlled perturbation injection across different purification horizons. We inject perturbations with $\epsilon_{\text{inj}} = 10^{-3}$ at normalized reverse positions q, and report the amplification ratio $R(t_{\text{inj}})$. The y-axis shows $R(t_{\text{inj}})$ in units of 10^3. Across all tested horizons, perturbations injected earlier in the reverse process lead to larger final deviations than perturbations injected later.

Table 2. Purification impact across different purification horizons. We report clean-image preservation, adversarial-image recovery, and the resulting trade-off. The adversarial examples are generated by 20-step PGD under the ℓ_∞ threat model with $\epsilon = 8/255$.

T_{pur}	Input Clean Acc	Input Adv Acc	Pur Clean Acc	Pur Adv Acc	Clean Dam Rate	Both Pur Corr
100	0.946	0.000	0.896	0.886	0.076	0.840
200	0.946	0.000	0.815	0.796	0.165	0.710
300	0.946	0.000	0.638	0.642	0.342	0.490
400	0.946	0.000	0.476	0.460	0.514	0.310

Figure 4 shows the full amplification curves for $\epsilon_{\text{inj}} = 10^{-3}$. For every tested horizon, the amplification ratio is largest near the beginning of the reverse process and decreases toward later reverse positions. In our tested score-based purification setting, this indicates a temporally non-uniform sensitivity pattern: perturbations introduced at early reverse stages have a disproportionately large effect on the final purified output.

5.3. Purification Horizon and Output Trade-off

The previous experiments analyze the internal behavior of the reverse process. We now ask whether these temporal effects are reflected in final classification outcomes. We evaluate how the purification horizon T_{pur} affects both clean-image preservation and adversarial-image recovery. This experiment is not intended to claim state-of-the-art robustness; it serves as an outcome-level check of how the amount of purification changes the final clean/adversarial trade-off.

We report purified clean accuracy, purified adversarial accuracy, clean damage rate, and Both purified correct. Clean damage rate measures the fraction of clean images that are correctly classified before purification but incorrectly classified after purification. Both purified correct, computed as $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[f(\mathcal{P}_\theta^{T_{\text{pur}}}(x_i)) = y_i \wedge f(\mathcal{P}_\theta^{T_{\text{pur}}}(x_i^{\text{adv}})) = y_i\right]$, is a paired metric that counts an image pair as successful only when both the purified clean image and the corresponding purified adversarial image are classified correctly.

Table 2 shows a clear clean–adversarial trade-off across purification horizons. At $T_{\text{pur}} = 100$, purification gives the best balance among the tested settings: it substantially recovers adversarial examples while keeping clean-image degradation moderate. In contrast, larger horizons such as $T_{\text{pur}} = 300$ and $T_{\text{pur}} = 400$ apply stronger purification but introduce more distortion, leading to lower purified clean accuracy and lower Both Pur Correct.

Table 3. Minimal early-stage damping intervention. We apply the damped update to the first K reverse steps with scale α. All results are evaluated on 500 CIFAR-10 clean/adversarial pairs using $T_{\mathrm{pur}} = 200$.

α	K	Pur Clean Acc	Pur Adv Acc	Clean Dam Rate	Both Pur Corr
0.5	20	0.776	0.786	0.205	0.686
	40	0.776	0.800	0.203	0.688
	60	**0.820**	0.790	**0.154**	**0.710**
	80	0.802	0.808	0.180	0.708
0.7	20	0.782	0.794	0.197	0.680
	40	0.784	0.784	0.190	0.690
	60	0.796	0.790	0.184	0.698
	80	**0.820**	0.766	0.163	0.692
0.9	20	0.778	0.802	0.203	0.702
	40	0.784	**0.814**	0.197	0.700
	60	0.788	0.786	0.184	0.676
	80	0.788	0.772	0.195	0.686

5.4. Minimal Early-Stage Damping

The trajectory and injection experiments show that early reverse stages are especially sensitive. We next test whether modifying only these stages changes the final purification outcome. This experiment uses the damping intervention as a diagnostic probe rather than a fully optimized defense method. We fix $T_{\mathrm{pur}} = 200$, evaluate 500 clean/adversarial image pairs, and vary $K \in \{20, 40, 60, 80\}, \alpha \in \{0.5, 0.7, 0.9\}$. For each setting, the first K reverse updates are damped with scale α, and the remaining updates use the original reverse process. We report purified clean accuracy, purified adversarial accuracy, clean damage rate, and Both Pur Correct.

