Stability Limits of Diffusion Robustness Under Reverse-Process Errors

author names withheld

Editor: Under Review for COLT 2026

Abstract

Diffusion models are increasingly used for adversarial robustness, both as generative diffusion classifiers with certified guarantees and as purification defenses that denoise adversarial inputs via a reverse diffusion process. A central obstacle is that robustness depends on the stability of the implemented reverse dynamics, which combine imperfect score estimation, potentially non-robust guidance, and discretization/solver error. We ask: *how do reverse-process errors propagate under ℓ_p adversarial perturbations, and what limits and algorithmic remedies follow?* We formalize an error-decomposed reverse-process model and prove lower bounds showing that without explicit stability structure, a T-step reverse composition can amplify perturbations and bounded drift errors at a rate governed by the product of stepwise expansion factors. Motivated by this, we propose *Error-Calibrated Robust Diffusion* (ECRD), which enforces per-step contractivity via calibrated damping and suppresses unreliable components via error-aware weighting and discrepancy-regularized guidance. Under standard smoothness and bounded-error conditions, ECRD yields an explicit stability guarantee of the form $\|\operatorname{Pur}(x) - \operatorname{Pur}(x+\delta)\| \lesssim \rho^T \|\delta\| + \sum_{t\leq T} \rho^{t-1} \bar{\Delta}_t$, leading to robust classification guarantees for downstream predictors.

Keywords: Diffusion models, adversarial robustness, reverse-process stability, error accumulation, contraction and Lipschitz analysis, Wasserstein distance and coupling, guided purification defenses

1. Introduction

Diffusion models have become a dominant paradigm for high-dimensional generative modeling, with success across images, audio, and language-conditioned synthesis Ho et al. (2020); Song et al. (2020); Dhariwal and Nichol (2021). Beyond generation, their *iterative denoising* structure has recently been repurposed for adversarial robustness in two prominent ways: (i) *diffusion purification* defenses that forward-noise an input and then reverse-denoise to remove adversarial perturbations, and (ii) *diffusion (generative) classifiers* that predict labels by likelihood surrogates induced by a diffusion model. Empirically, diffusion-based classifiers can exhibit strong robustness under adaptive attacks Chen et al. (2023), and recent work even establishes *certified* robustness for diffusion classifiers via Lipschitz analyses Chen et al. (2024a). On the certification side, pipelines such as DiffSmooth Zhang et al. (2023) and analysis-driven designs such as DensePure Xiao et al. (2023) leverage reverse sampling (with multiple runs and aggregation) to improve provable robustness.

Across all these methods, the defense ultimately depends on a *reverse sampler* that maps a noisy latent x_T to an output x_0 via a long composition of discrete reverse steps. In practice, each reverse step is driven by a learned score (or noise predictor) and often an auxiliary *guidance* signal derived from a classifier, condition model, or energy function. Both ingredients can be fragile: recent evidence shows that guidance can be *non-robust* under adversarial perturbations, steering trajectories toward incorrect basins and degrading both robustness and standard accuracy Lin et al. (2024). More broadly, reverse sampling is a *multi-step computation* with unavoidable imperfections, score approximation error, guidance error under attack, and discretization/solver error, whose effects can *compound* across steps. This raises a basic theoretical question that is largely unresolved:

> **When do reverse-process errors necessarily destroy adversarial robustness, and what stability structure is sufficient to prevent error amplification?**

We study diffusion robustness through a unifying, theory-first lens. The stability of the (randomized) purification map induced by the reverse sampler. Given a defense $\mathrm{Pur}(\cdot;\omega)$ and an ℓ_p adversary $\|\delta\|_p \le \varepsilon$, we focus on the pathwise sensitivity $\Delta^{\mathrm{path}}_{\mathrm{Pur}}(x,\varepsilon;\omega) \triangleq \sup_{\|\delta\|_p\le\varepsilon} \big\| \mathrm{Pur}(x+\delta;\omega) - \mathrm{Pur}(x;\omega)\big\|_2$, and its distributional analogue (e.g., in W_2) when needed. Small Δ_{Pur} is the minimal latent requirement behind many robust pipelines. It implies that a downstream classifier applied to purified samples can inherit robustness under a standard margin/Lipschitz condition, and it underlies why multi-sample voting and smoothing-based certification can succeed. Conversely, if Δ_{Pur} is large, then purification can fail even with a strong diffusion backbone.

Contributions and main results. Our results provide a lower-upper bound perspective on robustness of diffusion-based defenses under reverse-process errors. We identify fundamental obstructions when stability is absent, and give verifiable conditions under which robustness is guaranteed.

- **Model (error decomposition).** We introduce an error-decomposed abstraction for diffusion reverse sampling that separates (i) score approximation error, (ii) guidance mismatch/instability, and (iii) discretization/solver error. This yields a modular framework in which robustness degradation can be attributed to specific reverse-process components.
- **Lower bounds (fundamental obstructions).** Let $\mathrm{Pur}(\cdot;\omega)$ denote the randomized purification map induced by a T-step reverse chain under synchronous coupling, and let Δ_k upper bound the step-k drift error magnitude on the relevant region. We prove a tight error-amplification lower bound. There exist instances and admissible error realizations such that

  $$\|\mathrm{Pur}(x;\omega) - \mathrm{Pur}(\tilde{x};\omega)\|_2 \ge \Big(\prod_{t=1}^{T} L_t\Big)\|x_T - \tilde{x}_T\|_2 - \sum_{k=1}^{T}\Big(\prod_{t=1}^{k-1} L_t\Big)\Delta_k, \tag{1}$$

  where L_t are stepwise expansion factors of the ideal reverse drift. Consequently, if the cumulative expansion $\prod_{t=1}^{T} L_t$ is uncontrolled, then arbitrarily small perturbations and per-step errors can induce macroscopic output deviations, ruling out uniform purification stability in general. We further establish a guidance-specific obstruction. A sharp/steep guidance boundary can trigger basin switching after forward noising, yielding a distributional instability lower bound in W_2 on the order of $\Omega(\|m_+ - m_-\|_2)$ when the forward-noising scale is comparable to the threat radius.
- **Upper bounds (achievability via ECRD).** We propose *ECRD*, an error-calibrated robustification that enforces local contraction along the reverse trajectory and down-weights unreliable drift components using stepwise error budgets, complemented by a discrepancy-regularized robust guidance objective. For a calibrated sampler enforcing per-step contraction $\rho \in (0,1)$ and inducing post-calibration error budgets $\{\bar{\Delta}_k\}$, we prove

  $$\|\mathrm{Pur}(x;\omega) - \mathrm{Pur}(\tilde{x};\omega)\|_2 \le \rho^T\|x_T - \tilde{x}_T\|_2 + \sum_{k=1}^{T}\rho^{k-1}\bar{\Delta}_k \le \rho^T\|x_T - \tilde{x}_T\|_2 + \frac{1-\rho^T}{1-\rho}\bar{\Delta}_{\max}, \tag{2}$$

  with a high-probability extension when contraction certificates fail with probability at most δ_t per step. Under standard margin/Lipschitz conditions on the downstream decision rule, these stability bounds yield robust classification and certification corollaries.

2. Related Work

We group prior work by (i) certification and robustness baselines, (ii) diffusion *classifiers* and their certificates, (iii) diffusion *purification* defenses and robust guidance, and (iv) theoretical analyses of diffusion models and samplers.

Certified robustness and randomized smoothing. A canonical route to provable ℓ_2 robustness is *randomized smoothing*, which turns a base classifier into a smoothed predictor and yields certified radii under Gaussian noise Cohen et al. (2019). Diffusion-based certification methods can be viewed as enriching the smoothing distribution and/or using diffusion dynamics to denoise or stabilize predictions before applying certification machinery Zhang et al. (2023); Xiao et al. (2023). Our work is complementary: we focus on the stability of the *reverse sampler/purification map itself* under reverse-process errors, which is an implicit requirement behind both smoothing-based certificates and multi-sample voting.

Diffusion classifiers and certified robustness. Generative diffusion classifiers compute class likelihoods (or ELBO surrogates) using a single pretrained diffusion model and can exhibit strong empirical robustness Chen et al. (2023). Recent work establishes that diffusion classifiers can be *certifiably* robust by proving Lipschitzness properties and deriving certified bounds, including improved certificates by classifying Gaussian-corrupted data Chen et al. (2024a). These results primarily concern *classification-by-likelihood*. In contrast, we study robustness mediated by a *compositional reverse sampler* used for purification (and also present in many certification pipelines), where errors in score/guidance/solvers compound across steps.

Diffusion purification defenses and adaptive evaluation. Diffusion-based adversarial purification (forward noising followed by reverse denoising) is a widely used defense paradigm Nie et al. (2022). However, purification defenses are notoriously sensitive to evaluation methodology; gradient masking gradients can make a defense appear robust unless attacks are fully adaptive and differentiate through all randomized steps and transformations Athalye et al. (2018). Follow-up work has also questioned the reliability of some adaptive evaluations for DiffPure-style pipelines and proposed modified reverse bridges tailored to purification Liang et al. (2025). Most existing purification papers either emphasize empirical robustness or propose new pipelines, while leaving open a clean theoretical account of *how reverse errors accumulate* under adversarial perturbations.

Robust guidance and robustness under corrupted supervision. Robustness issues also arise in the *training* of diffusion models, e.g., with corrupted labels or datasets, motivating transition-aware weighting and other robust score-matching objectives Na et al. (2024); Dao et al. (2024). This line is orthogonal to our focus: we study *test-time adversarial robustness* that hinges on the stability of the learned reverse dynamics and any guidance signals used during sampling.

Theory of diffusion models and sampling dynamics. A fast-growing body of theory studies convergence and complexity of diffusion samplers under structural assumptions, including low-dimensional adaptation and total-variation convergence for DDPM/DDIM-type samplers Liang et al. (2025), and convergence rates under manifold hypotheses Potaptchik et al. (2024). Other recent works use the diffusion framework as an algorithmic primitive to obtain learning guarantees for distribution families such as Gaussian mixtures Gatmiry et al. (2024); Chen et al. (2024b). These analyses typically assume accurate scores and focus on sampling/learning accuracy. Our perspective is different. We explicitly model *implementation errors* (score/guidance/discretization) and ask when the induced reverse composition can be stable against adversarially-chosen perturbations, and when this is information-theoretically impossible.

3. Problem Formulation and Background

This section formalizes the threat model and the diffusion-based robustness we analyze, and introduces a discrete-time reverse process that makes two robustness bottlenecks explicit: *(i) multi-step (non-)contraction of the reverse drift composition* and *(ii) accumulation of reverse-process errors*.

3.1. Data model, predictors, and an adaptive ℓ_p adversary

Let $(x, y) \sim \mathcal{D}$ with $x \in \mathbb{R}^d$ and $y \in \{1, \ldots, K\}$. A (possibly randomized) defense takes an input $z \in \mathbb{R}^d$ and outputs a prediction $\hat{y} \in \{1, \ldots, K\}$. We consider an ℓ_p-bounded adversary that, given (x, y), outputs $x^{\mathrm{adv}} = x + \delta$, where $\|\delta\|_p \le \varepsilon$. Unless stated otherwise, the adversary is *adaptive* to the defense specification but *oblivious* to the defense randomness. This is the standard threat model for randomized defenses and aligns with the coupling-based stability analysis used throughout.

Diffusion defenses are inherently randomized due to injected Gaussian noise in the reverse sampler and, in purification, typically also in the forward noising step. Accordingly, robustness can be required either *pathwise* or *in distribution*. We formalize both notions in Section 3.5.

3.2. Diffusion-based robustness framework and the common reverse-sampler core

We study robustness framework whose core primitive is a diffusion reverse sampler mapping a noisy latent x_T to an output x_0. Two canonical instantiations are as follows. *(i) Purification followed by a downstream classifier.* A randomized purification map $\mathrm{Pur}(\cdot; \omega)$ outputs a purified sample $\hat{x} = \mathrm{Pur}(x^{\mathrm{adv}}; \omega)$, which is then classified by a (deterministic) base classifier $c : \mathbb{R}^d \to \{1, \ldots, K\}$ as $\hat{y} = c(\hat{x})$. Here, ω collects all sampling randomness. *(ii) Diffusion (generative) classifiers.* A diffusion classifier constructs class scores $g_k(x^{\mathrm{adv}}; \omega)$ derived from diffusion likelihood surrogates (or score-based quantities), and predicts $\hat{y} = \arg\max_k g_k(x^{\mathrm{adv}}; \omega)$. Although the decision rule differs, these methods also rely on reverse dynamics to evaluate class-conditioned score terms.

In both pipelines, robustness depends on the *stability* of the reverse sampler under input perturbations and implementation errors. We therefore abstract the reverse sampler as a discrete-time Markov chain and explicitly track how errors propagate through its composition.

3.3. Background: discrete-time diffusion dynamics and reverse updates

Forward noising. Let $\{\beta_t\}_{t=1}^T \subset (0, 1)$, define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. A forward diffusion admits the representation $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z$, where $z \sim \mathcal{N}(0, I_d)$, equivalently induced by a Markov chain $q(x_t \mid x_{t-1})$. In purification defenses, the (possibly adversarial) input $z = x^{\mathrm{adv}}$ is first mapped to a noisy latent $x_T \sim q_T(\cdot \mid z)$, and then denoised by a reverse sampler.

Reverse sampling as a Markov chain. We write the reverse sampler as

$$x_{t-1} = \Psi_t(x_t; \theta, \phi) + \tilde{\sigma}_t \xi_t, \xi_t \sim \mathcal{N}(0, I_d), \text{ for } t = T, \ldots, 1, \tag{3}$$

where Ψ_t is the deterministic drift determined by the learned score/noise-prediction model (parameters θ), optional guidance (parameters ϕ), and the chosen discretization/solver. The scalar $\tilde{\sigma}_t \ge 0$ captures the reverse-noise level (including deterministic samplers with $\tilde{\sigma}_t = 0$).

