TARO: TEMPORAL ADVERSARIAL RECTIFICATION OPTIMIZATION USING DIFFUSION MODELS AS PURIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial purification with diffusion models seeks to project adversarial examples back toward the data manifold, but balancing semantic preservation and robustness against adaptive attacks remains challenging. Recent work shows that standard diffusion purification can fail under adaptive evaluation, while test-time score-based optimization is more resilient. Existing optimization defenses, however, typically rely on a single diffusion noise regime or treat timesteps uniformly, overlooking the distinct roles of coarse and fine denoising scales. We propose Temporal Adversarial Rectification Optimization (TARO), an inference-time purification method that builds a temporally guided score prior from multiple denoising views along the diffusion trajectory. TARO forms a coarse-to-fine residual target: high-noise experts provide globally smoothed structure with reduced adversarial sensitivity, while low-noise experts restore image-specific, class-relevant details. A guidance strength controls this temporal correction, allowing TARO to balance robust global rectification with semantic preservation. Empirically, TARO improves robust accuracy across datasets and adaptive threat models in a zero-shot setting, while remaining compatible with complementary adversarial-likelihood objectives for further robustness gains.

1 INTRODUCTION

Deep neural networks remain highly susceptible to adversarial examples, where small, often imperceptible perturbations induce erroneous predictions under norm-bounded input changes (Szegedy et al., 2013; Goodfellow et al., 2015). Adversarial training is the dominant empirical defense, but its robustness is typically tied to the perturbation family used during training, it substantially increases training cost, and it often induces a clean-accuracy tradeoff (Madry et al., 2018; Schlarmann et al., 2024). These limitations have motivated adversarial purification, an inference-time defense that preprocesses a potentially adversarial input before classification by mapping it back toward the natural-image manifold.

Diffusion models provide a natural prior for adversarial purification because their denoising dynamics encode image structure across a continuum of noise scales. Standard diffusion purifiers inject Gaussian noise into the input to attenuate adversarial structure and then reconstruct the image through a pretrained reverse process (Song et al., 2021; Ho et al., 2020; Nie et al., 2022). This strategy is attractive because the purifier can be used with a fixed downstream classifier. However, the diffusion time axis introduces an intrinsic robustness–fidelity tension. Low-noise denoising preserves local image content but may retain adversarial artifacts, while high-noise denoising suppresses stronger perturbations but may discard class-relevant details or drift toward an incorrect semantic mode. Prior guided purification methods attempt to steer this reverse process using auxiliary classifiers, distance metrics, or other guidance objectives (Wang et al., 2022). Yet recent adaptive evaluations have shown that robustness claims for stochastic purification can be substantially overstated when attacks fail to differentiate through the purification process or fail to account for stochasticity (Lee & Kim, 2023; Kassis et al., 2025).

A recent alternative is to cast purification as test-time optimization. Rather than relying only on a fixed reverse diffusion trajectory, these methods directly update the input by optimizing an objective induced by a diffusion prior (Zhang et al., 2023; Chen et al., 2024). Adversary-aware optimization further strengthens this paradigm by augmenting the clean-image prior with an explicit likelihood

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

model for adversarial residuals (Wesego & Rooshenas, 2025). Despite their empirical strength, existing optimization-based purifiers still leave two structural limitations. First, they typically use a single diffusion noise level at each update, collapsing the multi-scale information encoded by the diffusion model into a single local estimate. Second, adversary-aware variants require an auxiliary perturbation model trained from generated adversarial examples, making the defense dependent on the attack distribution used during training.

We propose *Temporal Adversarial Rectification Optimization* (TARO), a test-time purification framework that treats diffusion timesteps as scale-dependent denoising experts. Different noise regimes provide complementary information: higher-noise denoisers recover globally smoothed structure with reduced sensitivity to fragile adversarial components, while lower-noise denoisers preserve image-specific, class-relevant details. Motivated by a temporal product-of-experts view of diffusion scores, TARO constructs an affine coarse-to-fine target from multiple denoising views, allowing the purifier to combine robust global structure with fine semantic detail instead of committing to a single point on the robustness–fidelity spectrum.

This temporal construction gives TARO a zero-shot rectification signal from the pretrained clean-image diffusion model itself. Empirically, this zero-shot setting is particularly effective under ℓ_∞ adaptive attacks, where TARO improves over prior diffusion purification and test-time optimization baselines while preserving high clean accuracy. At the same time, our results show that ℓ_2 perturbations are more challenging for raw zero-shot temporal rectification: smoother or more spatially distributed distortions can benefit from additional residual information or stronger stabilization. TARO is therefore complementary to adversary-aware likelihood models such as AAOpt (Wesego & Rooshenas, 2025): when an adversarial residual model is available, replacing the single-scale diffusion prior with TARO's temporal prior further improves robustness.

Reliable evaluation is equally important for this class of defenses. Optimization-based purifiers contain inner optimization loops, and attacks that remove these loops, omit higher-order gradient tracking, or fail to account for stochasticity may evaluate an easier surrogate defense rather than the actual purifier. We therefore use an optimization-aware evaluation protocol that attacks the full purifier–classifier composition, accounts for stochasticity through expectation over transformation, and preserves gradient flow through the inner loop when the optimization is evaluated. This protocol also reveals that improperly differentiated optimization-based purification can produce severely inflated robustness estimates, motivating our use of strong and properly applied adaptive attacks throughout the paper.

Our contributions are as follows:

- We introduce TARO, a test-time purifier that aggregates multiple diffusion noise regimes into an affine coarse-to-fine denoising target, combining robust high-noise structure with fine-scale semantic preservation.
- We interpret TARO as an affine extension of temporal product-of-experts aggregation, with a correction strength controlling the fine-to-coarse balance, and provide risk and local Gaussian interpretations.
- We evaluate TARO under PGD-EOT, BPDA-EOT, AutoAttack-EOT, and DiffBreak. TARO improves zero-shot robustness and further improves robustness with adversary-aware likelihoods or consistency regularization.
- We provide an optimization-aware evaluation study showing that reliable adaptive evaluation of optimization-based purifiers requires preserving higher-order gradients and is critical for robust evaluation.