Table 3 shows that changing the early reverse dynamics changes the final clean/adversarial trade-off. No single setting dominates all metrics. For example, $K = 40, \alpha = 0.9$ gives the strongest adversarial recovery in this grid, while $K = 60, \alpha = 0.5$ gives the best clean preservation and the highest Both Pur Correct score. Thus, early-stage damping shifts the operating point of purification rather than uniformly improving every metric.

The trajectory and injection experiments show that early reverse stages are especially sensitive, while the damping experiment shows that modifying these stages can change final classification outcomes. These findings suggest that diffusion purification should not always be viewed as a uniformly stabilizing denoiser; in our setting, its robustness behavior is shaped by temporally non-uniform sensitivity along the reverse process.

6. Conclusion

We studied diffusion-based purification through the temporal dynamics of its reverse process. Paired clean–adversarial trajectories show that adversarial discrepancies are not monotonically removed but can instead be amplified during early reverse stages before partially decreasing near the final output. Controlled perturbation injection further shows that early reverse stages have a disproportionately large influence on the purified sample. These temporal effects also appear in final classification outcomes: changing the purification horizon or damping early reverse updates shifts the trade-off between clean preservation and adversarial recovery. In general, our results suggest that, in the score-based purification setting studied, diffusion purification should not be viewed as a uniformly stabilizing denoiser. Instead, it can exhibit temporally non-uniform sensitivity, with early reverse dynamics playing an important role in the final robustness trade-off.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP26K21314 and JP26K02909.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Bai, M., Huang, W., Li, T., Wang, A., Gao, J., Caiafa, C. F., and Zhao, Q. Diffusion models demand contrastive guidance for adversarial purification to advance. In *Forty-first international conference on machine learning*, 2024.

Carlini, N., Tramèr, F., Dvijotham, K., Rice, L., Sun, M., and Kolter, Z. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023.

Chen, Y.-C., Huang, Z.-K., and Chen, J.-R. Stepbaq: Stepping backward as correction for quantized diffusion models. *Advances in Neural Information Processing Systems*, 37:54054–54078, 2024.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11472–11481, 2022.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–11, 2015.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kang, M., Song, D., and Li, B. Diffattack: Evasion attacks against diffusion-based adversarial purification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images, 2009.

Lee, M. and Kim, D. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 134–144, 2023.

Li, X., Sun, W., Chen, H., Li, Q., He, Y., Shi, J., and Hu, X. ADBM: Adversarial diffusion bridge model for reliable adversarial purification. In *The Thirteenth International Conference on Learning Representations*, 2025.

Li, Y. and Du, C. Optimizing quantized diffusion models via distillation with cross-timestep error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18530–18538, 2025.

Li, Y. and van der Schaar, M. On error propagation of diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Lin, G., Tao, Z., Zhang, J., Tanaka, T., and Zhao, Q. Adversarial guided diffusion models for adversarial purification. *Neural Networks*, 191:107705, 2025.

Liu, S., Zeng, C., Yan, C., Peng, X., Wang, X., Chen, F., and Mei, X. Error propagation mechanisms and compensation strategies for quantized diffusion. *arXiv preprint arXiv:2508.12094*, 2025.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Wang, J., Lyu, Z., Lin, D., Dai, B., and Fu, H. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

Yao, Y., Tian, F., Chen, J., Lin, H., Dai, G., Liu, Y., and Wang, J. Timestep-aware correction for quantized diffusion models. In *European Conference on Computer Vision*, pp. 215–232. Springer, 2024.

Yoon, J., Hwang, S. J., and Lee, J. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pp. 12062–12072. PMLR, 2021.