Canonical drift decomposition and robustness levers. A useful decomposition is

$$\Psi_t(x) = a_t x + b_t s_\theta(x,t) + g_t \,\mathrm{Guide}_\phi(x,t), \tag{4}$$

where $s_\theta(\cdot, t)$ denotes a score-like drift (possibly reparameterized from an ε-prediction network) and $\mathrm{Guide}_\phi(\cdot, t)$ denotes a guidance signal (e.g., classifier guidance $\nabla_x \log p_\phi(y \mid x)$). The coefficients (a_t, b_t, g_t) encode the sampler and connect the theory to step sizes and guidance scales. Equation (4) highlights two distinct robustness levers. First, *multi-step (non-)contraction* is governed by local expansion factors of Ψ_t (e.g., through $\|J\Psi_t(x)\|_{\mathrm{op}}$), which depend on the Jacobians of s_θ and Guide_ϕ and are amplified by large guidance scales $|g_t|$. Second, *injected error* arises from imperfect score/guidance models and numerical discretization, and accumulates through composition.

3.4. Ideal reverse dynamics and an error-decomposed reverse model

Let Ψ_t^* denote the *ideal* reverse deterministic update. In practice, the implemented drift differs due to three dominant sources: learned-model error, guidance fragility, and numerical error, i.e.,

$$\Psi_t(x) = \Psi_t^*(x) + e_t(x) \text{ and } e_t(x) = e_t^{(s)}(x) + e_t^{(g)}(x) + e_t^{(\mathrm{disc})}(x), \tag{5}$$

where $e_t^{(s)}$ (*score error*) captures approximation/optimization error in learning s_θ (or ε_θ), $e_t^{(g)}$ (*guidance error*) captures mismatch or instability of guidance (including adversarially induced misalignment or label flips), and $e_t^{(\mathrm{disc})}$ (*discretization error*) captures solver/time-discretization error and any mismatch between the assumed and implemented sampler. Many implementation differences (e.g., step-size changes, guidance scaling, solver approximation) can be expressed as an additive drift perturbation after reparameterization, up to higher-order terms. Our bounds therefore accommodate a broad class of samplers without committing to a specific solver.

3.5. Robustness and stability metrics

Because $\mathrm{Pur}(\cdot; \omega)$ is randomized, we distinguish *pathwise* and *distributional* stability. *(i) Pathwise purification stability.* For a fixed randomness realization ω, define the ℓ_2 sensitivity

$$\Delta_{\mathrm{Pur}}^{\mathrm{path}}(x, \varepsilon; \omega) \triangleq \sup_{\|\delta\|_p \le \varepsilon} \big\| \mathrm{Pur}(x+\delta; \omega) - \mathrm{Pur}(x; \omega) \big\|_2. \tag{6}$$

When comparing two reverse chains, we couple them using the same Gaussian seeds so that any output separation is attributable to drift differences and injected errors rather than independent noise. *(ii) Distributional purification stability.* Let $\mathcal{L}(\mathrm{Pur}(x))$ denote the output distribution. The metric

$$\Delta_{\mathrm{Pur}}^{\mathrm{dist}}(x, \varepsilon) \triangleq \sup_{\|\delta\|_p \le \varepsilon} W_2\big(\mathcal{L}(\mathrm{Pur}(x+\delta)), \mathcal{L}(\mathrm{Pur}(x))\big), \tag{7}$$

well matches diffusion-sampler analyses and can be upper bounded by explicit couplings.

For purification followed by a classifier, robust accuracy at radius ε is $\mathrm{RA}(\varepsilon) \triangleq \Pr_{(x,y)\sim\mathcal{D}}[\forall \delta : \|\delta\|_p \le \varepsilon, c(\mathrm{Pur}(x+\delta; \omega)) = y]$, where the probability is over (x, y) and defense randomness ω.

4. Impossibility Results: Error Amplification Imposes Robustness Limits

This section formalizes a fundamental limitation: *without stability structure in the reverse dynamics, diffusion-based defenses can be inherently non-robust*. The proofs are provided in the Appendix.

4.1. Lower-bound setup: synchronous coupling and error budgets

We study the randomized purification map $\mathrm{Pur}(\cdot;\omega)$ induced by the reverse chain (3). To isolate deterministic amplification from stochasticity, we use synchronous coupling: for two trajectories $\{x_t\}_{t=0}^T$ and $\{\tilde{x}_t\}_{t=0}^T$, we drive both chains with the same Gaussian seeds $\{\xi_t\}_{t=1}^T$, and $x_{t-1} = \Psi_t(x_t) + \tilde{\sigma}_t\xi_t$, $\tilde{x}_{t-1} = \Psi_t(\tilde{x}_t) + \tilde{\sigma}_t\xi_t$, for $t = T, \dots, 1$. Under this coupling, reverse noise cancels in the difference, so any separation is due solely to drift geometry and drift errors.

Fix a region $\mathcal{X} \subset \mathbb{R}^d$ containing the reverse trajectories. Recall $\Psi_t = \Psi_t^* + e_t$ from (5). The quantities that govern coupled differences are Lipschitz of Ψ_t^* on $\mathcal{X}$ and the oscillation of e_t on $\mathcal{X}$.

Assumption 1 (Stepwise expansion and error oscillation on a region) *For each $t \in \{1, \dots, T\}$, the ideal update Ψ_t^* is L_t-Lipschitz on $\mathcal{X}$, i.e., $\|\Psi_t^*(u) - \Psi_t^*(v)\|_2 \le L_t\|u - v\|_2$, $\forall u, v \in \mathcal{X}$. Moreover, the drift perturbation $e_t : \mathcal{X} \to \mathbb{R}^d$ has bounded oscillation, $\sup_{u,v\in\mathcal{X}} \|e_t(u) - e_t(v)\|_2 \le \Delta_t$.*

Assumption 1 imposes no contraction requirement, and it makes no statistical assumptions on e_t. It is therefore well suited for worst-case impossibility statements. Note that if one instead assumes a uniform magnitude bound $\sup_{x\in\mathcal{X}} \|e_t(x)\|_2 \le \bar{\Delta}_t$, then the oscillation budget satisfies $\Delta_t \le 2\bar{\Delta}_t$.

4.2. Tight pathwise instability lower bound

Let $d_t \triangleq x_t - \tilde{x}_t$ denote the coupled difference. Under synchronous coupling and the decomposition $\Psi_t = \Psi_t^* + e_t$, we have $d_{t-1} = \Psi_t^*(x_t) - \Psi_t^*(\tilde{x}_t) + (e_t(x_t) - e_t(\tilde{x}_t))$. Thus, expansive steps in Ψ_t^* and persistent oscillations in e_t can accumulate through composition.

Theorem 1 (Tight error-amplification lower bound) *Fix any horizon T and any nonnegative sequences $\{L_t\}_{t=1}^T$ and $\{\Delta_t\}_{t=1}^T$. There exist a dimension d, a region $\mathcal{X} \subset \mathbb{R}^d$, an ideal drift $\{\Psi_t^*\}$, and perturbations $\{e_t\}$ satisfying Assumption 1, together with initial states $x_T, \tilde{x}_T \in \mathcal{X}$, such that the synchronously coupled chains satisfy*

$$\|x_0 - \tilde{x}_0\|_2 \ge \Big(\prod_{t=1}^T L_t\Big)\|x_T - \tilde{x}_T\|_2 - \sum_{k=1}^T \Big(\prod_{t=1}^{k-1} L_t\Big)\Delta_k. \tag{8}$$

Moreover, the bound is tight. *Equality holds for a one-dimensional linear instance with appropriately aligned (piecewise-constant) perturbations.*

Proof (Proof sketch) We give an explicit tight construction. Let $d = 1$, $\mathcal{X} = \mathbb{R}$, and define $\Psi_t^*(u) = L_t u$. Choose perturbations $e_t(u) = -(\Delta_t/2)\,\mathrm{sgn}(u)$, for which $\sup_{u,v} |e_t(u) - e_t(v)| = \Delta_t$. Initialize $x_T = r$ and $\tilde{x}_T = -r$ for any $r > 0$, so $d_T = 2r > 0$ and the signs of $x_t, \tilde{x}_t$ remain opposite for all t (since Ψ_t^* is sign-preserving). Then, $d_{t-1} = L_t d_t + \big(e_t(x_t) - e_t(\tilde{x}_t)\big) = L_t d_t - \Delta_t$, hence $|d_{t-1}| = L_t|d_t| - \Delta_t$ holds with equality. Unrolling the recursion yields (8). ■

Theorem 1 isolates a necessary obstruction to uniform *pathwise* stability of $\mathrm{Pur}(\cdot;\omega)$. If the reverse drift admits steps with expansion factors $L_t > 1$ for a nontrivial portion of t and the oscillation budgets Δ_t do not decay fast enough, then there exist inputs (and admissible reverse-process errors) for which the purified outputs exhibit macroscopic deviations under synchronous coupling.

4.3. A guidance-specific impossibility: basin switching under adversarial perturbations

We next isolate a distinct failure mode that is specific to guided diffusion defenses. Guidance induces a *state-dependent* drift field. Near guidance decision boundaries, the guidance direction can flip rapidly (or discontinuously). An adversary can exploit this by causing the *noised latent* to fall on different sides of a boundary with nontrivial probability, after which a basin-attractive reverse dynamics steers the trajectories toward macroscopically separated outputs. The resulting instability is *distributional* (in the law of $\mathrm{Pur}(x)$), even when the perturbation is arbitrarily small.

A minimal two-basin construction. Fix a unit vector $v \in \mathbb{R}^d$, define the separator $H \triangleq \{x : \langle v, x\rangle = 0\}$ and basins $\mathcal{X}_+ = \{x : \langle v, x\rangle > 0\}$ and $\mathcal{X}_- = \{x : \langle v, x\rangle < 0\}$. Let two target modes be $m_+ = mv$ and $m_- = -mv$ with separation $\Delta_m \triangleq \|m_+ - m_-\|_2 = 2m$.

Theorem 2 (Guidance-induced basin switching yields distributional instability) *Fix $p \in [1, \infty]$ and $\varepsilon > 0$. Consider the purification defense, i.e., given input $z \in \mathbb{R}^d$, sample $x_T = z + \sigma z_0$, $z_0 \sim \mathcal{N}(0, I_d)$, for some $\sigma > 0$, and output $\mathrm{Pur}(z; z_0) = m \cdot \mathrm{sgn}\big(\langle v, x_T\rangle\big)v$. Then, there exist $x \in \mathbb{R}^d$ and $\delta \in \mathbb{R}^d$ with $\|\delta\|_p \le \varepsilon$ such that, letting $\mu = \mathcal{L}(\mathrm{Pur}(x))$ and $\nu = \mathcal{L}(\mathrm{Pur}(x + \delta))$ denote the output laws induced by z_0, we have*

$$W_2(\mu, \nu) = 2m\sqrt{2\Phi(\varepsilon/(2\sigma)) - 1}, \tag{9}$$

where Φ is the standard normal CDF. In particular, there exist constants $c_0, c_1 > 0$ such that

$$W_2(\mu, \nu) \ge c_0 m, \text{ for } \varepsilon/\sigma \ge c_1. \tag{10}$$

Consequently, $\Delta^{\mathrm{dist}}_{\mathrm{Pur}}(x, \varepsilon) \ge W_2(\mu, \nu)$, so basin switching can induce macroscopic output-law instability when the forward-noising scale is comparable to the threat radius.

Proof (Proof sketch) Choose $x = \frac{\varepsilon}{2}v$ and $\delta = -\varepsilon v$ so that $\|\delta\|_p = \varepsilon$ for all p. Let $G \triangleq \langle v, z_0\rangle \sim \mathcal{N}(0, 1)$ and define $t \triangleq \varepsilon/(2\sigma)$. Then, $\mathrm{Pur}(x) = m_+$ iff $G \ge -t$, so $\Pr(\mathrm{Pur}(x) = m_+) = \Phi(t)$, while $\Pr(\mathrm{Pur}(x + \delta) = m_+) = \Phi(-t) = 1 - \Phi(t)$. Thus, μ and ν are two-point mixtures on $\{m_+, m_-\}$ with swapped weights, and Lemma 3 yields (9). The lower bound (10) follows by choosing t to be a fixed constant. ■

Lemma 3 *Let $a, b \in \mathbb{R}^d$ and $\mu = p\delta_a + (1 - p)\delta_b$, $\nu = q\delta_a + (1 - q)\delta_b$ for $p, q \in [0, 1]$. Then, we have $W_2(\mu, \nu)^2 = \|a - b\|_2^2\,|p - q|$.*

Proof (Proof sketch) Assume wlog $p \ge q$. The optimal transport matches q mass at a and $(1 - p)$ mass at b at zero cost. The remaining mass $p - q$ must move from a to b, incurring cost $\|a - b\|_2^2$ per unit mass. Hence, the optimal cost is $\|a - b\|_2^2|p - q|$. ■

The instability magnitude is controlled by the mode separation $\|m_+ - m_-\|_2$ and the boundary-crossing probability induced by forward noising, captured by the ratio ε/σ. If $\sigma \gg \varepsilon$, then $2\Phi(\varepsilon/(2\sigma)) - 1 \approx \Theta(\varepsilon/\sigma)$ and the lower bound is necessarily small. Strong forward noise makes the latent nearly insensitive to a small adversarial perturbation. Theorem 2 highlights the practically relevant regime $\sigma = \Theta(\varepsilon)$, where the boundary-crossing probability is constant and a sharp guidance boundary can yield $\Omega(\|m_+ - m_-\|_2)$ distributional shifts.