2 Related Work

Adversarial purification is an inference-time defense motivated by the cost and threat-model specificity of adversarial training. It preprocesses a potentially corrupted input before classification by mapping it back toward the clean data manifold. Diffusion purification methods, such as DiffPure (Nie et al., 2022), inject noise and solve a reverse denoising process to remove adversarial structure. While effective against some attacks, standard reverse-sampling purification faces a robustness–fidelity

tradeoff: low-noise denoising may preserve adversarial artifacts, whereas high-noise denoising can distort class-relevant semantics.

Test-time optimization methods address this issue by directly updating the input using a pretrained diffusion prior. Score-Opt (Zhang et al., 2023) and LM (Chen et al., 2024) optimize score-induced objectives that move adversarially perturbed input $\mathbf{x}_{\mathrm{adv}}$ toward high-density regions of the clean data distribution. AAOpt (Wesego & Rooshenas, 2025) strengthens this MAP view by adding a learned adversarial likelihood $\log p_\phi(\mathbf{x}_{\mathrm{adv}} \mid \mathbf{x})$, allowing the purifier to model structured residual artifacts. However, these methods typically use a single score or noise scale at each update, and adversary-aware variants require a dedicated perturbation model trained on generated attacks. TARO instead uses multiple diffusion noise regimes at test time and can be used either as a zero-shot temporal prior or combined with adversary-aware likelihoods.

Robust evaluation of purification defenses is challenging because stochastic and optimization-based purifiers can lead to misleading gradients. Prior work has shown that diffusion purification robustness can be overestimated when attacks are weak, non-adaptive, or fail to account for stochasticity and differentiability through the defense (Lee & Kim, 2023; Kang et al., 2023; Kassis et al., 2025). DiffBreak (Kassis et al., 2025) further showed that low-frequency, perceptually guided perturbations can collapse standard diffusion purification. We follow adaptive evaluation best practices, including EOT, attacks on the full purifier–classifier composition, and graph-consistent evaluation for optimization-based purifiers whose inner loops must remain exposed to the attacker.

Diffusion guidance provides a related algebraic perspective. Classifier guidance modifies the score using an auxiliary classifier (Dhariwal & Nichol, 2021), while classifier-free guidance combines conditional and unconditional denoising estimates (Ho & Salimans, 2022). Self-Guidance (Epstein et al., 2023) and Auto-Guidance (Karras et al., 2024) derive guidance from internal representations or contrasts between denoisers. These methods primarily target controllability or sample quality in generation. In contrast, TARO uses multi-scale denoising estimates from different diffusion noise regimes as temporal experts for adversarial posterior rectification.

3 METHOD

Let $\mathbf{x}_{\mathrm{adv}} \in \mathbb{R}^d$ be an adversarially perturbed input and let $\mathbf{x}$ denote the unknown clean image we aim to recover. We pose purification as explicit posterior inference:

$$\arg\max_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{x}_{\mathrm{adv}}) = \arg\min_{\mathbf{x}} \underbrace{\mathcal{L}_{\mathrm{fid}}(\mathbf{x}, \mathbf{x}_{\mathrm{adv}})}_{\text{fidelity / corruption model}} + \lambda \underbrace{\mathcal{R}(\mathbf{x})}_{\text{image prior}}. \tag{1}$$

In the simplest setting, we use a Gaussian corruption model, $\mathcal{L}_{\mathrm{fid}}(\mathbf{x}, \mathbf{x}_{\mathrm{adv}}) = \frac{1}{2\sigma_{\mathrm{fid}}^2}\|\mathbf{x} - \mathbf{x}_{\mathrm{adv}}\|_2^2$. More generally, $\mathcal{L}_{\mathrm{fid}}$ can be any differentiable negative log-likelihood, including data-driven corruption models that learn $p_\phi(\mathbf{x}_{\mathrm{adv}} \mid \mathbf{x})$ (Wesego & Rooshenas, 2025). In that case, we set $\mathcal{L}_{\mathrm{fid}}(\mathbf{x}, \mathbf{x}_{\mathrm{adv}}) = -\log p_\phi(\mathbf{x}_{\mathrm{adv}} \mid \mathbf{x})$.

The MAP objective in Eq. equation 1 follows the standard inverse-problem view of purification. RED replaces an explicit prior with a denoiser-induced regularizer $\mathcal{R}_{\mathrm{RED}}(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - D(\mathbf{x})\|_2^2$. For a diffusion denoiser, Tweedie's formula gives $D_\theta(\mathbf{y}, \sigma_t) = \mathbf{y} + \sigma_t^2 \nabla_{\mathbf{y}} \log p_{\sigma_t}(\mathbf{y})$, linking denoising to score ascent. We use this connection to build a temporal product-of-experts prior over multiple noise levels.

3.1 TEMPORAL POE AND AFFINE TEMPORAL AGGREGATION

Let $\{p_t(\mathbf{x})\}_{t \in \mathcal{T}}$ denote the diffusion-induced smoothed data distributions indexed by timesteps $\mathcal{T} = [t_{\min}, t_{\max}]$. We define a continuous-time temporal product-of-experts prior:

$$\log p_{\mathrm{TARO}}(\mathbf{x}) = \int_{\mathcal{T}} w(t) \log p_t(\mathbf{x})\, dt - \log Z. \tag{2}$$

The function $w(t)$ controls how much each noise regime contributes to the effective prior. For example, placing more mass on small t emphasizes fine-scale fidelity, while placing more mass on large t emphasizes stronger global correction and robustness.