A. Additional controlled injection results.

In the main text, we show controlled injection curves for different purification horizons using $\epsilon_{\text{inj}} = 10^{-3}$. Here we provide the full set of injection-magnitude comparisons. Figure 5 shows that, for each fixed purification horizon, the same early-to-late decreasing pattern appears under all three injection magnitudes. Figure 6 further summarizes this trend using the early-to-late ratio. Across all tested settings, $R_{\text{early}}/R_{\text{late}} > 1$, confirming that perturbations injected near the beginning of the reverse process have consistently larger final-output influence than perturbations injected near the end.

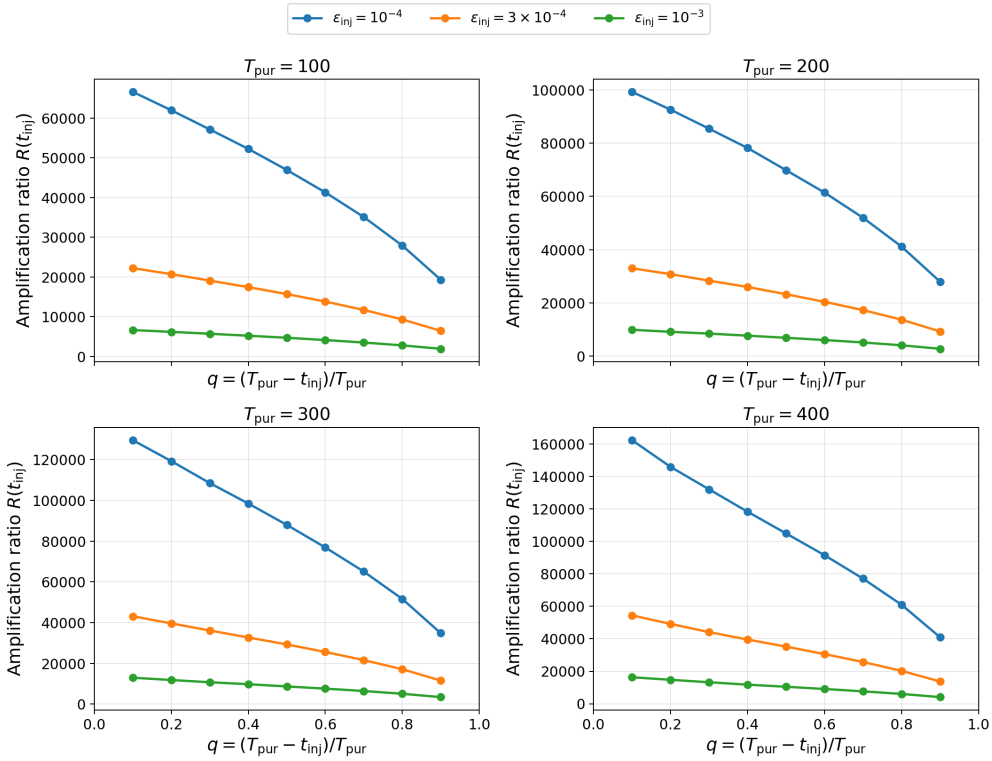


Figure 5. Controlled perturbation injection curves under different injection magnitudes. Each panel fixes a purification horizon T_{pur} and compares $\epsilon_{\text{inj}} \in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$. The x-axis is the normalized reverse position $q = (T_{\text{pur}} - t_{\text{inj}})/T_{\text{pur}}$, where small q corresponds to early reverse stages near the noised input and large q corresponds to late stages near the purified output. The y-axis reports the amplification ratio $R(t_{\text{inj}})$ in units of 10^3. Across all horizons and injection magnitudes, the amplification ratio decreases from early to late reverse positions.