Algorithm 1 ECRD: Error-Calibrated Robust Diffusion (reverse sampler)

```
Require: Input $x^{\mathrm{adv}}$, horizon $T$, schedule $\{a_t, b_t, g_t, \tilde{\sigma}_t\}_{t=1}^T$, models $s_\theta(\cdot, t)$ and $\mathrm{Guide}_\phi(\cdot, t)$, error budgets $\{\Delta_t^{(s)}, \Delta_t^{(g)}\}_{t=1}^T$, target contraction $\rho \in (0, 1)$, damping floor $\eta_{\min} \in (0, 1]$, neighborhood radius $\{r_t\}_{t=1}^T$, probes $J$, power-iterations $K$, slack $\{\gamma_t\}_{t=1}^T$, restarts $M$.
Sample $x_T \sim q_T(\cdot \mid x^{\mathrm{adv}})$ via forward noising (or set $x_T = x^{\mathrm{adv}}$ if purification starts at $T$).
for $m = 1, \dots, M$ do
    $x_T^{(m)} \leftarrow x_T$.
    for $t = T, T-1, \dots, 1$ do
        Set attenuation weights $w_t^{(s)}, w_t^{(g)}$ by (11) and form $\Psi_t^{(w)}$ by (12).
        Form probe points $x_{t,j} = x_t^{(m)} + r_t u_j$ ($u_0 = 0$, $u_j$ random unit vectors, $j = 1, \dots, J$).
        Compute $\widehat{L}_t^{(m)} = \max_{0 \le j \le J} \widehat{\ell}_t(x_{t,j}) + \gamma_t$, with $\widehat{\ell}_t(\cdot)$ from (15).
        Set $\eta_t \leftarrow \max\{\eta_{\min}, \min\{1, \rho/\widehat{L}_t^{(m)}\}\}$ and form $\Psi_t^{(\mathrm{calib})}$ via (13).
        Sample $\xi_t \sim \mathcal{N}(0, I_d)$ and update $x_{t-1}^{(m)} \leftarrow \Psi_t^{(\mathrm{calib})}(x_t^{(m)}) + \tilde{\sigma}_t \xi_t$.
    end for
end for
Aggregate $\{x_0^{(m)}\}_{m=1}^M$ to output $\hat{x}$ (e.g., coordinate-wise median/trimmed mean).
return $\hat{x}$
```

5. New Theory and Algorithm: Error-Calibrated Robust Diffusion (ECRD)

We design ECRD (Algorithm 1) to solve the failure mechanisms in Section 4 with two algorithmic principles. (i) *Control expansion (stability calibration):* Enforce a per-step contraction level by damping the reverse drift using a certified Lipschitz proxy. (ii) *Control injected error (error-aware attenuation):* Down-weight unreliable drift components (score and guidance) using stepwise error budgets, and train guidance to be locally invariant under the threat model. We emphasize that ECRD is a drop-in robustification of existing reverse samplers. It does not assume a particular solver, and the reverse noise level $\tilde{\sigma}_t$ can be chosen to match the desired sampling behavior.

5.1. Reverse-step parameterization and the quantity we certify

We use the canonical drift decomposition from (4). The lower bound in Theorem 1 suggests that *trajectory-wise* control of local expansion is essential. Rather than assuming a global Lipschitz constant for Ψ_t, ECRD aims to certify an upper bound on a *region-wise* Lipschitz factor around the current trajectory point. Specifically, for a measurable set $\mathcal{X}_t \subset \mathbb{R}^d$, define $L_t(\mathcal{X}_t) \triangleq \sup_{x \in \mathcal{X}_t} \|J\Psi_t(x)\|_{\mathrm{op}}$, where $J\Psi_t(x)$ is the Jacobian and $\|\cdot\|_{\mathrm{op}}$ is the spectral (operator) norm. If $L_t(\mathcal{X}_t) \le \rho < 1$ holds along the trajectory-relevant region, then Ψ_t is contractive on that region and multi-step amplification is suppressed. ECRD implements this by attenuating unreliable components and damping the resulting map using a certified proxy $\widehat{L}_t \gtrsim L_t(\mathcal{X}_t)$.

5.2. Error-aware attenuation: reliability weights in $[0, 1]$

Let $\Delta_t^{(s)}$ and $\Delta_t^{(g)}$ denote stepwise error budgets for the score and guidance components, respectively (see Section 5.5). ECRD converts these budgets into *attenuation* weights in $[0, 1]$, i.e.,

$$w_t^{(s)} \triangleq \tau/(\Delta_t^{(s)} + \tau), w_t^{(g)} \triangleq \tau/(\Delta_t^{(g)} + \tau), \tau > 0, \tag{11}$$

so larger error budgets imply smaller weights. We then form the weighted drift

$$\Psi_t^{(w)}(x) \triangleq a_t x + b_t w_t^{(s)} s_\theta(x,t) + g_t w_t^{(g)}\, \mathrm{Guide}_\phi(x,t). \tag{12}$$

Under synchronous coupling, drift errors enter through differences $e_t(x_t) - e_t(\tilde{x}_t)$. Attenuation reduces the influence of components whose errors (and/or sensitivity) are large, and it also shrinks the Jacobian contribution of those components, thereby improving both terms in the stability recursion.

5.3. Stability calibration by damping

Even after attenuation, $\Psi_t^{(w)}$ may be locally expansive. ECRD enforces a target per-step contraction level $\rho \in (0,1)$ via a damped update:

$$\Psi_t^{(\mathrm{calib})}(x) \triangleq (1-\eta_t)x + \eta_t \Psi_t^{(w)}(x), \tag{13}$$

where $\eta_t \in [\eta_{\min}, 1]$ is chosen using a certified expansion proxy $\widehat{L}_t$, $\eta_t \triangleq \max\{\eta_{\min}, \min\{1, \rho/\widehat{L}_t\}\}$. If $\widehat{L}_t \geq L_t(\mathcal{X}_t)$ for a region $\mathcal{X}_t$, then $\sup_{x\in\mathcal{X}_t} \|J\Psi_t^{(\mathrm{calib})}(x)\|_{\mathrm{op}} \leq (1-\eta_t) + \eta_t\, L_t(\mathcal{X}_t) \leq (1-\eta_t) + \eta_t \widehat{L}_t \leq \rho$, so $\Psi_t^{(\mathrm{calib})}$ is ρ-Lipschitz on $\mathcal{X}_t$. In addition, damping scales down *all* additive drift errors, including discretization error. If $\Psi_t = \Psi_t^* + e_t$, then $\Psi_t^{(\mathrm{calib})} = (1-\eta_t)x + \eta_t \Psi_t^*(x) + \eta_t e_t(x)$, so the effective perturbation becomes $\eta_t e_t$. The calibrated reverse sampler used in ECRD is

$$x_{t-1} = \Psi_t^{(\mathrm{calib})}(x_t) + \tilde{\sigma}_t \xi_t, \xi_t \sim \mathcal{N}(0, I_d), \tag{14}$$

where $\tilde{\sigma}_t$ matches the desired sampling regime (stochastic or deterministic).

5.4. Estimating a certified expansion proxy $\widehat{L}_t$

A pointwise Jacobian norm $\|J\Psi_t^{(w)}(x_t)\|_{\mathrm{op}}$ is insufficient for a Lipschitz guarantee on arbitrary pairs. ECRD therefore targets a *region-wise* proxy $\widehat{L}_t \gtrsim L_t(\mathcal{X}_t)$ for a small neighborhood $\mathcal{X}_t$ around the current iterate. Given the current state $x_t^{(m)}$, define a local region $\mathcal{X}_t^{(m)} \triangleq B_2(x_t^{(m)}, r_t)$, where $r_t > 0$ is a tunable radius. ECRD estimates $\widehat{L}_t^{(m)}$ as an upper bound on $\sup_{x\in\mathcal{X}_t^{(m)}} \|J\Psi_t^{(w)}(x)\|_{\mathrm{op}}$. We use J probe points $x_{t,j} = x_t^{(m)} + r_t u_j$ with $u_0 = 0$ and u_j i.i.d. uniform on the unit sphere (or Rademacher-normalized), compute a local spectral-norm estimate at each probe, and take the maximum plus a small slack $\gamma_t \geq 0$, $\widehat{L}_t^{(m)} \triangleq \max_{0\leq j\leq J} \widehat{\ell}_t(x_{t,j}) + \gamma_t$, $\widehat{\ell}_t(x) \approx \|J\Psi_t^{(w)}(x)\|_{\mathrm{op}}$. Each $\widehat{\ell}_t(x)$ is computed by K power iterations using Jacobian-vector products (JVP) and vector-Jacobian products (VJP), i.e., $A_t(x) \triangleq J\Psi_t^{(w)}(x)^\top J\Psi_t^{(w)}(x)$, and

$$v^{(0)} \sim \mathcal{N}(0, I_d), v^{(k+1)} \leftarrow \frac{A_t(x)v^{(k)}}{\|A_t(x)v^{(k)}\|_2}, \widehat{\ell}_t(x) \leftarrow \sqrt{(v^{(K)})^\top A_t(x) v^{(K)}}, \tag{15}$$

implemented by computing $u = J\Psi_t^{(w)}(x)\, v$ (JVP) and then $A_t(x)v = J\Psi_t^{(w)}(x)^\top u$ (VJP).

Per reverse step and per restart, the overhead is $O(JK)$ JVP/VJP passes through $\Psi_t^{(w)}$. The total overhead is $O(MTJK)$ passes. When computation is constrained, one may set $J = 0$ (pointwise proxy) and increase γ_t as a safety margin, or use a finite-difference directional proxy.

5.5. Robust guidance training and setting $\Delta_t^{(g)}$

Section 4.3 shows that guidance can induce basin switching when it is steep or unstable *near decision boundaries*. ECRD reduces guidance error by discrepancy-regularized training, i.e.,

$$\min_\phi \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\ell(f_\phi(x),y)+\lambda\cdot\max_{\|\delta\|_p\le\varepsilon} D\big(f_\phi(x),f_\phi(x+\delta)\big)\Big], \tag{16}$$

where D can be KL divergence, ℓ_2 on logits, or Jensen–Shannon. This objective penalizes changes in the classifier outputs under ℓ_p perturbations, which in turn stabilizes guidance signals derived from f_ϕ (e.g., $\mathrm{Guide}_\phi(x,t)=\nabla_x \log p_\phi(y\mid x)$) under standard smoothness assumptions.

To mitigate occasional underestimation of $\widehat{L}_t$ (or rare large errors), ECRD uses M restarts and aggregates $\{x_0^{(m)}\}_{m=1}^M$ with robust rules such as coordinate-wise median or trimmed mean, which are well aligned with stability-based guarantees.

6. Achievable Guarantees: Certified Stability Under Calibrated Steps

This section proves that ECRD yields provable purification stability. Throughout, we use the norm-conversion constant $C_{p,d}\triangleq d^{\left(\frac{1}{2}-\frac{1}{p}\right)_+}$, so that $\|u\|_2\le C_{p,d}\|u\|_p$, $\forall u\in\mathbb{R}^d$, $\forall p\in[1,\infty]$.

6.1. Calibration implies contraction (region-wise)

We restate the contraction certificate in a form aligned with the algorithm: $\widehat{L}_t$ must upper bound the maximal Jacobian norm on a trajectory-relevant neighborhood (not merely at a single point).

Assumption 2 (Conservative region-wise expansion proxy) *Fix a step t and a measurable region $\mathcal{X}_t\subset\mathbb{R}^d$. The estimator returns $\widehat{L}_t$ such that, with probability at least $1-\delta_t$,*

$$\widehat{L}_t\ge\sup_{x\in\mathcal{X}_t}\|J\Psi_t^{(w)}(x)\|_{\mathrm{op}}. \tag{17}$$

Lemma 4 (Damping yields a contractive step on $\mathcal{X}_t$) *Fix t and let $\Psi_t^{(\mathrm{calib})}$ be defined by (13). Under Assumption 2, on the event (17) we have, for all $u,v\in\mathcal{X}_t$,*

$$\|\Psi_t^{(\mathrm{calib})}(u)-\Psi_t^{(\mathrm{calib})}(v)\|_2\le\big((1-\eta_t)+\eta_t\widehat{L}_t\big)\|u-v\|_2\ \le\ \rho\,\|u-v\|_2. \tag{18}$$

Equivalently, $\Psi_t^{(\mathrm{calib})}$ is ρ-Lipschitz on $\mathcal{X}_t$.

Let $\mathcal{E}_t$ denote the event (17) at step t and $\mathcal{E}\triangleq\cap_{t=1}^T\mathcal{E}_t$. By a union bound, $\Pr(\mathcal{E})\ge 1-\sum_{t=1}^T\delta_t$.

6.2. Pathwise purification stability under calibrated contractions

We analyze two reverse chains driven by the same Gaussian seeds, so reverse noise cancels in the difference. We allow post-calibration drift errors induced by score/guidance/solver imperfections.

Assumption 3 (Trajectory containment and post-calibration error oscillation) *There exist regions $\mathcal{X}_t\subset\mathbb{R}^d$ such that, for the coupled chains of interest, $x_t,\tilde{x}_t\in\mathcal{X}_t$ for all t. Moreover, there exist reference maps $\Psi_t^{(\mathrm{calib}),*}:\mathcal{X}_t\to\mathbb{R}^d$ and perturbations $\bar{e}_t:\mathcal{X}_t\to\mathbb{R}^d$ such that $\Psi_t^{(\mathrm{calib})}(x)=\Psi_t^{(\mathrm{calib}),*}(x)+\bar{e}_t(x),\forall x\in\mathcal{X}_t$, and the perturbations have bounded oscillation on $\mathcal{X}_t$, $\sup_{u,v\in\mathcal{X}_t}\|\bar{e}_t(u)-\bar{e}_t(v)\|_2\le\bar{\Delta}_t$.*

Theorem 5 (Pathwise stability under calibrated contractive steps) *Consider two reverse chains $\{x_t\}_{t=0}^T$ and $\{\tilde{x}_t\}_{t=0}^T$ following (14) under synchronous coupling. Suppose Assumptions 2. Then, on the event $\mathcal{E}$, for any starting points $x_T, \tilde{x}_T$, we have*

$$\|x_0 - \tilde{x}_0\|_2 \le \rho^T \|x_T - \tilde{x}_T\|_2 + \sum\nolimits_{k=1}^T \rho^{k-1} \bar{\Delta}_k \le \rho^T \|x_T - \tilde{x}_T\|_2 + (1-\rho^T)\bar{\Delta}_{\max}/(1-\rho), \quad (19)$$

where $\bar{\Delta}_{\max} \triangleq \max_{1\le k\le T} \bar{\Delta}_k$.

Proof Let $d_t \triangleq x_t - \tilde{x}_t$. Under synchronous coupling, the Gaussian terms cancel and $d_{t-1} = \Psi_t^{(\mathrm{calib})}(x_t) - \Psi_t^{(\mathrm{calib})}(\tilde{x}_t)$. Using Assumption 3 and triangle inequality, $\|d_{t-1}\|_2 \le \|\Psi_t^{(\mathrm{calib}),*}(x_t) - \Psi_t^{(\mathrm{calib}),*}(\tilde{x}_t)\|_2 + \|\bar{e}_t(x_t) - \bar{e}_t(\tilde{x}_t)\|_2$. On $\mathcal{E}$, Lemma 4 gives ρ-Lipschitzness of $\Psi_t^{(\mathrm{calib})}$ on $\mathcal{X}_t$; applying it to the reference map $\Psi_t^{(\mathrm{calib}),*}$ (which shares the same damping choice) yields $\|\Psi_t^{(\mathrm{calib}),*}(x_t) - \Psi_t^{(\mathrm{calib}),*}(\tilde{x}_t)\|_2 \le \rho \|d_t\|_2$ for $x_t, \tilde{x}_t \in \mathcal{X}_t$. Assumption 3 gives $\|\bar{e}_t(x_t) - \bar{e}_t(\tilde{x}_t)\|_2 \le \bar{\Delta}_t$. Thus, $\|d_{t-1}\|_2 \le \rho\|d_t\|_2 + \bar{\Delta}_t$, and unrolling yields (19). ■

6.3. Distributional stability and end-to-end purification sensitivity

Theorem 5 is pathwise under synchronous coupling. It immediately yields a clean distributional consequence when the calibration certificates hold (deterministically or with very high probability).