By Tweedie's formula, the score of the continuous-time PoE prior can be written as $\nabla_{\mathbf{x}} \log p_{\mathrm{TARO}}(\mathbf{x}) = \int_{\mathcal{T}} w(t) \frac{D_\theta(\mathbf{x},\sigma_t)-\mathbf{x}}{\sigma_t^2}\, dt$. Defining the variance-corrected weight $\tilde{w}(t) = w(t)/\sigma_t^2$, this becomes $\nabla_{\mathbf{x}} \log p_{\mathrm{TARO}}(\mathbf{x}) = \int_{\mathcal{T}} \tilde{w}(t)\, (D_\theta(\mathbf{x},\sigma_t) - \mathbf{x})\, dt$. We approximate the continuous temporal prior using a sparse K-expert quadrature: $\tilde{w}(t)\, dt \approx \sum_{k=1}^K \hat{\gamma}_k \delta(t - t_k)$. This gives $\nabla_{\mathbf{x}} \log p_{\mathrm{TARO}}(\mathbf{x}) \approx \sum_{k=1}^K \hat{\gamma}_k\, (D_\theta(\mathbf{x},\sigma_{t_k}) - \mathbf{x})$. Letting $\Lambda = \sum_{k=1}^K \hat{\gamma}_k$ and $\alpha_k = \frac{\hat{\gamma}_k}{\Lambda}$, we obtain

$$\nabla_{\mathbf{x}} \log p_{\mathrm{TARO}}(\mathbf{x}) \approx \Lambda \left(\sum_{k=1}^K \alpha_k D_\theta(\mathbf{x},\sigma_{t_k}) - \mathbf{x} \right) = \Lambda(\bar{D}_\alpha(\mathbf{x}) - \mathbf{x}), \tag{3}$$

where we define the temporal consensus denoiser $\bar{D}_\alpha(\mathbf{x}) \triangleq \sum_{k=1}^K \alpha_k D_\theta(\mathbf{x},\sigma_{t_k})$.

Standard PoE requires nonnegative expert weights, which yields a convex interpolation of the temporal denoisers. TARO relaxes this convexity constraint by allowing the effective aggregation weights to be negative while preserving $\sum_k \alpha_k = 1$. When some $\alpha_k < 0$, the temporal target becomes an affine extrapolation of the experts, enabling coarse-to-fine corrections beyond standard convex aggregation while retaining the same denoiser-aggregation form, $\bar{D}_\alpha(\mathbf{x})$. A local Gaussian interpretation of this affine temporal aggregation is provided in Appendix A.1.

Choice of temporal experts. The discrete experts $\{t_k\}_{k=1}^K$ are interpreted as a sparse quadrature of the continuous temporal prior. The theory only requires ordered timesteps $t_{\min} \le t_1 < \cdots < t_K \le t_{\max}$, which may be chosen by any monotone schedule $t_k = \psi(k; t_{\min}, t_{\max}, K)$, such as linear, logarithmic, or problem-specific spacing. In our sampled-time implementation, we generate effective timesteps by $t_k = \mathrm{clip}(\sqrt{t\,\kappa_k}, t_{\min}, t_{\max})$, with $\kappa_1 < \cdots < \kappa_K$. We evaluate both $K = 2$ and $K = 3$: $K = 2$ uses fine and coarse experts, while $K = 3$ adds a middle expert to stabilize the coarse-to-fine transition.

3.2 Optimization Algorithm

Algorithm 1 gives the score-prior-only TARO update. At each iteration, we sample a base timestep, construct K effective timesteps, and evaluate the denoiser at each scale. TARO anchors the target at the coarsest expert u_K and adds lower-noise residual corrections, $\mathbf{x}_{\mathrm{target}} = u_K + \sum_{j=1}^{K-1} \gamma^{K-j}(u_j - u_{j+1})$, where $u_j - u_{j+1}$ transfers information from a finer expert to the next coarser one. The image is updated toward this target with the RED-style fixed-target loss: $\ell = \|\mathrm{sg}(\mathbf{x}_{\mathrm{target}}) - \mathbf{x}_i\|_2^2$. Here, $\mathrm{sg}(\cdot)$ treats $\mathbf{x}_{\mathrm{target}}$ as fixed only during the purifier update and makes the optimization efficient similar to Wesego & Rooshenas (2025). But to show that are our improvements are not coming from the effects this, we expose the gradients to the attacker by removing the $\mathrm{sg}(\cdot)$ during attack. Fidelity or adversary-aware likelihood terms can also be added to recover the full MAP objective as discussed in section 3.5.

3.3 Guidance Interpretation of Affine Temporal Extrapolation

The affine temporal aggregation used by TARO admits a guidance-style interpretation (Ho & Salimans, 2022; Karras et al., 2024). Let $t_f < t_c$, where t_f denotes a lower-noise fine scale and t_c denotes a higher-noise coarse scale. We write $u_f = D_\theta(x_{t_f}, \sigma_{t_f})$ and $u_c = D_\theta(x_{t_c}, \sigma_{t_c})$. The coarse expert u_c provides a stronger global projection and tends to suppress adversarial artifacts, but may introduce semantic drift or remove fine details. The fine expert u_f is more input-grounded and better preserves details, but may retain part of the adversarial structure. TARO uses the temporal discrepancy $u_f - u_c$ as a correction direction.