B. Additional Results for Clean-vs-Adversarial Trajectory Divergence

Figure 7 provides three complementary views of the clean–adversarial trajectory discrepancy. First, Figure 7a compares the mean discrepancy at the first recorded state, the final output, and the peak point. The peak discrepancy is much larger than both the initial and final discrepancies, while the final discrepancy remains above the initial value. Second, Figure 7b shows the distribution of per-image peak positions along the recorded reverse trajectory. Most samples reach their maximum discrepancy around recorded reverse positions 60–65, corresponding roughly to discrete score-SDE timesteps $t \approx 135$–140 when $T_{\text{pur}} = 200$. Third, Figure 7c reports the mean, median, and 90th percentile trajectories. These curves follow the same overall pattern: the clean–adversarial discrepancy first increases, reaches a clear peak, and then partially decreases.

Together, these additional results support the conclusion that the clean–adversarial discrepancy is not monotonically removed during reverse diffusion. Instead, it can be transiently amplified during the early-to-middle reverse process before being partially reduced later.

Figure 8 and Figure 9 provide additional qualitative examples of clean and adversarial reverse trajectories under

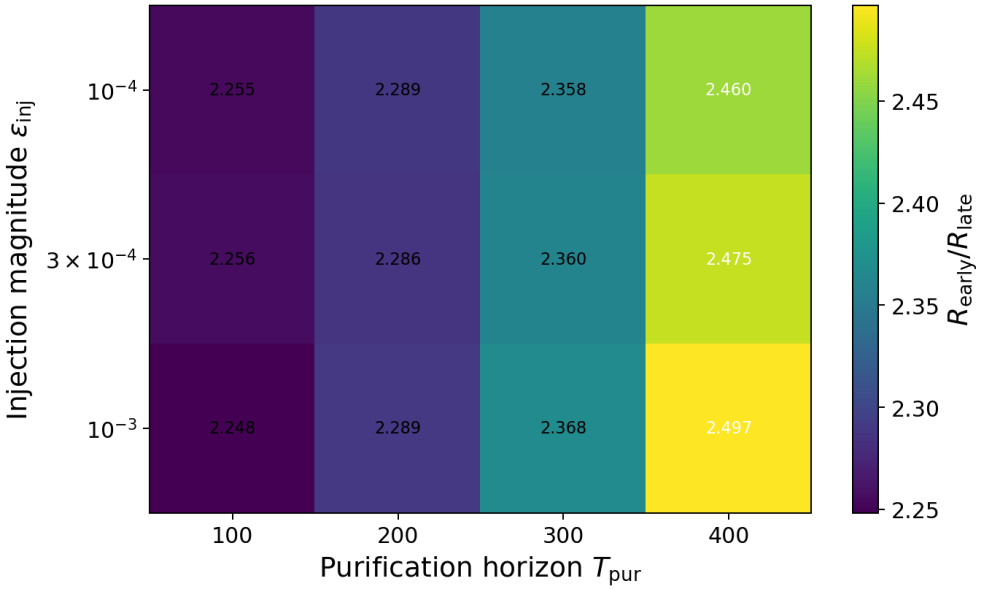


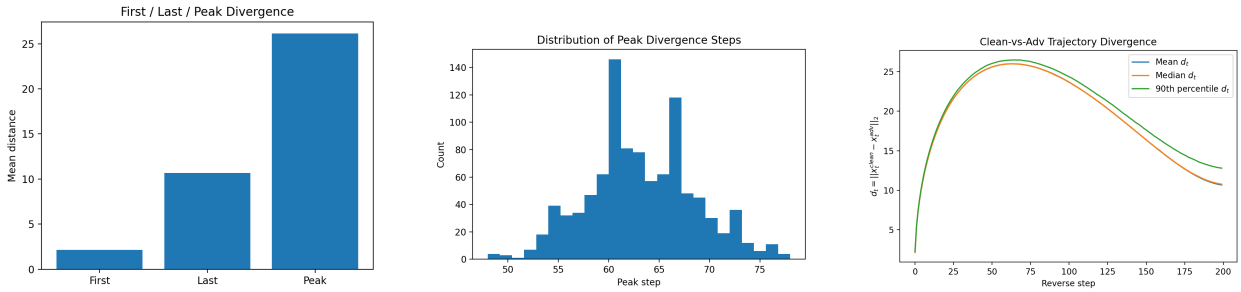
Figure 6. Heatmap of the early-to-late amplification ratio $R_{\text{early}}/R_{\text{late}}$ across purification horizons and injection magnitudes. Here R_{early} averages $R(t_{\text{inj}})$ over early normalized positions $q \in \{0.1, 0.2, 0.3\}$, while R_{late} averages over late positions $q \in \{0.7, 0.8, 0.9\}$. All tested settings have ratios well above 1, showing that early reverse stages are consistently more sensitive than late reverse stages.