Corollary 6 (Wasserstein stability via synchronous coupling) *Fix $x_T, \tilde{x}_T$ and consider the output laws $\mu = \mathcal{L}(\mathrm{Pur}(x))$ and $\nu = \mathcal{L}(\mathrm{Pur}(\tilde{x}))$ induced by all defense randomness. If (19) holds almost surely over the defense randomness, then $W_2(\mu, \nu) \le \rho^T \|x_T - \tilde{x}_T\|_2 + \sum_{k=1}^T \rho^{k-1} \bar{\Delta}_k$. More generally, if (19) holds on an event $\mathcal{E}$ with probability at least $1 - \delta_{\mathrm{tot}}$ and $\|\mathrm{Pur}(\cdot)\|_2 \le R$ almost surely for the inputs considered, then $W_2(\mu, \nu) \le \rho^T \|x_T - \tilde{x}_T\|_2 + \sum_{k=1}^T \rho^{k-1} \bar{\Delta}_k + 2R\sqrt{\delta_{\mathrm{tot}}}$.*

Under the forward marginal representation, a natural coupling for forward noising uses the same Gaussian z for both inputs, which yields $x_T(x+\delta) - x_T(x) = \sqrt{\bar{\alpha}_T}\delta$. Therefore, for any $\|\delta\|_p \le \varepsilon$,

$$\|x_T(x+\delta) - x_T(x)\|_2 \le \sqrt{\bar{\alpha}_T} C_{p,d} \varepsilon. \quad (20)$$

Combining (20) with Theorem 5 yields a sufficient bound on $\Delta_{\mathrm{Pur}}^{\mathrm{path}}(x, \varepsilon; \omega)$ in terms of $(\rho, \bar{\Delta}_t)$.

6.4. From stability to robust classification and certification

Theorem 5 controls purification sensitivity, which can be converted into robustness of downstream.

Proposition 7 (Robust classification via stability and margin) *Let $c : \mathbb{R}^d \to \{1, \ldots, K\}$ be a classifier whose score map is L_c-Lipschitz in ℓ_2, and suppose the score margin at $\mathrm{Pur}(x;\omega)$ is at least $\gamma(x) > 0$ (top-1 score exceeds runner-up by $\gamma(x)$). If, for a fixed ω,*

$$\sup\nolimits_{\|\delta\|_p \le \varepsilon} \|\mathrm{Pur}(x+\delta;\omega) - \mathrm{Pur}(x;\omega)\|_2 \le \gamma(x)/(2L_c), \quad (21)$$

then $c(\mathrm{Pur}(x+\delta;\omega)) = c(\mathrm{Pur}(x;\omega))$ for all $\|\delta\|_p \le \varepsilon$. In particular, on the event $\mathcal{E}$, Theorem 5 and (20) provide an explicit sufficient condition in terms of ρ, $\bar{\Delta}_t$, and $\sqrt{\bar{\alpha}_T} C_{p,d} \varepsilon$.

Diffusion-smoothing certificates (DiffSmooth-style) and multi-run denoise-and-vote pipelines (DensePure-style) both benefit from concentrated, stable purified outputs. ECRD provides an explicit stability handle, control of $(\rho, \bar{\Delta}_t)$, that directly enforces such concentration under the threat model, and thereby reduces brittleness of downstream aggregation and certification.

6.5. From Discrepancy control to guidance-error budgets

We connect the discrepancy-regularized training objective (16) to a quantitative guidance error budget. Let $h_\phi(x) \in \mathbb{R}^K$ denote logits, $p_\phi(x) = \mathrm{softmax}(h_\phi(x))$, and for a class y consider guidance

$$\mathrm{Guide}_\phi(x,t) \triangleq \nabla_x \log p_\phi(y \mid x) = Jh_\phi(x)^\top \big(e_y - p_\phi(x)\big), \tag{22}$$

where $Jh_\phi(x) \in \mathbb{R}^{K\times d}$ is the Jacobian of logits and e_y is the y-th standard basis vector.

Lemma 8 (Discrepancy control to guidance-error budget) *Fix $y \in \{1,\dots,K\}$. Assume that on a region $\mathcal{X}$, $\|Jh_\phi(x)\|_{\mathrm{op}} \le B$, $\|Jh_\phi(x) - Jh_\phi(x')\|_{\mathrm{op}} \le L_J\|x - x'\|_2$, $\forall x, x' \in \mathcal{X}$. Then, we have*

$$\|\,\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t)\|_2 \le B\|p_\phi(x) - p_\phi(x')\|_2 + 2L_J\,\|x - x'\|_2. \tag{23}$$

Moreover, if D in (16) is KL *and $D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x')) \le \kappa$, then by Pinsker's inequality $\|p_\phi(x) - p_\phi(x')\|_2 \le \|p_\phi(x) - p_\phi(x')\|_1 \le \sqrt{2\kappa}$, hence $D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x')) \le \kappa$ and then*

$$\|\,\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t)\|_2 \le B\sqrt{2\kappa} + 2L_J\|x - x'\|_2. \tag{24}$$

In particular, for any $\|x - x'\|_p \le \varepsilon$ we have $\|x - x'\|_2 \le C_{p,d}\varepsilon$, so

$$\|\,\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t)\|_2 \le B\sqrt{2\kappa} + 2L_J C_{p,d}\varepsilon. \tag{25}$$

Proof From (22) and add/subtract $Jh_\phi(x)^\top(e_y - p_\phi(x'))$, we have $\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t) = Jh_\phi(x)^\top\big(p_\phi(x') - p_\phi(x)\big) + \big(Jh_\phi(x) - Jh_\phi(x')\big)^\top(e_y - p_\phi(x'))$. Bound these terms and $\|e_y - p\|_2 \le \|e_y - p\|_1 \le 2$, yielding (23). Pinsker gives $\|p_\phi(x) - p_\phi(x')\|_1 \le \sqrt{2D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x'))}$. ■

Lemma 8 provides a measurable way to translate discrepancy control into a guidance-error budget. If validation diagnostics certify $\max_{\|\delta\|_p \le \varepsilon} D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x+\delta)) \le \kappa_t$ along reverse states and if step-dependent bounds $(B_t, L_{J,t})$ hold, then one may set $\Delta_t^{(g)} = B_t\sqrt{2\kappa_t} + 2L_{J,t}C_{p,d}\varepsilon$, which then feeds into the attenuation weights (11) and the post-calibration budget.

7. Conclusion and Future Work

This work develops a stability-first theory of adversarial robustness for diffusion-based defenses by viewing the reverse sampler as a multi-step composition with *error-decomposed* drift perturbations. We prove impossibility results showing that without explicit stability structure, reverse dynamics can amplify small ℓ_p perturbations and bounded reverse-process errors into macroscopic output deviations, and we isolate a guidance-specific basin-switching mechanism that yields $\Omega(1)$ distributional instability when the forward-noising scale is comparable to the threat radius. Motivated by these limits, we propose *Error-Calibrated Robust Diffusion* (ECRD), which enforces per-step contractivity via damping calibrated by conservative region-wise expansion proxies and attenuates unreliable drift components using stepwise error budgets. Under synchronous coupling, we establish explicit pathwise and distributional stability guarantees with transparent constants, and we connect discrepancy-regularized guidance training to quantitative guidance-error budgets that enter the stability bounds. Together, our results clarify when diffusion defenses can be provably stable and provide a constructive route to robustness by jointly controlling expansion and injected error.

References

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

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. *arXiv preprint arXiv:2305.15241*, 2023.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. *Advances in Neural Information Processing Systems*, 37:50062–50097, 2024a.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024b.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

Quan Dao, Binh Ta, Tung Pham, and Anh Tran. A high-quality robust diffusion framework for corrupted dataset. In *European Conference on Computer Vision*, pages 107–123. Springer, 2024.

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

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025.

Guang Lin, Zerui Tao, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Robust diffusion models for adversarial purification. *arXiv preprint arXiv:2403.16067*, 436, 2024.

Byeonghu Na, Yeongmin Kim, HeeSun Bae, Jung Hyun Lee, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Label-noise robust diffusion models. *arXiv preprint arXiv:2402.17517*, 2024.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models for adversarial robustness. In *The Eleventh International Conference on Learning Representations*, 2023.

Jiawei Zhang, Zhongzhu Chen, Huan Zhang, Chaowei Xiao, and Bo Li. {DiffSmooth}: Certifiably robust learning via diffusion models and local smoothing. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 4787–4804, 2023.

Appendix A. Proof of Theorem 1

Proof We give an explicit one-dimensional construction and show that it attains (8) with equality whenever the right-hand side is nonnegative (otherwise the inequality is vacuous since norms are nonnegative).

Step 0: recall the perturbation assumption. Assumption 1 (as used in the paper) requires that for each t: (i) the ideal map $\Psi_t^\star$ is L_t-Lipschitz on $\mathcal{X}$ in ℓ_2, and (ii) the perturbation e_t has bounded oscillation on $\mathcal{X}$:

$$\sup_{u,v\in\mathcal{X}} \|e_t(u) - e_t(v)\|_2 \le \Delta_t.$$

(Equivalently, under synchronous coupling, the perturbation contribution per step is bounded by Δ_t.)

Step 1: if the RHS is negative, the claim is immediate. If $\prod_{t=1}^T L_t = 0$, then for any choice of $x_T, \tilde{x}_T$ the first term on the RHS of (8) vanishes and the remaining sum is nonnegative, hence the RHS is ≤ 0. Since $\|x_0 - \tilde{x}_0\|_2 \ge 0$, the inequality holds trivially. More generally, even if $\prod_{t=1}^T L_t > 0$, if we choose $x_T = \tilde{x}_T$ then the RHS equals $-\sum_{k=1}^T(\prod_{t=1}^{k-1} L_t)\Delta_k \le 0$, and the inequality again holds trivially. Therefore, the only interesting case for *tightness* is when the RHS is nonnegative, and we now give a construction that achieves equality.

Step 2: a tight one-dimensional construction. Fix any T and any nonnegative sequences $\{L_t\}_{t=1}^T$ and $\{\Delta_t\}_{t=1}^T$. We construct an instance with $d = 1$, $\mathcal{X} = \mathbb{R}$, and a deterministic reverse chain (set $\tilde{\sigma}_t \equiv 0$ in (3), which is allowed since the theorem is existential).

Define the ideal drift maps by

$$\Psi_t^\star(u) \triangleq L_t u, \qquad u \in \mathbb{R}.$$

Then $\Psi_t^\star$ is L_t-Lipschitz on $\mathbb{R}$.

Next define the perturbations by the piecewise-constant functions

$$e_t(u) \triangleq -\frac{\Delta_t}{2}\,\mathrm{sgn}(u), \qquad \mathrm{sgn}(u) \triangleq \begin{cases} +1, & u \ge 0, \\ -1, & u < 0. \end{cases}$$

Then $e_t(u) \in \{-\Delta_t/2, +\Delta_t/2\}$, and for all $u, v \in \mathbb{R}$,

$$|e_t(u) - e_t(v)| \le \Delta_t,$$

with equality when $u \ge 0$ and $v < 0$. Hence $\sup_{u,v\in\mathcal{X}} |e_t(u) - e_t(v)| = \Delta_t$, so Assumption 1 holds.

Finally choose initial states

$$x_T = r, \qquad \tilde{x}_T = -r,$$

where $r > 0$ will be chosen below. Define the (implemented) reverse dynamics

$$x_{t-1} = \Psi_t^\star(x_t) + e_t(x_t), \qquad \tilde{x}_{t-1} = \Psi_t^\star(\tilde{x}_t) + e_t(\tilde{x}_t), \qquad t = T, T-1, \dots, 1.$$

This corresponds to the synchronously coupled chain with zero reverse noise.

Step 3: ensure the signs stay separated (choose r). To obtain equality, we want $x_t \geq 0$ and $\tilde{x}_t < 0$ for all t so that $e_t(x_t) - e_t(\tilde{x}_t) = -\Delta_t$ at every step.

Assume first that $\prod_{t=1}^{T} L_t > 0$ (i.e., $L_t > 0$ for all t); this is the regime in which the RHS of (8) can be made positive by choosing a large enough initial separation. Define for $t \in \{0, 1, \dots, T\}$ the coefficients

$$A_t \triangleq \prod_{s=t+1}^{T} L_s \quad (\text{with } A_T = 1), \qquad B_t \triangleq \frac{1}{2} \sum_{k=t+1}^{T} \Big(\prod_{s=t+1}^{k-1} L_s \Big) \Delta_k \quad (\text{with } B_T = 0).$$

A direct induction shows that, as long as $x_t \geq 0$ for all steps encountered,

$$x_t = A_t r - B_t, \qquad \tilde{x}_t = -x_t, \qquad t = 0, 1, \dots, T. \tag{26}$$

Indeed, (26) holds at $t = T$. If it holds at t, then using $x_t \geq 0$ we have $x_{t-1} = L_t x_t - \Delta_t/2$ and substituting the inductive form yields the same closed form at $t-1$.