For $K = 2$, the TARO target is $u_\gamma = u_c + \gamma(u_f - u_c) = \gamma u_f + (1-\gamma)u_c$. When $0 \le \gamma \le 1$, this is a convex interpolation between fine and coarse experts. When $\gamma > 1$, it becomes a guidance-style extrapolation that amplifies the fine-minus-coarse correction. This has the same algebraic form as guidance methods such as CFG and Auto-Guidance, which also scale a discrepancy between two predictions. However, the source and purpose of the discrepancy are different: in TARO, both predictions are produced by the same diffusion denoiser at different temporal noise regimes, and the extrapolated target is used for test-time MAP adversarial purification rather than conditional sample generation. Thus, TARO should be viewed as a temporal affine extrapolation with a guidance-

Algorithm 1 Test-time Optimization Algorithm with TARO

Input: Input image $\mathbf{x}_{\text{adv}}$; pretrained score denoiser D_θ; guidance strength γ; iterations M; timestep range $[t_{\min}, t_{\max}]$; timestep sampling schedule; noise-level schedule $\{\sigma_t\}_{t=1}^{T}$; ordered scale multipliers $[\kappa_1, \ldots, \kappa_K]$

$\mathbf{x}_0 \leftarrow \mathbf{x}_{\text{adv}}$
for $i = 0$ **to** $M - 1$ **do**
  Sample base timestep t from the predefined timestep schedule
  for $j = 1$ **to** K **do**
    $t_j \leftarrow \text{clip}(\sqrt{t\,\kappa_j}, t_{\min}, t_{\max})$ ▷ Set effective timestep
    Sample $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{u}_j \leftarrow D_\theta\big(\mathbf{x}_i + \sigma_{t_j}\boldsymbol{\epsilon}_j, \sigma_{t_j}\big)$ ▷ Compute temporal expert
  end for
  $\mathbf{x}_{\text{target}} \leftarrow \mathbf{u}_K + \sum_{j=1}^{K-1} \gamma^{K-j}(\mathbf{u}_j - \mathbf{u}_{j+1})$ ▷ Form TARO target
  $\ell \leftarrow \|\text{sg}(\mathbf{x}_{\text{target}}) - \mathbf{x}_i\|_2^2$ ▷ Calculate score-prior loss
  $\mathbf{x}_{i+1} \leftarrow \text{OptimizerUpdate}(\mathbf{x}_i, \ell)$ ▷ Update image
end for
Return: Final image $\mathbf{x}_M$

style form: the resemblance to guidance is algebraic, while the correction direction is defined by fine-to-coarse denoising disagreement for adversarial rectification.

3.4 Risk Interpretation of Temporal Consensus

The PoE derivation motivates aggregating temporal denoising experts. Our implemented TARO target uses a residual coarse-to-fine correction controlled by γ. We therefore analyze the risk of the guided fine-to-coarse target directly, which makes the role of γ explicit.

Proposition 1 (Risk of the guided fine-to-coarse target): Let $x_{t_f} = x + \sigma_{t_f}\epsilon_f$ and $x_{t_c} = x + \sigma_{t_c}\epsilon_c$. We use $u_f = D_\theta(x_{t_f}, \sigma_{t_f})$ and $u_c = D_\theta(x_{t_c}, \sigma_{t_c})$ to denote the fine and coarse temporal experts, where $t_f < t_c$. Suppose $u_f = x^\star + b_f(x^\star) + \xi_f$ and $u_c = x^\star + b_c(x^\star) + \xi_c$ with $\mathbb{E}[\xi_f \mid x^\star] = \mathbb{E}[\xi_c \mid x^\star] = 0$. Define the TARO guided target $u_\gamma = u_c + \gamma(u_f - u_c)$. Then

$$u_\gamma = x^\star + \underbrace{b_c + \gamma(b_f - b_c)}_{b_\gamma} + \underbrace{\xi_c + \gamma(\xi_f - \xi_c)}_{\xi_\gamma}, \quad \text{and}$$

$$\mathbb{E}[\|u_\gamma - x^\star\|_2^2 \mid x^\star] = \|b_c + \gamma(b_f - b_c)\|_2^2 + \mathbb{E}[\|\xi_c + \gamma(\xi_f - \xi_c)\|_2^2 \mid x^\star].$$

Proof of the Proposition is provided in Appendix A.2.

Proposition 3.4 shows that the correction strength γ changes both the effective bias b_γ and the residual variation ξ_γ of the temporal target. Small γ keeps u_γ close to the coarse expert u_c, which can inherit oversmoothing bias and remove class-relevant details. Larger γ increases the influence of the fine expert u_f through the correction $u_f - u_c$, restoring image-specific structure but potentially amplifying cross-timestep residual variation. This motivates selecting γ on validation data.

3.5 Using Data-Driven Adversarial Models

In addition to temporal correction, purification can also benefit from both attack-specific residual modeling. Adversary-aware methods such as AAOpt (Wesego & Rooshenas, 2025) address the latter by explicitly modeling the residual between the adversarial input and the purified estimate. In particular, AAOpt augments the diffusion prior with a learned perturbation likelihood by minimizing an objective of the form $\mathcal{L}_{\text{AAOpt}}(\mathbf{x}) = -\log p_\theta(\mathbf{x}) - \lambda \log p_{\text{pert}}(\mathbf{x} - \mathbf{x}_{\text{adv}})$. Here, $p_{\text{pert}}(\mathbf{x} - \mathbf{x}_{\text{adv}})$ models the likelihood of the residual between the current purified estimate and the adversarial input.

TARO is complementary: it uses the zero-shot discrepancy $\Delta u = u_f - u_c$ between fine and coarse temporal experts as an internal proxy for scale-sensitive adversarial structure. Thus, AAOpt provides an external learned residual model, while TARO provides an internal temporal rectification signal from the pre-trained diffusion model itself. The two can be combined by adding specific adverarial likelihood term to the TARO temporal consensus prior (see Appendix A.3).

4 EXPERIMENTS

Setup. We evaluate TARO on CIFAR-10, CIFAR-100, CIFAR-10-C, and Imagenette. CIFAR-10-C, Imagenette, and extended hyperparameter details are discussed in Appendix. We report clean accuracy on benign inputs and robust accuracy under adaptive attacks. Unless otherwise stated, TARO results are averaged over independent trials sampled from the test set.