shared noise. These examples follow the same visualization format as Figure 2: the first two rows show the clean and adversarial trajectories, while the last row shows the amplified absolute difference between them. Across different samples, we observe a similar non-monotonic pattern: the clean–adversarial discrepancy becomes more visible during the early reverse process and is later partially reduced near the final purified output.

C. Additional results for early-stage damping.

In Section 5.4, we report the main intervention results in table form. Here we visualize the full grid over K and α. Figure 11 shows heatmaps for purified adversarial accuracy, clean damage rate, and Both Purified Correct. Figure 12 further shows the clean–adversarial trade-off across all intervention settings. These plots confirm that early-stage damping changes the final purification behavior, but does not produce a single universally best setting across all metrics.

Figure 12 visualizes the clean–adversarial trade-off across all early-stage damping settings. The scatter plot shows that the intervention changes the operating point of purification rather than uniformly improving all metrics.



(a) First / last / peak divergence. *(b)* Distribution of peak steps. *(c)* Mean, median, and 90th percentile trajectories.

Figure 7. Additional statistics for clean–adversarial trajectory divergence. All results are computed over 1000 CIFAR-10 clean/adversarial image pairs with $T_{\text{pur}} = 200$.

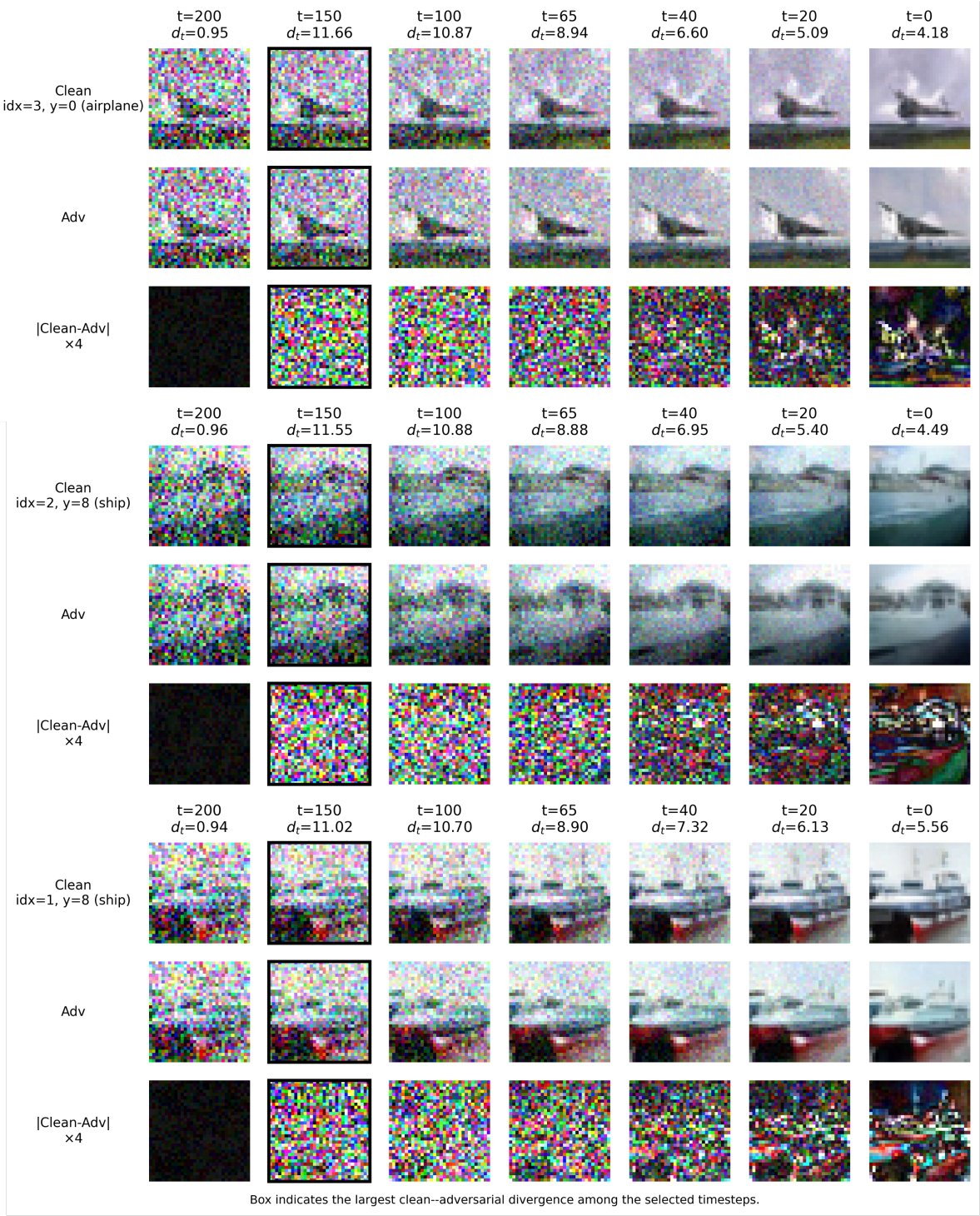


Figure 8. Additional qualitative trajectory examples, part I. Each block shows clean trajectory, adversarial trajectory, and amplified absolute difference.