Now choose

$$r \; > \; \max_{0 \leq t \leq T} \frac{B_t}{A_t}.$$

Since $A_t > 0$ for all t under $\prod_{t=1}^{T} L_t > 0$, this choice is finite and ensures $x_t = A_t r - B_t > 0$ for all t. Consequently $\tilde{x}_t = -x_t < 0$ for all t, and therefore for every t,

$$e_t(x_t) - e_t(\tilde{x}_t) = -\frac{\Delta_t}{2} \cdot (+1) - \Big(-\frac{\Delta_t}{2} \cdot (-1) \Big) = -\Delta_t. \tag{27}$$

Step 4: exact recursion for the coupled difference. Let $d_t \triangleq x_t - \tilde{x}_t$; then $d_T = 2r$ and $d_t > 0$ for all t by the sign separation above. Moreover, for each step t,

$$\begin{aligned} d_{t-1} &= \big(\Psi_t^\star(x_t) - \Psi_t^\star(\tilde{x}_t) \big) + \big(e_t(x_t) - e_t(\tilde{x}_t) \big) \\ &= L_t(x_t - \tilde{x}_t) - \Delta_t = L_t d_t - \Delta_t, \end{aligned}$$

where we used $\Psi_t^\star(u) = L_t u$ and (27). Thus, the recursion holds *with equality*:

$$d_{t-1} = L_t d_t - \Delta_t, \qquad t = T, T-1, \dots, 1. \tag{28}$$

Step 5: unroll the recursion. Unrolling (28) gives

$$d_0 = \Big(\prod_{t=1}^{T} L_t \Big) d_T - \sum_{k=1}^{T} \Big(\prod_{t=1}^{k-1} L_t \Big) \Delta_k. \tag{29}$$

Since $d_T = \|x_T - \tilde{x}_T\|_2$ (in $d = 1$) and $d_0 = \|x_0 - \tilde{x}_0\|_2$ because $d_0 > 0$, (29) is exactly (8) with equality:

$$\|x_0 - \tilde{x}_0\|_2 = \Big(\prod_{t=1}^{T} L_t \Big) \|x_T - \tilde{x}_T\|_2 - \sum_{k=1}^{T} \Big(\prod_{t=1}^{k-1} L_t \Big) \Delta_k.$$

Step 6: conclude existence and tightness. This construction satisfies Assumption 1 and achieves (8). Moreover, when the RHS is nonnegative (in particular, when $\prod_{t=1}^{T} L_t > 0$ and r is chosen sufficiently large), the inequality is achieved with equality, showing the bound is tight. When the RHS is negative, the inequality is automatically true for any instance because $\|x_0 - \tilde{x}_0\|_2 \geq 0$. ■

Appendix B. Proof of Theorem 2

This section provides a complete proof of Theorem 2 together with intuition. The construction isolates a guidance-specific instability mechanism: a sharp decision boundary in the guided reverse map partitions the latent space into two basins that deterministically map to well-separated outputs. Even an arbitrarily small ℓ_p perturbation can flip which side of the boundary the *noised* latent lands on with nontrivial probability, creating a macroscopic shift in the output law.

B.1. Setup and geometric intuition

Fix a unit vector $v \in \mathbb{R}^d$. The purification defense first perturbs the input z by isotropic Gaussian noise and then applies a sign-thresholding map along direction v:

$$x_T \;=\; z + \sigma z_0, \qquad z_0 \sim \mathcal{N}(0, I_d), \qquad \mathrm{Pur}(z; z_0) \;=\; m \cdot \mathrm{sgn}\big(\langle v, x_T \rangle\big)\, v. \tag{30}$$

Thus, the hyperplane $H = \{u : \langle v, u \rangle = 0\}$ is a *guidance boundary*: inputs whose noised latents fall in $\mathcal{X}_+ = \{u : \langle v, u \rangle \geq 0\}$ map to the mode $m_+ = mv$, and those falling in $\mathcal{X}_- = \{u : \langle v, u \rangle < 0\}$ map to $m_- = -mv$. The outputs are separated by

$$\|m_+ - m_-\|_2 \;=\; 2m. \tag{31}$$

Adversarially perturbing the input by δ shifts the distribution of $\langle v, x_T \rangle$ by $\langle v, \delta \rangle$. By choosing δ aligned with v, the adversary maximizes the chance that the two noised latents fall on opposite sides of H, which forces the outputs to switch modes.

B.2. A lemma on Wasserstein distance for two-point mixtures

The proof uses a closed form for W_2 between two distributions supported on two points.

Lemma 9 (Exact W_2 for two-point mixtures) *Let $a, b \in \mathbb{R}^d$ and let $\mu = p\,\delta_a + (1-p)\,\delta_b$ and $\nu = q\,\delta_a + (1-q)\,\delta_b$ with $p, q \in [0,1]$. Then*

$$W_2(\mu, \nu)^2 \;=\; \|a - b\|_2^2\, |p - q|. \tag{32}$$

Proof Assume without loss of generality that $p \geq q$. Consider the coupling that matches mass q at point a and mass $(1-p)$ at point b at zero cost. The remaining unmatched mass equals $p - q$; under any coupling it must be transported from a to b, incurring squared cost $\|a - b\|_2^2$ per unit mass. Hence, the optimal transport cost equals $\|a - b\|_2^2 (p - q)$, proving (32). ■

B.3. Complete proof of Theorem 2

Proof [Proof of Theorem 2] Fix $p \in [1, \infty]$ and $\varepsilon > 0$. Choose

$$x \triangleq \frac{\varepsilon}{2} v, \qquad \delta \triangleq -\varepsilon\, v. \tag{33}$$

Then $x + \delta = -\frac{\varepsilon}{2} v$. Since v is unit norm,

$$\|\delta\|_p = \varepsilon \|v\|_p \leq \varepsilon \|v\|_2 = \varepsilon, \tag{34}$$

where we used $\|v\|_p \leq \|v\|_2$ for all $p \geq 2$, and for $p \in [1, 2)$ we can (if desired) choose $v = e_1$ to get $\|v\|_p = \|v\|_2 = 1$. In either case, the perturbation is feasible under the ℓ_p threat model.

Let $G \triangleq \langle v, z_0 \rangle$. Since $z_0 \sim \mathcal{N}(0, I_d)$ and v is unit norm, we have

$$G \sim \mathcal{N}(0, 1). \tag{35}$$

Define the scalar threshold

$$t \triangleq \frac{\varepsilon}{2\sigma}. \tag{36}$$

Step 1: compute the output distributions. For input $x = \frac{\varepsilon}{2} v$, the noised latent satisfies

$$\langle v, x_T \rangle = \left\langle v, \frac{\varepsilon}{2} v + \sigma z_0 \right\rangle = \frac{\varepsilon}{2} + \sigma \langle v, z_0 \rangle = \frac{\varepsilon}{2} + \sigma G.$$

Therefore, using the definition of Pur in (30), we have

$$\mathrm{Pur}(x; z_0) = m_+ \iff \langle v, x_T \rangle \geq 0 \iff G \geq -t,$$

so

$$\Pr(\mathrm{Pur}(x) = m_+) = \Pr(G \geq -t) = \Phi(t), \qquad \Pr(\mathrm{Pur}(x) = m_-) = 1 - \Phi(t), \tag{37}$$

where Φ is the standard normal CDF.

Similarly, for input $x + \delta = -\frac{\varepsilon}{2} v$,

$$\langle v, (x + \delta)_T \rangle = \left\langle v, -\frac{\varepsilon}{2} v + \sigma z_0 \right\rangle = -\frac{\varepsilon}{2} + \sigma G,$$

so

$$\mathrm{Pur}(x + \delta; z_0) = m_+ \iff -\frac{\varepsilon}{2} + \sigma G \geq 0 \iff G \geq t,$$

and hence

$$\Pr(\mathrm{Pur}(x+\delta) = m_+) = \Pr(G \geq t) = 1 - \Phi(t) = \Phi(-t), \qquad \Pr(\mathrm{Pur}(x+\delta) = m_-) = \Phi(t). \tag{38}$$

Equations (37)–(38) show that the two output laws are two-point mixtures with swapped weights:

$$\mu = \Phi(t)\, \delta_{m_+} + (1 - \Phi(t))\, \delta_{m_-}, \qquad \nu = (1 - \Phi(t))\, \delta_{m_+} + \Phi(t)\, \delta_{m_-}. \tag{39}$$

Step 2: compute $W_2(\mu, \nu)$ in closed form. Apply Lemma 3 with $a = m_+$, $b = m_-$, $p = \Phi(t)$, and $q = 1 - \Phi(t)$. Using (31),

$$\|a - b\|_2 = \|m_+ - m_-\|_2 = 2m, \qquad |p - q| = |\Phi(t) - (1 - \Phi(t))| = 2\Phi(t) - 1.$$

Therefore

$$W_2(\mu, \nu)^2 = (2m)^2 \, (2\Phi(t) - 1),$$

and taking square roots gives

$$W_2(\mu, \nu) \;=\; 2m\,\sqrt{2\Phi(t) - 1} \;=\; 2m\,\sqrt{2\Phi\Big(\frac{\varepsilon}{2\sigma}\Big) - 1}, \tag{40}$$

which is exactly (9).

Step 3: derive the $\Omega(m)$ lower bound. Choose any constant $t_0 > 0$ (e.g., $t_0 = 1$) and let $c_1 \triangleq 2t_0$. If $\varepsilon/\sigma \geq c_1$, then $t = \varepsilon/(2\sigma) \geq t_0$ and hence

$$2\Phi(t) - 1 \;\geq\; 2\Phi(t_0) - 1 \;\triangleq\; c_\Phi,$$

where $c_\Phi \in (0, 1)$ is a universal constant. Plugging into (40) yields

$$W_2(\mu, \nu) \;\geq\; 2m\sqrt{c_\Phi} \;\triangleq\; c_0\, m,$$

which proves (10) for constants $c_0 = 2\sqrt{2\Phi(t_0) - 1}$ and $c_1 = 2t_0$.

Step 4: relate to $\Delta^{\text{dist}}_{\text{Pur}}$. By definition,

$$\Delta^{\text{dist}}_{\text{Pur}}(x, \varepsilon) = \sup_{\|\delta'\|_p \leq \varepsilon} W_2(\mathcal{L}(\text{Pur}(x + \delta')), \mathcal{L}(\text{Pur}(x))).$$

Since the constructed δ in (33) is feasible, we have $\Delta^{\text{dist}}_{\text{Pur}}(x, \varepsilon) \geq W_2(\mu, \nu)$, completing the proof. ■

B.4. Interpretation and takeaways

What causes the instability? The reverse map in (30) has a *hard* boundary at $H = \{u : \langle v, u\rangle = 0\}$. Once the noised latent x_T crosses H, the output deterministically flips from m_+ to m_-. Thus the robustness question reduces to: how often do the two inputs produce noised latents on different sides of H under the *same* Gaussian draw? The answer is controlled by the signal-to-noise ratio $t = \varepsilon/(2\sigma)$.

Why does $\sigma = \Theta(\varepsilon)$ matter? If $\sigma \gg \varepsilon$, then $t \ll 1$ and $2\Phi(t) - 1 \approx \Theta(t)$, so $W_2(\mu, \nu)$ shrinks. In this regime, forward noise overwhelms the small adversarial shift, making basin switching rare under synchronous coupling. In contrast, if $\sigma = \Theta(\varepsilon)$, then $t = \Theta(1)$ and the crossing probability is constant, so $W_2(\mu, \nu) = \Omega(m)$: the output law changes at the scale of the mode separation $\|m_+ - m_-\|_2$.

How does this map to guided diffusion? The map (30) is a stylized proxy for guidance-dominated reverse dynamics in which the guidance term induces a steep state-dependent drift direction with a boundary separating two semantic basins. The theorem shows that even when the reverse map is "stable within each basin," the existence of a steep boundary alone is sufficient to generate macroscopic *distributional* instability under ℓ_p perturbations.

Appendix C. Proof of Lemma 3

Lemma 3 gives a closed form for the 2-Wasserstein distance between two distributions supported on the same two points. Intuitively, since both measures place all mass on $\{a, b\}$, the only possible transportation is to move some fraction of mass from one atom to the other. The optimal plan therefore matches as much mass as possible at zero cost, and transports only the *unmatched* mass across the gap $\|a - b\|_2$.

Proof Recall the definition of the squared 2-Wasserstein distance:

$$W_2(\mu, \nu)^2 \;=\; \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, \pi(\mathrm{d}x, \mathrm{d}y),$$

where $\Pi(\mu, \nu)$ denotes the set of couplings (transport plans) with marginals μ and ν.

Since μ and ν are supported on $\{a, b\}$, any coupling $\pi \in \Pi(\mu, \nu)$ is supported on the four pairs $\{(a, a), (a, b), (b, a), (b, b)\}$ and can be identified with a 2×2 nonnegative matrix

$$\pi \equiv \begin{pmatrix} \pi_{aa} & \pi_{ab} \\ \pi_{ba} & \pi_{bb} \end{pmatrix}, \qquad \pi_{ij} \geq 0,$$

where (for example) π_{ab} is the amount of mass transported from a (under μ) to b (under ν). The marginal constraints are:

$$\pi_{aa} + \pi_{ab} = p, \qquad \pi_{ba} + \pi_{bb} = 1 - p, \tag{41}$$

$$\pi_{aa} + \pi_{ba} = q, \qquad \pi_{ab} + \pi_{bb} = 1 - q. \tag{42}$$

The transport cost induced by π is

$$\begin{aligned} \int \|x - y\|_2^2 \, \mathrm{d}\pi &= \pi_{aa}\|a - a\|_2^2 + \pi_{bb}\|b - b\|_2^2 + \pi_{ab}\|a - b\|_2^2 + \pi_{ba}\|b - a\|_2^2 \\ &= (\pi_{ab} + \pi_{ba}) \, \|a - b\|_2^2, \end{aligned} \tag{43}$$

since the diagonal terms cost zero and $\|a - b\|_2^2 = \|b - a\|_2^2$.

Thus, minimizing the Wasserstein cost is equivalent to minimizing $\pi_{ab} + \pi_{ba}$ subject to (41)–(42). We now solve this exactly.

Step 1: lower bound on $\pi_{ab} + \pi_{ba}$. Subtract the first equation in (42) from the first equation in (41):

$$(\pi_{aa} + \pi_{ab}) - (\pi_{aa} + \pi_{ba}) = p - q,$$

which yields

$$\pi_{ab} - \pi_{ba} = p - q. \tag{44}$$

By the triangle inequality for real numbers,

$$|\pi_{ab} - \pi_{ba}| \leq \pi_{ab} + \pi_{ba}.$$

Combining with (44) gives the universal lower bound

$$\pi_{ab} + \pi_{ba} \;\geq\; |p - q|, \qquad \forall \pi \in \Pi(\mu, \nu). \tag{45}$$

Step 2: achievability by an explicit optimal coupling. Assume without loss of generality that $p \geq q$ (the case $p < q$ follows by symmetry). Define a coupling by

$$\pi_{aa} = q, \quad \pi_{ab} = p - q, \quad \pi_{ba} = 0, \quad \pi_{bb} = 1 - p. \tag{46}$$

It is immediate that all entries are nonnegative and that the row sums and column sums match (41)–(42): row sums are $q + (p - q) = p$ and $0 + (1 - p) = 1 - p$; column sums are $q + 0 = q$ and $(p - q) + (1 - p) = 1 - q$. Hence $\pi \in \Pi(\mu, \nu)$.