We compare against adversarial training (AT) baselines, including Pang et al. (2022), Gowal et al. (2020), Gowal et al. (2021), and Wang et al. (2023); standard adversarial purification (AP) methods, including Nie et al. (2022), Yoon et al. (2021), Lin et al. (2024), and Li et al. (2025); and test-time optimization baselines, including LM (Chen et al., 2024), Score-Opt (Zhang et al., 2023), and AAOpt (Wesego & Rooshenas, 2025). For direct comparisons, we reimplement AAOpt using the same classifier, diffusion checkpoint, and attack settings as TARO.

We use pretrained EDM denoisers that predict the clean estimate $\hat{x}_0 = D_\theta(x_t; \sigma)$, together with clean-trained WRN classifiers. Thus, robustness improvements for AP methods come from purification rather than adversarially trained classifiers. TARO-K denotes aggregation over K diffusion noise scales, and TARO-K+AA denotes the hybrid variant of TARO by adding adversarial likelihood term from AAOpt.

Evaluation protocol. Because TARO is an optimization-based purifier, attacks must target the full purifier–classifier composition and account for stochasticity with EOT. We use PGD-EOT, BPDA-EOT, AutoAttack-EOT, and low freqruency adaptive attacks from DiffBreak.

Because TARO belongs to the class of test-time optimization defenses, we include a dedicated evaluation study in Appendix A.4. Optimization-based purifiers differ from standard diffusion defenses because the purifier itself contains an inner optimization loop. If an attack omits higher-order gradient tracking, or otherwise breaks the computational graph, it may present a strong result against a weak evaluation. We formalize this failure mode and provide an optimization-aware evaluation protocol that tracks higher-order gradient flow through the inner loop required for a strong attack. We evaluate all our experiments throughout this work with this strong setup and as a case study, Appendix A.4 revisits CLIPure (Zhang et al., 2025) and shows that its previously reported high robust accuracies collapse to 0% under corrected graph-consistent evaluation. The analysis motivates the stronger evaluation suite used throughout this work.

4.1 ADAPTIVE PGD-EOT ATTACKS

We evaluate end-to-end robustness against adaptive Projected Gradient Descent with Expectation over Transformation (PGD-EOT) (Athalye et al., 2018). The attack targets the purifier–classifier composition and estimates gradients through the stochastic purification process. We consider both ℓ_∞ and ℓ_2 perturbations with budgets $\epsilon = 8/255$ and $\epsilon = 0.5$, respectively. Following Wesego & Rooshenas (2025); Zhang et al. (2023); Lee & Kim (2023), we use 20 attack iterations and a surrogate one-step gradient approximation through the iterative purification trajectory to generate the attacks. Attacks using the full iteration are also studied in section 4.3. When generating the attacks, we remove the $\mathrm{sg}(\cdot)$ so that the attacker can also have access to the gradients coming from the diffusion model. Unlike the original AAOpt evaluation, which uses 20 EOT samples for the attack and a larger number of EOT samples for the PGD-EOT defense, we use 20 EOT samples for both attack and defense across TARO and all directly reimplemented baselines including TARO and AAOpt. This gives a matched and computationally comparable evaluation setting.

Table 1 reports CIFAR-10 results under ℓ_∞ and ℓ_2 PGD-EOT using a WRN-28-10 classifier. Under ℓ_∞, all TARO variants achieves the strongest robustness among the evaluated purification methods while maintaining clean accuracy close to the undefended classifier. TARO-3+AA achieves the strongest of all from the combination of TARO with adversary likelihood term. Under ℓ_2, the TARO variants achieve very close results to the strongest one. The table also shows that TARO using different timescales improves the results of TARO-1 to TARO-2 and TARO-3. The strongest result is achieved when going from TARO-1 to TARO-2 suggesting adding more scales above that provide diminishing returns. To study that the improvements of TARO-2 and TARO-3 are not just from additional computate resources, we conduct experiments in Appendix A.7 that the improvements are arising from the multi-scale temporal prior and not as a result of more compute.

Table 1: Standard and robust accuracy on CIFAR-10 using WRN-28-10 under PGD-EOT-20. Robustness is evaluated under ℓ_∞ ($\epsilon = 8/255$) and ℓ_2 ($\epsilon = 0.5$) perturbations. Avg. Robustness averages the available robust accuracies; Avg. All averages the available clean and robust accuracies. Missing results are excluded from averages.

Type	Method	Clean	ℓ_∞ ($\epsilon = 8/255$)	ℓ_2 ($\epsilon = 0.5$)	Avg. Robustness
	Default	94.78	0.0	0.0	0.00
AT	Pang et al. (2022)	88.62	64.95	-	64.95
	Gowal et al. (2020)	88.54	65.10	-	65.10
	Gowal et al. (2021)	87.51	66.01	-	66.01
	Rebuffi et al. (2021)	91.79	-	85.05	85.05
	Augustin et al. (2020)	93.96	-	**86.14**	86.14
AP	Yoon et al. (2021)	85.66	37.27	-	37.27
	Nie et al. (2022)	91.41	51.25	82.11	66.68
	Score-Opt-O	91.21	64.96	79.09	72.03
	Score-Opt-N	94.43	65.62	84.86	75.24
	Lin et al. (2024)	90.42	64.06	85.55	74.81
	AAOpt	92.18	84.76	85.97	85.37
	TARO-1	$91.66_{\pm 0.6}$	$86.28_{\pm 0.8}$	$84.76_{\pm 0.2}$	85.52
	TARO-2	$93.78_{\pm 1.0}$	$87.26_{\pm 1.5}$	$85.46_{\pm 0.8}$	86.36
	TARO-3	$94.02_{\pm 0.9}$	$87.41_{\pm 1.1}$	$85.70_{\pm 1.3}$	86.56
	TARO-1+AA	$92.10_{\pm 0.6}$	$86.55_{\pm 1.0}$	$84.99_{\pm 1.3}$	85.77
	TARO-2+AA	$93.59_{\pm 0.8}$	$87.73_{\pm 1.7}$	$85.26_{\pm 1.4}$	86.50
	TARO-3+AA	$93.90_{\pm 1.0}$	**88.12**$_{\pm 1.2}$	$85.74_{\pm 1.5}$	**86.93**