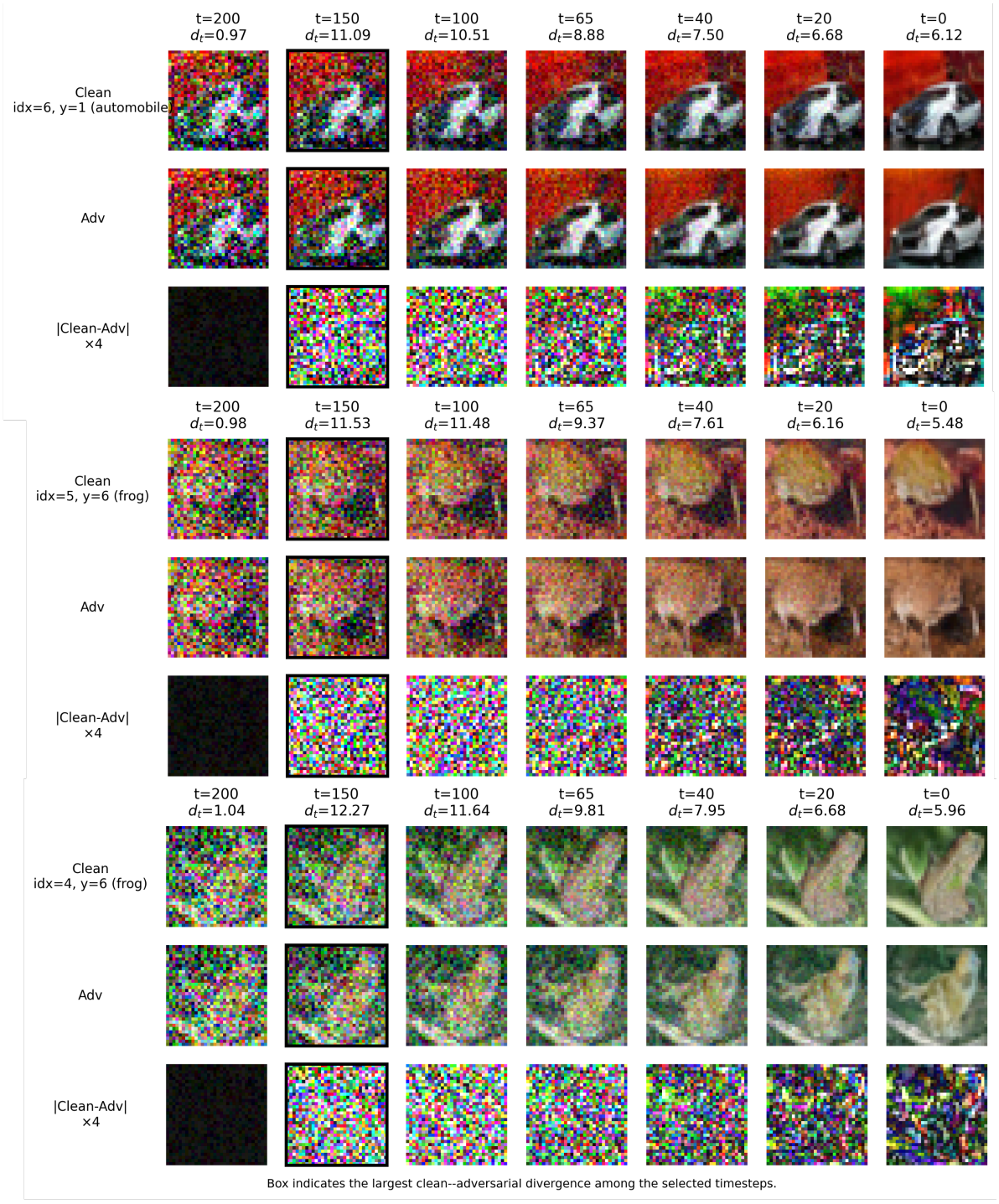


Figure 9. Additional qualitative trajectory examples, part II. The examples show that intermediate clean–adversarial divergence is consistently more pronounced than the final discrepancy.

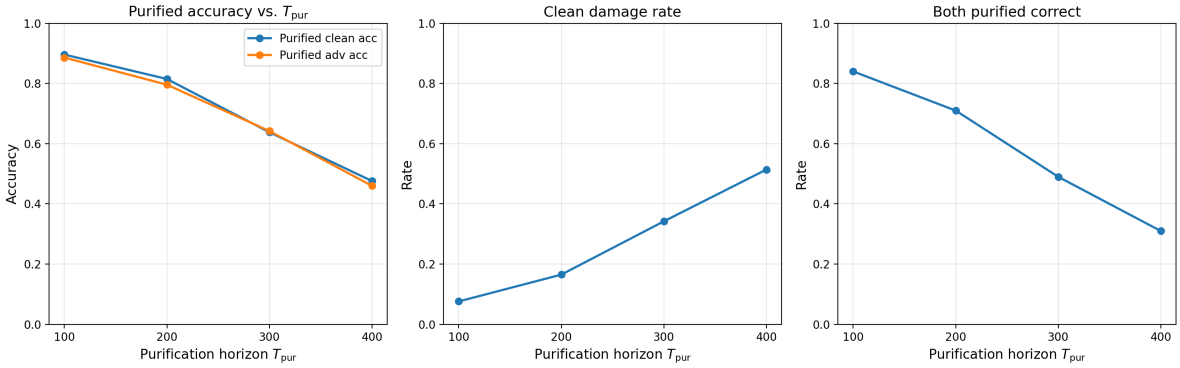
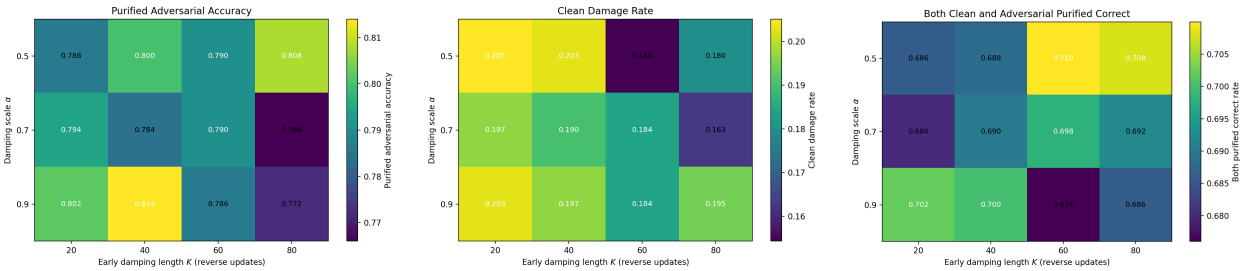


Figure 10. Purification impact across different purification horizons. Here T_{pur} denotes the starting discrete score-SDE timestep for purification. Larger T_{pur} applies stronger forward noising before reverse denoising. As T_{pur} increases, clean and adversarial purified accuracies both decrease, while the clean damage rate increases, showing the clean–adversarial preservation trade-off induced by the purification horizon.