Moreover, for this coupling,

$$\pi_{ab} + \pi_{ba} = (p - q) + 0 = |p - q|,$$

which matches the lower bound (45). Therefore (46) is optimal and the optimal value is $\inf_{\pi \in \Pi(\mu,\nu)}(\pi_{ab} + \pi_{ba}) = |p - q|$.

Step 3: compute $W_2(\mu, \nu)^2$. Plugging the optimal value into (43) yields

$$W_2(\mu, \nu)^2 = \|a - b\|_2^2 \cdot |p - q|,$$

which proves (32). ■

The quantity $|p - q|$ is exactly the *unmatched mass* at atom a (equivalently at atom b) between μ and ν. Optimal transport keeps the matched mass in place (zero cost) and moves only this unmatched mass across distance $\|a - b\|_2$, so the squared cost is moved mass times squared distance.

Appendix D. Proof of Lemma 4

This appendix section provides a complete proof of Lemma 4 and explains why the calibration (damping) step is the correct stability primitive for ECRD.

D.1. Statement recap and calibration map

Fix a reverse step t and a measurable region $\mathcal{X}_t \subset \mathbb{R}^d$. Recall the weighted drift map $\Psi_t^{(w)} : \mathbb{R}^d \to \mathbb{R}^d$ and the calibrated (damped) map

$$\Psi_t^{(\mathrm{calib})}(x) \;\triangleq\; (1 - \eta_t)\, x + \eta_t\, \Psi_t^{(w)}(x), \qquad \eta_t \in (0, 1]. \tag{47}$$

Equivalently,

$$\Psi_t^{(\mathrm{calib})}(x) \;=\; x + \eta_t\big(\Psi_t^{(w)}(x) - x\big), \tag{48}$$

so η_t is a *step size* on the residual update $\Psi_t^{(w)}(x) - x$.

Assumption 2 states that with probability at least $1 - \delta_t$ (over any randomness in the estimator), the computed proxy $\widehat{L}_t$ upper bounds the maximal Jacobian operator norm of $\Psi_t^{(w)}$ on $\mathcal{X}_t$:

$$\widehat{L}_t \;\geq\; \sup_{x \in \mathcal{X}_t} \|J\Psi_t^{(w)}(x)\|_{\mathrm{op}}. \tag{49}$$

We analyze the lemma *on this event*.

D.2. A standard calculus fact: Jacobian bound implies Lipschitzness on convex regions

We will use the following standard implication.

Lemma 10 (Jacobian bound ⇒ Lipschitz on convex sets) *Let $\mathcal{X} \subset \mathbb{R}^d$ be convex. If $f : \mathcal{X} \to \mathbb{R}^d$ is continuously differentiable and $\sup_{x\in\mathcal{X}} \|Jf(x)\|_{\mathrm{op}} \le L$, then f is L-Lipschitz on $\mathcal{X}$, i.e., $\|f(u) - f(v)\|_2 \le L\|u - v\|_2$ for all $u, v \in \mathcal{X}$.*

Proof Fix $u, v \in \mathcal{X}$ and define the segment $\gamma(\tau) = v + \tau(u - v)$ for $\tau \in [0, 1]$. By convexity, $\gamma(\tau) \in \mathcal{X}$ for all τ. By the fundamental theorem of calculus,

$$f(u) - f(v) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau} f(\gamma(\tau))\,\mathrm{d}\tau = \int_0^1 Jf(\gamma(\tau))(u - v)\,\mathrm{d}\tau.$$

Taking norms and using submultiplicativity yields

$$\|f(u) - f(v)\|_2 \le \int_0^1 \|Jf(\gamma(\tau))\|_{\mathrm{op}}\,\|u - v\|_2\,\mathrm{d}\tau \le \Big(\sup_{x\in\mathcal{X}} \|Jf(x)\|_{\mathrm{op}}\Big)\|u - v\|_2 \le L\|u - v\|_2.$$

■

In our application, $\mathcal{X}_t$ is typically chosen as a ball $B_2(x_t, r_t)$ or another convex neighborhood around the current state.

D.3. Complete proof of Lemma 4

Proof [Proof of Lemma 4] Fix t and condition on the event (49). Assume $\mathcal{X}_t$ is convex and $\Psi_t^{(w)}$ is continuously differentiable on an open set containing $\mathcal{X}_t$. (These conditions are standard and are satisfied for neural-network drifts almost everywhere; see the discussion below.)

Step 1: bound the Lipschitz constant of the calibrated map. Take any $u, v \in \mathcal{X}_t$. Using the definition (47),

$$\Psi_t^{(\mathrm{calib})}(u) - \Psi_t^{(\mathrm{calib})}(v) = (1 - \eta_t)(u - v) + \eta_t\big(\Psi_t^{(w)}(u) - \Psi_t^{(w)}(v)\big). \tag{50}$$

Taking ℓ_2 norms and applying the triangle inequality gives

$$\|\Psi_t^{(\mathrm{calib})}(u) - \Psi_t^{(\mathrm{calib})}(v)\|_2 \le (1 - \eta_t)\|u - v\|_2 + \eta_t\|\Psi_t^{(w)}(u) - \Psi_t^{(w)}(v)\|_2. \tag{51}$$

By Assumption 2 on the event (49),

$$\sup_{x\in\mathcal{X}_t} \|J\Psi_t^{(w)}(x)\|_{\mathrm{op}} \le \widehat{L}_t.$$

Hence, by Lemma 10, $\Psi_t^{(w)}$ is $\widehat{L}_t$-Lipschitz on $\mathcal{X}_t$, i.e.,

$$\|\Psi_t^{(w)}(u) - \Psi_t^{(w)}(v)\|_2 \le \widehat{L}_t\,\|u - v\|_2, \qquad \forall u, v \in \mathcal{X}_t. \tag{52}$$

Substituting (52) into (51) yields the first inequality in (18):

$$\|\Psi_t^{(\mathrm{calib})}(u) - \Psi_t^{(\mathrm{calib})}(v)\|_2 \le \big((1 - \eta_t) + \eta_t\widehat{L}_t\big)\|u - v\|_2. \tag{53}$$

Step 2: enforce contraction by choosing η_t. ECRD chooses η_t to satisfy

$$\eta_t \;\le\; \min\left\{1,\; \frac{\rho-(1-\eta_t)}{\widehat{L}_t}\right\} \qquad \text{equivalently} \qquad (1-\eta_t)+\eta_t\widehat{L}_t \le \rho, \tag{54}$$

which is achieved by the explicit rule used in the algorithm:

$$\eta_t \;\triangleq\; \max\{\eta_{\min},\; \min\{1,\; \rho/\widehat{L}_t\}\}, \tag{55}$$

together with the standard convention that the calibration objective is to ensure $\big((1-\eta_t)+\eta_t\widehat{L}_t\big) \le \rho$.

Under (55) (and ignoring the optional floor $\eta_{\min}$ for the theoretical statement), we have $\eta_t\widehat{L}_t \le \rho$, and therefore

$$(1-\eta_t)+\eta_t\widehat{L}_t \;\le\; (1-\eta_t)+\rho \;\le\; \rho,$$

whenever $\eta_t = 1$ only occurs when $\widehat{L}_t \le \rho$. Thus the contraction factor in (53) is at most ρ, proving the second inequality in (18):

$$\|\Psi_t^{(\mathrm{calib})}(u)-\Psi_t^{(\mathrm{calib})}(v)\|_2 \le \rho\,\|u-v\|_2, \qquad \forall u,v\in\mathcal{X}_t. \tag{56}$$

Step 3: conclude ρ-Lipschitzness. Since the above bound holds for all $u,v\in\mathcal{X}_t$, the map $\Psi_t^{(\mathrm{calib})}$ is ρ-Lipschitz on $\mathcal{X}_t$. ■

D.4. Insights and intuition

Why does damping work? Equation (47) shows that $\Psi_t^{(\mathrm{calib})}$ is a convex combination of the identity map and the potentially expansive map $\Psi_t^{(w)}$. The identity has Lipschitz constant 1, while $\Psi_t^{(w)}$ has (certified) Lipschitz constant at most $\widehat{L}_t$ on $\mathcal{X}_t$. The Lipschitz constant of a convex combination is at most the same convex combination of Lipschitz constants, giving

$$\mathrm{Lip}(\Psi_t^{(\mathrm{calib})};\mathcal{X}_t) \;\le\; (1-\eta_t)\cdot 1+\eta_t\cdot\widehat{L}_t.$$

Thus η_t is a *stability knob*: decreasing η_t shrinks the local expansion factor toward 1, and if $\widehat{L}_t > 1$ it can be made strictly contractive by taking η_t sufficiently small.

Why region-wise rather than pointwise? Pointwise control of $\|J\Psi_t^{(w)}(x_t)\|_{\mathrm{op}}$ at a single trajectory state is not enough for a Lipschitz guarantee between two coupled chains, because the two chains may explore nearby but distinct states. Region-wise control on $\mathcal{X}_t$ ensures that (52) holds uniformly for all pairs $(u,v)\in\mathcal{X}_t\times\mathcal{X}_t$, which is exactly what is needed in the coupling recursion.

Non-smooth drifts (ReLU networks). For piecewise-linear networks, the Jacobian exists almost everywhere. The same conclusion can be stated using Rademacher's theorem: Lipschitz functions are differentiable a.e., and a bound on the essential supremum of the Jacobian norm over $\mathcal{X}_t$ implies Lipschitzness. Assumption 2 can be rephrased directly in terms of a Lipschitz proxy $\widehat{L}_t \ge \|\Psi_t^{(w)}\|_{\mathrm{Lip}(\mathcal{X}_t)}$, avoiding differentiability altogether. This preserves the lemma verbatim.

What does the contraction guarantee buy downstream? Once each calibrated step is ρ-contractive on the relevant region, the reverse chain becomes a composition of contractive maps (plus bounded error), so synchronous coupling yields geometric decay of perturbations across steps. This is the key mechanism behind the end-to-end stability bounds in Theorem 5 and is the direct antidote to the multiplicative amplification exhibited by the lower bound in Theorem 1.

Appendix E. Proof of Theorem 5

This appendix section gives a complete proof of Theorem 5 and explains the underlying mechanism. The key idea is *synchronous coupling*: when two reverse chains share the same Gaussian seeds, the stochastic terms cancel in the difference, so stability is governed entirely by (i) the Lipschitz/contractivity of the calibrated deterministic update and (ii) the oscillation of the residual (post-calibration) drift error.

E.1. Setup and the good event

Recall the calibrated reverse update (Algorithm 1):

$$x_{t-1} \;=\; \Psi_t^{(\mathrm{calib})}(x_t) + \tilde{\sigma}_t\,\xi_t, \qquad \tilde{x}_{t-1} \;=\; \Psi_t^{(\mathrm{calib})}(\tilde{x}_t) + \tilde{\sigma}_t\,\xi_t, \qquad \xi_t \sim \mathcal{N}(0, I_d), \tag{57}$$

where the *same* ξ_t is used in both chains (synchronous coupling).

Define the event $\mathcal{E}$ as the intersection of the stepwise proxy events from Assumption 2:

$$\mathcal{E} \;\triangleq\; \bigcap_{t=1}^{T} \left\{ \widehat{L}_t \;\geq\; \sup_{x \in \mathcal{X}_t} \|J\Psi_t^{(w)}(x)\|_{\mathrm{op}} \right\}. \tag{58}$$

By a union bound,

$$\Pr(\mathcal{E}) \;\geq\; 1 - \sum_{t=1}^{T} \delta_t. \tag{59}$$

Theorem 5 is stated *on* $\mathcal{E}$, i.e., conditionally on the contraction certificates being valid on the trajectory-relevant regions.

E.2. A precise contraction recursion

We restate the key assumption and show how it yields a one-step recursion on the coupled difference.

Assumption 4 (Trajectory containment and post-calibration error oscillation) *There exist regions $\mathcal{X}_t \subset \mathbb{R}^d$ such that for the coupled chains in (57), we have $x_t, \tilde{x}_t \in \mathcal{X}_t$ for all t. Moreover, there exist reference maps $\Psi_t^{(\mathrm{calib}),*} : \mathcal{X}_t \to \mathbb{R}^d$ and perturbations $\bar{e}_t : \mathcal{X}_t \to \mathbb{R}^d$ such that*

$$\Psi_t^{(\mathrm{calib})}(x) \;=\; \Psi_t^{(\mathrm{calib}),*}(x) + \bar{e}_t(x), \qquad \forall x \in \mathcal{X}_t, \tag{60}$$

and the perturbations have bounded oscillation on $\mathcal{X}_t$:

$$\sup_{u,v \in \mathcal{X}_t} \|\bar{e}_t(u) - \bar{e}_t(v)\|_2 \;\leq\; \bar{\Delta}_t. \tag{61}$$

Theorem 11 (Pathwise stability under calibrated contractive steps) *Consider two reverse chains $\{x_t\}_{t=0}^T$ and $\{\tilde{x}_t\}_{t=0}^T$ following (57) under synchronous coupling. Suppose Assumptions 2 and 4 hold. Then, on the event $\mathcal{E}$ defined in (58), for any starting points $x_T, \tilde{x}_T$, we have*

$$\|x_0 - \tilde{x}_0\|_2 \;\leq\; \rho^T \|x_T - \tilde{x}_T\|_2 \;+\; \sum_{k=1}^{T} \rho^{k-1} \bar{\Delta}_k \;\leq\; \rho^T \|x_T - \tilde{x}_T\|_2 \;+\; \frac{1-\rho^T}{1-\rho}\,\bar{\Delta}_{\max}, \tag{62}$$

where $\bar{\Delta}_{\max} \triangleq \max_{1 \leq k \leq T} \bar{\Delta}_k$.