4.2 BPDA-EOT Attacks

We next evaluate robustness against Backward Pass Differentiable Approximation with EOT (BPDA-EOT) (Athalye et al., 2018). BPDA-EOT tests defenses with non-differentiable, stochastic, or approximate components by replacing the backward pass with a differentiable surrogate while averaging gradients over stochastic transformations. We use 50 BPDA iterations and 20 EOT samples under an ℓ_∞ perturbation budget of $\epsilon = 8/255$. This setting tests whether TARO remains robust when the attack accounts for randomized purification and approximate gradients. If TARO was obfuscating gradients in attacks passing through the purification loop, we will see inferior results here. Tables 2 and 3 summarize the BPDA-EOT results on CIFAR-10 and CIFAR-100. As shown in the tables, TARO with TARO+AA achieve robust results here also and shows that improvements seen in the PGD experiments are not achieved from gradient obfuscation. On CIFAR-10, TARO-2-AA achieves the strongest robust accuracy while maintaining high clean accuracy. On CIFAR-100, where classification is more challenging and the data manifold is more complex, TARO-2 improves over AAOpt and Score-Opt baselines. These results support the claim that temporal multi-scale purification provides a stronger optimization signal than the baselines' single score guidance.

4.3 Fully Exposed Attacks

In addition to the above attacks, we rigorously evaluate our method against fully exposed, adaptive white-box adversaries to rule out any weak evaluation. In this setting, the adversarial attacks are generated by backpropagating through all the purification iterations with no $\mathrm{sg}(\cdot)$ applied to either the attack or the defense mechanisms. By utilizing optimization-aware gradient tracking, the adversary computes exact gradients through the fully unrolled reverse diffusion trajectory, ensuring complete access to the defense's internal computational graph.

4.3.1 Full-Reverse AutoAttack-EOT

We evaluate TARO using AutoAttack-EOT with gradients propagated through all purification iterations, rather than the one-step approximation used in the main PGD-EOT evaluation. This evaluates robustness in a fully unrolled white-box setting. The attack targets the complete purifier–classifier pipeline and averages gradients over stochastic purification samples. To reduce computational cost,

Table 2: BPDA+EOT attack on CIFAR-10 under ℓ_∞ perturbations ($\epsilon = 8/255$), WRN-28-10.

Method	Clean	Robust
Default	94.78	0.0
Wang et al. (2022)	93.50	79.83
Yoon et al. (2021)	86.14	70.01
Nie et al. (2022)	89.02	81.40
Score-Opt-O	90.23	81.36
Score-Opt-N	93.94	90.07
ADBM	91.93	70.51
AAOpt	92.10	91.36
TARO-2	$90.62_{\pm 0.6}$	$88.90_{\pm 0.6}$
TARO-3	$89.40_{\pm 0.4}$	$86.87_{\pm 0.6}$
TARO-2+AA	$92.68_{\pm 0.4}$	$\mathbf{92.30}_{\pm 0.4}$
TARO-3+AA	$92.29_{\pm 0.3}$	$91.71_{\pm 0.6}$

Table 3: BPDA+EOT attack on CIFAR-100 under ℓ_∞ perturbations ($\epsilon = 8/255$), WRN-28-10.

Method	Clean	Robust
Default	81.55	0.0
Yoon et al. (2021)	60.66	39.72
Hill et al. (2021)	51.66	26.10
Score-Opt-O	70.53	66.11
Score-Opt-N	74.18	60.21
AAOpt	69.68	66.20
TARO-2	$73.78_{\pm 1.5}$	$66.71_{\pm 1.9}$
TARO-3	$72.61_{\pm 1.8}$	$\mathbf{68.66}_{\pm 2.2}$

we evaluate 128 samples per trial and report results averaged over three independent trials. Table 4 reports results for the no-stop-gradient setting, in which $\mathrm{sg}(\cdot)$ is removed from both the defense and the attack computation. TARO-2 achieves the highest robust accuracy among the evaluated methods at 76.73%, compared with 74.22% for Score-Opt-O and 57.42% for AAOpt. TARO-2+AA achieves 75.44%. These results show that TARO remains effective under fully unrolled adaptive attacks even when stop-gradient is absent from the purification objective. Table 5 compares the settings with and without stop-gradient. TARO-2 achieves 71.09% robust accuracy with stop-gradient and 76.73% without it, whereas AAOpt achieves 85.54% and 57.42%, respectively. Thus, the relative performance depends on the stop-gradient configuration: AAOpt performs better in the stop-gradient setting, while TARO-2 performs well on average in both settings. This shows that TARO's observed robustness is not confined to the stop-gradient formulation.

Table 4: AutoAttack-EOT with full unrolled optimization on CIFAR-10, WRN-28-10.