(a) Purified adversarial accuracy. *(b)* Clean damage rate. *(c)* Both purified correct.

Figure 11. Full grid of early-stage damping interventions. Each heatmap varies the early damping length K and scale α. The intervention changes adversarial recovery, clean damage, and paired correctness in different ways, showing that early reverse dynamics affect the final purification trade-off.

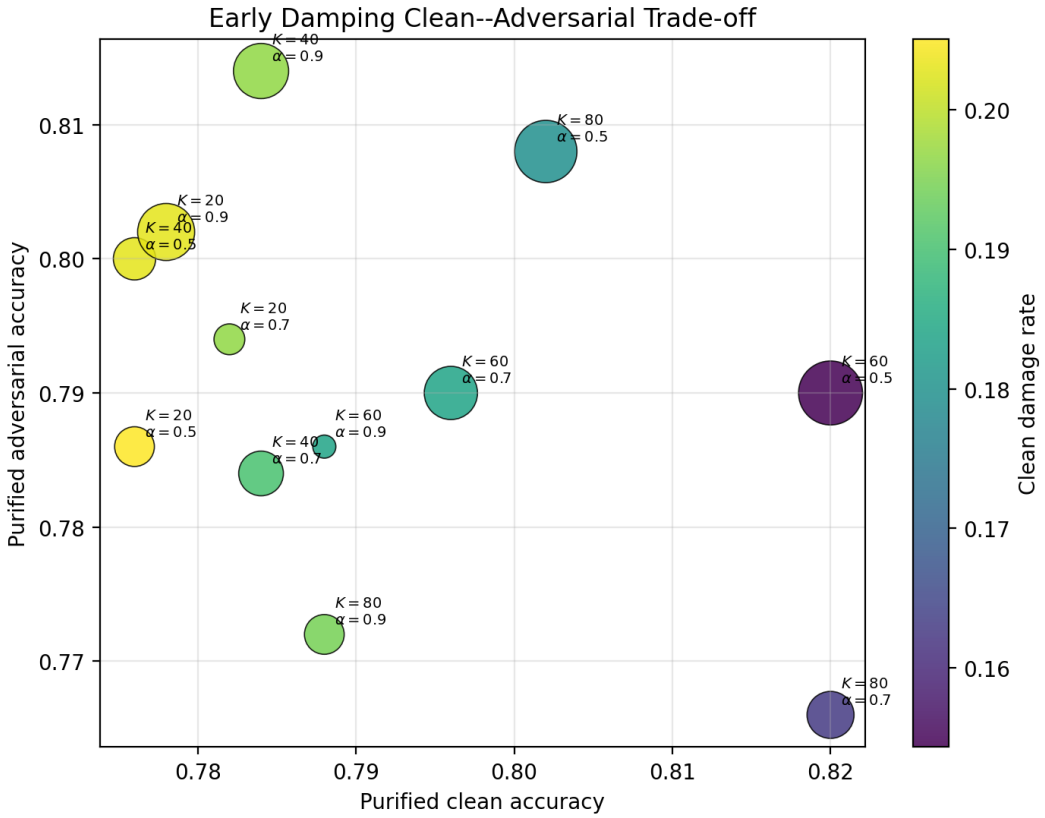


Figure 12. Clean–adversarial trade-off under early-stage damping. Each point is one (K, α) setting. The x-axis shows purified clean accuracy, and the y-axis shows purified adversarial accuracy. Marker size corresponds to Both Purified Correct, and color corresponds to clean damage rate. The grid shows that early-stage damping changes the operating point of diffusion purification rather than uniformly improving all metrics.