Proof Define the coupled difference $d_t \triangleq x_t - \tilde{x}_t$. Under synchronous coupling (57), the Gaussian noise cancels:

$$d_{t-1} = \big(\Psi_t^{(\text{calib})}(x_t) + \tilde{\sigma}_t\xi_t\big) - \big(\Psi_t^{(\text{calib})}(\tilde{x}_t) + \tilde{\sigma}_t\xi_t\big) = \Psi_t^{(\text{calib})}(x_t) - \Psi_t^{(\text{calib})}(\tilde{x}_t). \tag{63}$$

Step 1: split into reference dynamics plus residual error. Using the decomposition (60) from Assumption 4, for $x_t, \tilde{x}_t \in \mathcal{X}_t$ we have

$$d_{t-1} = \Psi_t^{(\text{calib}),*}(x_t) - \Psi_t^{(\text{calib}),*}(\tilde{x}_t) \;+\; \big(\bar{e}_t(x_t) - \bar{e}_t(\tilde{x}_t)\big). \tag{64}$$

Taking ℓ_2 norms and applying the triangle inequality yields

$$\|d_{t-1}\|_2 \;\le\; \|\Psi_t^{(\text{calib}),*}(x_t) - \Psi_t^{(\text{calib}),*}(\tilde{x}_t)\|_2 \;+\; \|\bar{e}_t(x_t) - \bar{e}_t(\tilde{x}_t)\|_2. \tag{65}$$

Step 2: contractivity of the reference map on $\mathcal{E}$. On the event $\mathcal{E}$, Lemma 4 applies on each region $\mathcal{X}_t$ and yields that the calibrated map is ρ-Lipschitz on $\mathcal{X}_t$. In particular, since $\Psi_t^{(\text{calib}),*}$ shares the same damping coefficient η_t and the same region-wise certificate logic,[1] we have for all $u, v \in \mathcal{X}_t$,

$$\|\Psi_t^{(\text{calib}),*}(u) - \Psi_t^{(\text{calib}),*}(v)\|_2 \;\le\; \rho\,\|u - v\|_2. \tag{66}$$

Applying (66) with $u = x_t$ and $v = \tilde{x}_t$ gives

$$\|\Psi_t^{(\text{calib}),*}(x_t) - \Psi_t^{(\text{calib}),*}(\tilde{x}_t)\|_2 \;\le\; \rho\,\|d_t\|_2. \tag{67}$$

Step 3: bound the residual oscillation term. By Assumption 4 and since $x_t, \tilde{x}_t \in \mathcal{X}_t$,

$$\|\bar{e}_t(x_t) - \bar{e}_t(\tilde{x}_t)\|_2 \;\le\; \bar{\Delta}_t. \tag{68}$$

Step 4: one-step recursion. Substituting (67) and (68) into (65) yields the key one-step inequality:

$$\|d_{t-1}\|_2 \;\le\; \rho\,\|d_t\|_2 \;+\; \bar{\Delta}_t, \qquad t = T, T-1, \dots, 1. \tag{69}$$

Step 5: unroll the recursion. Iterating (69) yields

$$\|d_0\|_2 \le \rho^T\|d_T\|_2 + \sum_{k=1}^{T} \rho^{k-1}\bar{\Delta}_k, \tag{70}$$

which is the first inequality in (62).

Step 6: simplify via $\bar{\Delta}_{\max}$. Since $\bar{\Delta}_k \le \bar{\Delta}_{\max}$ for all k and $\sum_{k=1}^{T}\rho^{k-1} = (1-\rho^T)/(1-\rho)$,

$$\sum_{k=1}^{T} \rho^{k-1}\bar{\Delta}_k \;\le\; \bar{\Delta}_{\max}\sum_{k=1}^{T}\rho^{k-1} \;=\; \frac{1-\rho^T}{1-\rho}\,\bar{\Delta}_{\max}, \tag{71}$$

proving the second inequality in (62). Finally, note that $\|d_0\|_2 = \|x_0 - \tilde{x}_0\|_2$ and $\|d_T\|_2 = \|x_T - \tilde{x}_T\|_2$, completing the proof. ■

[1] Formally, it suffices that $\Psi_t^{(\text{calib}),*}$ is a calibrated map obtained from some $\Psi_t^{(w),*}$ whose Jacobian norm on $\mathcal{X}_t$ is also upper bounded by $\widehat{L}_t$ on $\mathcal{E}$. This is the intended meaning of “reference map” in Assumption 4.

E.3. Insights and intuition

Why synchronous coupling is the right lens. Diffusion samplers are randomized, but in robustness we care about *sensitivity to the input* rather than sensitivity to fresh randomness. Synchronous coupling compares two runs of the defense that use identical noise realizations. This removes stochastic variance from the comparison and isolates the deterministic amplification mechanism: all instability must come from expansion in the reverse drift and from drift errors.

Contraction-error tradeoff. The recursion (69) has the canonical "stable system with bounded disturbance" form: the ρ term shrinks the current separation, while $\bar{\Delta}_t$ injects new separation. Unrolling yields two terms:

- $\rho^T \|x_T - \tilde{x}_T\|_2$: the initial difference after forward noising, geometrically damped by T contractive steps;
- $\sum_{k=1}^{T} \rho^{k-1} \bar{\Delta}_k$: accumulated injected error, with earlier errors amplified less than later errors because subsequent steps contract them.

This is the precise opposite of the lower bound in Theorem 1: there, stepwise expansion factors multiply errors; here, stepwise contractions geometrically *attenuate* them.

Role of the regions $\mathcal{X}_t$. The guarantee is local in the sense that it requires: (i) the coupled trajectories remain inside $\mathcal{X}_t$ (containment), and (ii) the expansion proxy upper bounds the Jacobian norm on $\mathcal{X}_t$ (certificate). This is exactly why ECRD estimates *region-wise* expansion: it is the correct object for controlling the Lipschitz constant between two nearby trajectories.

From pathwise to distributional stability. Because (70) holds for the coupled sample paths on $\mathcal{E}$, it immediately upper bounds a coupling cost. Consequently,

$$W_2\big(\mathcal{L}(\mathrm{Pur}(x)), \mathcal{L}(\mathrm{Pur}(\tilde{x}))\big) \;\le\; \Big(\mathbb{E}\|x_0 - \tilde{x}_0\|_2^2\Big)^{1/2}$$

can be controlled by the same bound, and if $\mathcal{E}$ fails with small probability, a standard conditioning argument yields a high-probability or "in expectation and failure probability" variant.

Appendix F. Proof of Corollary 6

This appendix section provides a complete proof of Corollary 6 and explains the coupling intuition. The key observation is that W_2 is an *infimum over couplings*. Thus, any explicit coupling yields an upper bound. Synchronous coupling is particularly natural for diffusion defenses because it reuses the same Gaussian seeds, and therefore the pathwise stability bound immediately becomes a bound on the transport cost.

F.1. Preliminaries: a standard inequality for W_2

We will use the basic fact that for any coupling (X, Y) with marginals μ, ν,

$$W_2(\mu, \nu) \;\le\; \big(\mathbb{E}\|X - Y\|_2^2\big)^{1/2}. \tag{72}$$

This follows directly from the definition: $W_2(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi \|X - Y\|_2^2$, hence choosing any π gives an upper bound.

We also use the simple tail-splitting inequality: if A is an event and $Z \geq 0$ is any random variable, then

$$\mathbb{E}[Z] \;=\; \mathbb{E}[Z\mathbf{1}_A] + \mathbb{E}[Z\mathbf{1}_{A^c}] \;\leq\; \mathbb{E}[Z \mid A]\,\Pr(A) + \mathbb{E}[Z \mid A^c]\,\Pr(A^c). \tag{73}$$

F.2. Corollary statement

Corollary 12 (Wasserstein stability via synchronous coupling) *Fix $x_T, \tilde{x}_T$ and consider the output laws $\mu = \mathcal{L}(\mathrm{Pur}(x))$ and $\nu = \mathcal{L}(\mathrm{Pur}(\tilde{x}))$ induced by all defense randomness. Let*

$$B \;\triangleq\; \rho^T \|x_T - \tilde{x}_T\|_2 + \sum_{k=1}^{T} \rho^{k-1} \bar{\Delta}_k.$$

(i) If (19) holds almost surely over the defense randomness, then

$$W_2(\mu, \nu) \;\leq\; B. \tag{74}$$

(ii) More generally, if (19) holds on an event $\mathcal{E}$ with probability at least $1 - \delta_{\mathrm{tot}}$ and $\|\mathrm{Pur}(\cdot)\|_2 \leq R$ almost surely for the inputs considered, then

$$W_2(\mu, \nu) \;\leq\; B \;+\; 2R\sqrt{\delta_{\mathrm{tot}}}. \tag{75}$$

F.3. Complete proof

Proof We prove the two parts separately.

Part (i): almost sure pathwise bound ⇒ W_2 bound. Let ω denote all defense randomness (forward noising randomness, reverse-chain Gaussian seeds, and any estimator randomness). Consider the *synchronous* coupling that uses the *same* ω to generate both outputs:

$$X(\omega) \triangleq \mathrm{Pur}(x; \omega), \qquad Y(\omega) \triangleq \mathrm{Pur}(\tilde{x}; \omega).$$

By construction, $X \sim \mu$ and $Y \sim \nu$. Since (19) is assumed to hold almost surely, we have

$$\|X(\omega) - Y(\omega)\|_2 \;\leq\; B, \qquad \text{for almost every } \omega. \tag{76}$$

Therefore, we have

$$\mathbb{E}\|X - Y\|_2^2 \;\leq\; \mathbb{E}[B^2] \;=\; B^2,$$

and applying (72) yields

$$W_2(\mu, \nu) \;\leq\; \left(\mathbb{E}\|X - Y\|_2^2\right)^{1/2} \;\leq\; B,$$

which proves (74).

Part (ii): high-probability pathwise bound. Again define the synchronous coupling $(X, Y) = (\mathrm{Pur}(x; \omega), \mathrm{Pur}(\tilde{x}; \omega))$. Assume that on an event $\mathcal{E}$ of probability at least $1 - \delta_{\mathrm{tot}}$, the pathwise bound holds:

$$\|X - Y\|_2 \ \le\ B \qquad \text{on } \mathcal{E}. \tag{77}$$

Also assume $\|X\|_2 \le R$ and $\|Y\|_2 \le R$ almost surely (for the inputs considered). Then by the triangle inequality,

$$\|X - Y\|_2 \ \le\ \|X\|_2 + \|Y\|_2 \ \le\ 2R \qquad \text{almost surely,} \tag{78}$$

and hence $\|X - Y\|_2^2 \le 4R^2$ almost surely.

Split the second moment across $\mathcal{E}$ and $\mathcal{E}^c$:

$$\begin{aligned} \mathbb{E}\|X - Y\|_2^2 &= \mathbb{E}[\|X - Y\|_2^2 \mathbf{1}_{\mathcal{E}}] + \mathbb{E}[\|X - Y\|_2^2 \mathbf{1}_{\mathcal{E}^c}] \\ &\le B^2 \Pr(\mathcal{E}) + (2R)^2 \Pr(\mathcal{E}^c) \qquad \text{(by (77) and (78))} \\ &\le B^2 + 4R^2 \delta_{\mathrm{tot}}. \end{aligned} \tag{79}$$

Applying (72) yields

$$W_2(\mu, \nu) \ \le\ \sqrt{B^2 + 4R^2 \delta_{\mathrm{tot}}}. \tag{80}$$

Finally, use $\sqrt{u + v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$ to obtain

$$W_2(\mu, \nu) \le B + 2R\sqrt{\delta_{\mathrm{tot}}},$$

which is (75). ■

F.4. Intuition: why the $2R\sqrt{\delta_{\mathrm{tot}}}$ term appears

When the contraction certificate fails (event $\mathcal{E}^c$), we fall back on the trivial worst-case bound $\|X - Y\|_2 \le 2R$. Wasserstein distance depends on the *average squared transport cost* under a coupling. Thus, failure events contribute at most $(2R)^2$ cost but only with probability δ_{tot}, leading to a variance-type penalty $\sqrt{(2R)^2 \delta_{\mathrm{tot}}} = 2R\sqrt{\delta_{\mathrm{tot}}}$ after taking square roots.

F.5. Forward noising coupling

Under the standard DDPM marginal, a natural forward coupling uses the same Gaussian z for both inputs:

$$x_T(x) = \sqrt{\bar{\alpha}_T}\, x + \sqrt{1 - \bar{\alpha}_T}\, z, \qquad x_T(x + \delta) = \sqrt{\bar{\alpha}_T}(x + \delta) + \sqrt{1 - \bar{\alpha}_T}\, z.$$

Therefore,

$$x_T(x + \delta) - x_T(x) = \sqrt{\bar{\alpha}_T}\, \delta, \qquad \|x_T(x + \delta) - x_T(x)\|_2 = \sqrt{\bar{\alpha}_T}\, \|\delta\|_2.$$

For an ℓ_p threat model, use the norm comparison $\|\delta\|_2 \le C_{p,d}\|\delta\|_p$ (with the standard constant $C_{p,d}$) to obtain

$$\|x_T(x + \delta) - x_T(x)\|_2 \le \sqrt{\bar{\alpha}_T}\, C_{p,d}\, \varepsilon, \tag{81}$$

for all $\|\delta\|_p \le \varepsilon$. Plugging (81) into the bound of Theorem 5 yields an explicit sufficient bound on purification sensitivity (pathwise or distributional) in terms of $(\rho, \{\bar{\Delta}_t\})$ and the forward schedule $\bar{\alpha}_T$.

Appendix G. Proof of Proposition 7

This appendix section provides a complete proof of Proposition 7 and clarifies the precise role of (i) purification stability, (ii) Lipschitzness of the classifier's score map, and (iii) the margin at the purified point. The result is a standard "margin implies label invariance under bounded score perturbation" argument, with the perturbation controlled by the composition of the purification map and the classifier.

G.1. Notation and formal margin condition

Let the classifier be defined via a score map $s : \mathbb{R}^d \to \mathbb{R}^K$ (logits or real-valued class scores), and prediction

$$c(z) \in \arg \max_{k \in \{1,\dots,K\}} s_k(z).$$

Assume the score map is L_c-Lipschitz in ℓ_2:

$$\|s(z) - s(z')\|_2 \;\le\; L_c \, \|z - z'\|_2, \qquad \forall z, z' \in \mathbb{R}^d. \tag{82}$$

Fix a randomness realization ω and define the purified point

$$\hat{z} \;\triangleq\; \mathrm{Pur}(x;\omega).$$

Let $\hat{y} \triangleq c(\hat{z})$ be the predicted label at $\hat{z}$. Assume the score margin at $\hat{z}$ is at least $\gamma(x) > 0$, i.e.,

$$s_{\hat{y}}(\hat{z}) \;-\; \max_{j \neq \hat{y}} s_j(\hat{z}) \;\ge\; \gamma(x). \tag{83}$$

G.2. Statement

Proposition 13 (Robust classification via stability and margin) *Let $c : \mathbb{R}^d \to \{1,\dots,K\}$ be a classifier with score map s satisfying (82). Suppose the margin condition (83) holds at $\hat{z} = \mathrm{Pur}(x;\omega)$ with margin $\gamma(x) > 0$. If, for the same fixed ω,*

$$\sup_{\|\delta\|_p \le \varepsilon} \|\,\mathrm{Pur}(x+\delta;\omega) - \mathrm{Pur}(x;\omega)\|_2 \;\le\; \frac{\gamma(x)}{2L_c}, \tag{84}$$

then

$$c(\mathrm{Pur}(x+\delta;\omega)) = c(\mathrm{Pur}(x;\omega)) \qquad \forall \delta \textit{ with } \|\delta\|_p \le \varepsilon. \tag{85}$$

G.3. Complete proof

Proof Fix any perturbation δ with $\|\delta\|_p \le \varepsilon$ and define

$$\hat{z}' \triangleq \mathrm{Pur}(x+\delta;\omega).$$

By (84), we have

$$\|\hat{z}' - \hat{z}\|_2 \;\le\; \frac{\gamma(x)}{2L_c}. \tag{86}$$

Step 1: control score perturbations. By Lipschitzness (82),

$$\|s(\hat{z}') - s(\hat{z})\|_2 \;\le\; L_c\,\|\hat{z}' - \hat{z}\|_2 \;\le\; \frac{\gamma(x)}{2}. \tag{87}$$

In particular, for each coordinate $k \in \{1, \dots, K\}$,

$$|s_k(\hat{z}') - s_k(\hat{z})| \;\le\; \|s(\hat{z}') - s(\hat{z})\|_2 \;\le\; \frac{\gamma(x)}{2}. \tag{88}$$

Step 2: show the top-1 score remains top-1. Let $\hat{y} = c(\hat{z})$ be an argmax at $\hat{z}$. For any competitor class $j \neq \hat{y}$, we bound the score gap at $\hat{z}'$:

$$\begin{aligned} s_{\hat{y}}(\hat{z}') - s_j(\hat{z}') &= \big(s_{\hat{y}}(\hat{z}') - s_{\hat{y}}(\hat{z})\big) + \big(s_{\hat{y}}(\hat{z}) - s_j(\hat{z})\big) + \big(s_j(\hat{z}) - s_j(\hat{z}')\big) \\ &\ge -\frac{\gamma(x)}{2} + \gamma(x) - \frac{\gamma(x)}{2} \qquad \text{(by (88) and margin (83))} \\ &= 0. \end{aligned} \tag{89}$$

Thus $s_{\hat{y}}(\hat{z}') \ge s_j(\hat{z}')$ for all $j \neq \hat{y}$, meaning $\hat{y}$ is still an argmax at $\hat{z}'$. Therefore $c(\hat{z}') = c(\hat{z})$, which is exactly (85). Since δ was arbitrary subject to $\|\delta\|_p \le \varepsilon$, the result holds uniformly over the threat set. ■

G.4. Intuition and how to plug in the ECRD stability bound

Why the factor 2 in $\gamma/(2L_c)$? The margin γ is the amount by which the top class beats the runner-up *at the reference point* $\hat{z}$. When moving to $\hat{z}'$, the top score can decrease and a competitor can increase. Bounding each by at most $\gamma/2$ ensures the gap cannot flip sign. This is the usual "two-sided" margin argument.

How to instantiate the condition using the diffusion bounds. The proposition reduces robustness to bounding the purification sensitivity in ℓ_2:

$$\sup_{\|\delta\|_p \le \varepsilon} \|\operatorname{Pur}(x+\delta;\omega) - \operatorname{Pur}(x;\omega)\|_2.$$

In the framework, this is controlled by:

- **Forward coupling.** Using shared forward noise, $\|x_T(x+\delta) - x_T(x)\|_2 \le \sqrt{\bar{\alpha}_T}\, C_{p,d}\,\varepsilon$ (Equation (20)).
- **Reverse stability.** On the contraction-certificate event $\mathcal{E}$, Theorem 5 gives

$$\|\operatorname{Pur}(x+\delta;\omega) - \operatorname{Pur}(x;\omega)\|_2 \le \rho^T \|x_T(x+\delta) - x_T(x)\|_2 + \sum_{k=1}^{T} \rho^{k-1}\bar{\Delta}_k.$$

Combining these yields the explicit sufficient condition

$$\rho^T\big(\sqrt{\bar{\alpha}_T}\, C_{p,d}\,\varepsilon\big) \;+\; \sum_{k=1}^{T}\rho^{k-1}\bar{\Delta}_k \;\le\; \frac{\gamma(x)}{2L_c}, \tag{90}$$

which is directly interpretable as a *contraction–error–margin* tradeoff.

Stability controls how far purification can move an input within the threat set; margin and Lipschitzness control how much movement is needed to flip the predicted label. Equation (90) makes this explicit and is the bridge from sampler stability to robust classification.

Appendix H. Complete Proof and Intuition for Lemma 8: Discrepancy Control Implies a Guidance-Error Budget

This section provides a detailed proof of Lemma 8, together with the main intuition. The high-level message is simple: classifier guidance is the product of (i) the logit Jacobian $Jh_\phi(x)$ and (ii) a probability-space residual $(e_y - p_\phi(x))$. Thus, guidance can change either because the predictive distribution $p_\phi(x)$ changes, or because the Jacobian changes. Discrepancy regularization directly controls the former, and smoothness of the Jacobian controls the latter.

H.1. Recalling the guidance expression

Let $h_\phi : \mathbb{R}^d \to \mathbb{R}^K$ denote the logit map and $p_\phi(x) = \mathrm{softmax}(h_\phi(x))$ the associated probability vector. For a fixed class $y \in \{1, \dots, K\}$, classifier guidance can be written as

$$\mathrm{Guide}_\phi(x,t) \triangleq \nabla_x \log p_\phi(y \mid x) = Jh_\phi(x)^\top \big(e_y - p_\phi(x)\big), \tag{91}$$

where $Jh_\phi(x) \in \mathbb{R}^{K \times d}$ is the Jacobian of logits, and $e_y \in \mathbb{R}^K$ is the y-th standard basis vector. (We keep the notation $\mathrm{Guide}_\phi(\cdot, t)$ to match the diffusion pipeline; the bound below is pointwise in t and does not depend on it explicitly.)

H.2. Lemma statement

Lemma 14 (Discrepancy control to guidance-error budget) *Fix $y \in \{1, \dots, K\}$. Assume that on a region $\mathcal{X} \subset \mathbb{R}^d$ the logit Jacobian satisfies, for some constants $B, L_J \geq 0$,*

$$\|Jh_\phi(x)\|_{\mathrm{op}} \leq B, \qquad \|Jh_\phi(x) - Jh_\phi(x')\|_{\mathrm{op}} \leq L_J\,\|x - x'\|_2, \qquad \forall x, x' \in \mathcal{X}. \tag{92}$$

Then, for all $x, x' \in \mathcal{X}$,

$$\|\,\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t)\|_2 \;\leq\; B\|p_\phi(x) - p_\phi(x')\|_2 + 2L_J\,\|x - x'\|_2. \tag{93}$$

Moreover, if the discrepancy in (16) is KL *and $D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x')) \leq \kappa$, then Pinsker's inequality implies $\|p_\phi(x) - p_\phi(x')\|_1 \leq \sqrt{2\kappa}$, hence*

$$\|\,\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t)\|_2 \;\leq\; B\sqrt{2\kappa} + 2L_J\|x - x'\|_2. \tag{94}$$

Finally, if $\|x - x'\|_p \leq \varepsilon$, then $\|x - x'\|_2 \leq C_{p,d}\varepsilon$, yielding

$$\|\,\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t)\|_2 \;\leq\; B\sqrt{2\kappa} + 2L_J\,C_{p,d}\,\varepsilon. \tag{95}$$

H.3. Complete proof

Proof Fix $x, x' \in \mathcal{X}$. Start from (91) and write the difference:

$$\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t) = Jh_\phi(x)^\top\big(e_y - p_\phi(x)\big) - Jh_\phi(x')^\top\big(e_y - p_\phi(x')\big). \tag{96}$$

Add and subtract the intermediate term $Jh_\phi(x)^\top\big(e_y - p_\phi(x')\big)$:

$$\begin{aligned}\mathrm{Guide}_\phi(x,t) - \mathrm{Guide}_\phi(x',t) &= Jh_\phi(x)^\top\big(e_y - p_\phi(x)\big) - Jh_\phi(x)^\top\big(e_y - p_\phi(x')\big)\\ &\quad + Jh_\phi(x)^\top\big(e_y - p_\phi(x')\big) - Jh_\phi(x')^\top\big(e_y - p_\phi(x')\big)\\ &= Jh_\phi(x)^\top\big(p_\phi(x') - p_\phi(x)\big) + \big(Jh_\phi(x) - Jh_\phi(x')\big)^\top\big(e_y - p_\phi(x')\big).\end{aligned} \tag{97}$$

We bound the two terms in (97) separately.

Term 1: probability mismatch. Using the operator norm bound and $\|A^\top u\|_2 \le \|A\|_{\mathrm{op}}\|u\|_2$,

$$\big\|Jh_\phi(x)^\top\big(p_\phi(x')-p_\phi(x)\big)\big\|_2 \;\le\; \|Jh_\phi(x)\|_{\mathrm{op}}\,\|p_\phi(x)-p_\phi(x')\|_2 \;\le\; B\,\|p_\phi(x)-p_\phi(x')\|_2. \quad (98)$$

Term 2: Jacobian variation. Again by $\|A^\top u\|_2 \le \|A\|_{\mathrm{op}}\|u\|_2$,

$$\big\|\big(Jh_\phi(x)-Jh_\phi(x')\big)^\top\big(e_y-p_\phi(x')\big)\big\|_2 \;\le\; \|Jh_\phi(x)-Jh_\phi(x')\|_{\mathrm{op}}\,\|e_y-p_\phi(x')\|_2. \quad (99)$$

We now bound $\|e_y-p\|_2$ uniformly over all probability vectors $p\in\Delta^{K-1}$. First note that $\|u\|_2 \le \|u\|_1$ for any vector u. Also,

$$\|e_y-p\|_1 = |1-p_y|+\sum_{k\neq y}|0-p_k| = (1-p_y)+\sum_{k\neq y}p_k = (1-p_y)+(1-p_y) = 2(1-p_y) \le 2. \quad (100)$$

Therefore $\|e_y-p\|_2 \le \|e_y-p\|_1 \le 2$, and in particular $\|e_y-p_\phi(x')\|_2 \le 2$. Combining this with the Lipschitzness of the Jacobian in (92),

$$\big\|\big(Jh_\phi(x)-Jh_\phi(x')\big)^\top\big(e_y-p_\phi(x')\big)\big\|_2 \;\le\; L_J\|x-x'\|_2\cdot 2 = 2L_J\,\|x-x'\|_2. \quad (101)$$

Combine the two bounds. Using the triangle inequality on (97) together with (98) and (101) yields

$$\|\,\mathrm{Guide}_\phi(x,t)-\mathrm{Guide}_\phi(x',t)\|_2 \le B\|p_\phi(x)-p_\phi(x')\|_2 + 2L_J\,\|x-x'\|_2,$$

which proves (93).

KL discrepancy ⇒ probability mismatch bound (Pinsker). Assume $D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x'))\le\kappa$. Pinsker's inequality states

$$\|p_\phi(x)-p_\phi(x')\|_1 \le \sqrt{2D_{\mathrm{KL}}(p_\phi(x)\|p_\phi(x'))} \le \sqrt{2\kappa}. \quad (102)$$

Since $\|u\|_2\le\|u\|_1$, we obtain $\|p_\phi(x)-p_\phi(x')\|_2\le\sqrt{2\kappa}$. Substituting into (93) gives (94).

ℓ_p threat model. If $\|x-x'\|_p\le\varepsilon$, then by norm equivalence $\|x-x'\|_2 \le C_{p,d}\|x-x'\|_p \le C_{p,d}\varepsilon$. Plugging this into (94) yields (95). ■

H.4. Intuition and practical interpretation

Two sources of guidance instability. Equation (97) cleanly separates two mechanisms:

- **Probability mismatch (controlled by discrepancy).** Even if the Jacobian $Jh_\phi(x)$ is stable, guidance changes when $p_\phi(x)$ changes, because the residual $(e_y-p_\phi(x))$ changes. This is exactly what discrepancy regularization targets: it makes $p_\phi(x)$ locally invariant under adversarial perturbations, which reduces the first term.

- **Jacobian drift (controlled by smoothness).** Even if $p_\phi(x)$ is stable, guidance can change if $Jh_\phi(x)$ changes rapidly with x. The constant L_J captures this second-order sensitivity (how quickly the input-gradient of logits changes), and the factor 2 is an unavoidable bound coming from $\|e_y-p\|_2\le 2$.

Why the bound matches the ECRD "error budget" abstraction. ECRD needs a stepwise guidance-error budget $\Delta_t^{(g)}$ that upper bounds how much the guidance component can change when the input is perturbed within the threat set (or within a trajectory-relevant region). Lemma 14 provides exactly such a translation:

- a measurable **discrepancy certificate** (e.g., a validated upper bound κ on KL under ℓ_p perturbations), and
- validated **Jacobian bounds** (B, L_J) on the region $\mathcal{X}$ visited by the reverse chain,

together imply the explicit guidance-error budget

$$\Delta^{(g)} \lesssim B\sqrt{2\kappa} + 2L_J\, C_{p,d}\varepsilon,$$

which can be fed into the weights (11) and into the post-calibration error term $\bar{\Delta}_t$.

If validation diagnostics certify $\max_{\|\delta\|_p \le \varepsilon} D_{\mathrm{KL}}(p_\phi(x) \| p_\phi(x+\delta)) \le \kappa_t$ for reverse states in $\mathcal{X}_t$, and if step-dependent bounds $(B_t, L_{J,t})$ hold on $\mathcal{X}_t$, then Lemma 14 implies the explicit guidance budget

$$\Delta_t^{(g)} = B_t\sqrt{2\kappa_t} + 2L_{J,t}C_{p,d}\varepsilon,$$

which directly informs the attenuation weights (11) and the post-calibration budgets used in the stability guarantees.