Model	Robust Acc.
Default	0.0
DiffPure	39.45
Score-Opt-O	74.22
AAOpt	57.42
TARO-2	$76.73_{\pm 3.2}$
TARO-2+AA	$75.44_{\pm 1.6}$

Table 5: AutoAttack-EOT Robust Accuracy with full unrolled attack on CIFAR-10 with and without $\mathrm{sg}(\cdot)$ on the diffusion model

Model	No sg	With sg	Avg.
AAOpt	57.42	85.54	71.48
TARO-2	76.73	71.09	**73.91**

4.3.2 DIFFBREAK LOW-FREQUENCY ATTACKS

Recently, Kassis et al. (2025) introduced DiffBreak, a fully-exposed adaptive attack designed specifically to bypass diffusion-based purification defenses. Unlike standard ℓ_p-bounded attacks, DiffBreak uses a low-frequency (LF) perturbation strategy guided by perceptual similarity objectives such as LPIPS. These perturbations remain closer to natural image statistics, making them difficult for standard diffusion denoisers to remove. As shown in Table 6, the evaluated standard diffusion purification baselines collapse under this attack: both one-step denoising and DiffPure-EDM achieve 0.0% robust accuracy. In contrast, TARO-2 and TARO-2-AA remain substantially more robust, with TARO-2 achieving the best result. We also evaluate a transfer-DiffBreak setting. We generate low-frequency adversarial examples against DiffPure-EDM using the same diffusion model and evaluate the resulting examples on TARO. Table 7 shows that TARO-2 and TARO-2+AA remain robust under transfer, again outperforming standard diffusion purification which completely collapse.

Table 6: Robust accuracy on CIFAR-10 under the DiffBreak low-frequency (LF) attack with WRN-28-10 and EOT=20.

Model	Robust Accuracy
1-step Denoise	0.0$\pm$0.0
DiffPure-EDM	0.0$\pm$0.0
GDMP	0.0$\pm$0.0
Score-Opt-O	80.20$\pm$1.9
TARO-2	**80.72**$\pm$1.9
TARO-2+AA	77.08 $\pm$4.8

Table 7: Robust accuracy on CIFAR-10 under transfer-DiffBreak low-frequency (LF) attack with similar setup as Table 6

Model	Robust Accuracy
DiffPure-EDM	0.0$\pm$0.0
TARO-2	**93.74**$\pm$1.2
TARO-2+AA	93.22$\pm$1.4

Table 8: Guidance-strength ablation for TARO-2 on CIFAR-10 using WRN-28-10 under PGD-EOT-20. We report clean accuracy and robust accuracy under ℓ_∞ and ℓ_2 attacks.

γ	Clean Acc.	ℓ_∞ Robust Acc.	ℓ_2 Robust Acc.
0 (TARO-coarse)	81.90$\pm$0.6	70.24$\pm$1.0	71.80$\pm$1.2
0.5	82.54$\pm$0.6	74.14$\pm$0.3	74.92$\pm$0.9
1.0 (TARO-fine)	91.66$\pm$0.6	86.28$\pm$0.8	84.76$\pm$0.2
1.5	94.02$\pm$0.7	**87.95**$\pm$0.7	**85.93**$\pm$0.3
2.0	**94.52**$\pm$0.7	81.70$\pm$0.5	82.61$\pm$0.9
3.0	92.89$\pm$0.1	60.47$\pm$1.6	74.14$\pm$0.3

4.4 Guidance Strength Ablation

We study the temporal correction strength γ in TARO-2. Small γ keeps the target closer to the high-noise coarse expert, while larger γ increases the contribution of the low-noise fine expert. Table 8 reports results on CIFAR-10 with WRN-28-10 under PGD-EOT-20 for both ℓ_∞ ($\epsilon = 8/255$) and ℓ_2 ($\epsilon = 0.5$). It shows that the fine expert plays an important role in both clean and robust accuracy. When $\gamma < 1$, clean accuracy and robust accuracy both drop, which is consistent with the target remaining too close to the coarse expert and losing class-relevant detail. The robustness results peak at $\gamma = 1.5$, while larger values reduce robustness, suggesting that excessive correction can over-amplify unstable temporal residuals.

5 Conclusion

We proposed TARO, a test-time purification framework that treats diffusion timesteps as complementary denoising experts. TARO builds an affine coarse-to-fine target that combines robust global structure from high-noise experts with class-relevant detail from low-noise experts. Empirically, TARO improves robustness under strong adaptive attacks while maintaining high clean accuracy. Empirically, TARO achieves strong robustness across multiple datasets against ℓ_∞, ℓ_2, PGD-EOT, BPDA-EOT, and structural low-frequency attacks, significantly outperforming prior purification and test-time optimization baselines while preserving high clean accuracy. We further showed that TARO is complementary to adversary-aware residual modeling and that optimization-based purification remains robust in settings where standard diffusion purification collapses. We also highlight the need for proper adaptive evaluation of optimization-based purifiers which is discussed in detail in the Appendix. Additional experiments, studies and proofs are provided in the Appendix section.

Limitations. TARO increases inference latency due to iterative multi-scale optimization, and its correction strength must be selected carefully for the threat model; see Appendix A.13 for a detailed discussion.

AI use statement

In this work, we used generative AI tools for editing purposes. The main idea and the algorithms of the presented work are written by us and generative AI was used just for editing and general feedback

about the work. We take responsibility for the final content of this work, including text, claims or the attached code in this work.

Reproducibility statement

To make our work reproducible, we have attached our source code as a supplementary file. In addition to listing the details of the datasets, models, and hyperparameters used in Experiments and Appendix section, our code can be run using the attached zip file. Basic readme on how to run the scripts is also provided.

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*, pp. 274–283. PMLR, 2018.

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. *arXiv preprint arXiv:2003.09461*, 2020.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6643–6665. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24k.html.

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

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in neural information processing systems*, 34:4218–4233, 2021.

Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *International Conference on Learning Representations*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Andre Kassis, Urs Hengartner, and Yaoliang Yu. Diffbreak: Is diffusion-based purification robust? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Z7eXOBcxE9.

Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 134–144. IEEE, 2023.

Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yingzhe He, Jie Shi, and Xiaolin Hu. ADBM: Adversarial diffusion bridge model for reliable adversarial purification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=g0rnZeBguq.

Guang Lin, Zerui Tao, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Robust diffusion models for adversarial purification. *CoRR*, abs/2403.16067, 2024. URL https://doi.org/10.48550/arXiv.2403.16067.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:248811081.

Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International conference on machine learning*, pp. 17258–17277. PMLR, 2022.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International conference on machine learning*, pp. 36246–36263. PMLR, 2023.

Daniel Wesego and Pedram Rooshenas. Adversary aware optimization for robust defense. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VNbvk5pMS7.

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

Boya Zhang, Weijian Luo, and Zhihua Zhang. Enhancing adversarial robustness via score-based optimization. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 11547–11561, 2023. URL https://arxiv.org/pdf/2307.04333.

Mingkun Zhang, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. CLIPure: Purification in latent space via CLIP for adversarially robust zero-shot classification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TQ2ZOy6miT.

A Appendix

This appendix provides additional algorithmic, evaluation, and experimental details omitted from the main text due to space constraints. Appendix A.1 provides a local Gaussian interpretation of affine temporal extrapolation. Appendix A.2 shows proof of proposition 1, then Appendix A.3 describes TARO+AA variant that combines temporal multi-scale purification with the adversary-aware perturbation likelihood of AAOpt. Appendix A.4 formalizes the evaluation protocol used for optimization-based purifiers, and Appendix A.5 presents a corrected-evaluation case study showing that improperly differentiated optimization loops can lead to severely inflated robustness estimates. Appendix A.6 reports results using the Imagenette dataset. Appendix A.7 studies compute matched experiments showing the robustness of TARO does not arise from increasing the compute resource. Appendix A.8 includes additional ablations on time scales, Appendix A.9 studies how applying the multi-scale prior of TARO can be similarly applied to other baselines. Appendix A.10 reports the results against common corruptions using the CIFAR-10-C dataset, Appendix A.11 discusses the computation requirement and compares the wall clock time the algorithm takes, Appendix A.12 lists the hyperparameters used in our experiments, and finally, Appendix A.13 discusses the limitation and societal impact in detail.

A.1 Local Gaussian View of Affine Temporal Extrapolation

Positive-weight temporal PoE motivates aggregating diffusion experts, but it only yields interpolation. TARO uses a more general affine log-density combination,

$$\log p_{\text{aff}}(x) = \sum_{k=1}^{K} \gamma_k \log p_{t_k}(x) - \log Z, \qquad \sum_{k=1}^{K} \gamma_k = 1,$$

where the coefficients need not be nonnegative. When all $\gamma_k \geq 0$, this reduces to standard PoE interpolation. When some coefficients are negative, it becomes a contrastive or ratio-of-experts prior. For $K = 2$,

$$\log p_\gamma(x) = \gamma \log p_f(x) + (1-\gamma) \log p_c(x),$$

and for $\gamma > 1$,

$$p_\gamma(x) \propto \frac{p_f(x)^\gamma}{p_c(x)^{\gamma-1}}.$$

Thus, affine temporal extrapolation emphasizes structures supported by the fine expert relative to the coarse expert.

Under a local Gaussian approximation $p_{t_k}(x) \approx \mathcal{N}(\mu_{t_k}, \Sigma_{t_k})$, the affine log-density has local precision

$$\Lambda_{\text{aff}} = \sum_{k=1}^{K} \gamma_k \Lambda_{t_k},$$

provided this matrix is positive definite. If the fine expert has larger local precision than the coarse experts, and the affine weights place extrapolative positive mass on the fine expert while subtracting coarser experts, then

$$\Lambda_{\text{aff}} = \Lambda_f + \sum_{k>1} a_k (\Lambda_f - \Lambda_{t_k}) \succeq \Lambda_f.$$

This gives local intuition for why emphasizing the fine expert can sharpen the target and restore class-relevant structure. However, excessive sharpening may also amplify unstable cross-scale residuals, motivating validation-based selection of γ and the consistency regularizer used in our ablations.

It should be noted that unlike positive-weight PoE, affine extrapolation does not automatically preserve positive definiteness. The local Gaussian interpretation is valid only when $\Lambda_{\text{aff}} \succ 0$. In the two-expert case with $\gamma > 1$, a sufficient condition is $\Lambda_f \succeq \Lambda_c$, in which case

$$\Lambda_{\text{aff}} = \Lambda_f + (\gamma - 1)(\Lambda_f - \Lambda_c) \succeq \Lambda_f \succ 0.$$

A.2 Proof of Proposition 1

Proof. By assumption, the fine and coarse temporal experts admit the decompositions

$$u_f = x^\star + b_f(x^\star) + \xi_f, \qquad u_c = x^\star + b_c(x^\star) + \xi_c,$$

with $\mathbb{E}[\xi_f \mid x^\star] = \mathbb{E}[\xi_c \mid x^\star] = 0$. The TARO guided target is

$$u_\gamma = u_c + \gamma(u_f - u_c).$$

Substituting the decompositions of u_f and u_c, we obtain

$$u_f - u_c = b_f(x^\star) - b_c(x^\star) + \xi_f - \xi_c.$$

Therefore,

$$u_\gamma = x^\star + b_c(x^\star) + \xi_c + \gamma\left(b_f(x^\star) - b_c(x^\star) + \xi_f - \xi_c\right).$$

Rearranging terms gives

$$u_\gamma = x^\star + \underbrace{b_c(x^\star) + \gamma\big(b_f(x^\star) - b_c(x^\star)\big)}_{b_\gamma(x^\star)} + \underbrace{\xi_c + \gamma(\xi_f - \xi_c)}_{\xi_\gamma}.$$

Thus,

$$u_\gamma - x^\star = b_\gamma(x^\star) + \xi_\gamma.$$

Taking the squared norm,

$$\|u_\gamma - x^\star\|_2^2 = \|b_\gamma(x^\star) + \xi_\gamma\|_2^2.$$

Expanding,

$$\|u_\gamma - x^\star\|_2^2 = \|b_\gamma(x^\star)\|_2^2 + 2b_\gamma(x^\star)^\top \xi_\gamma + \|\xi_\gamma\|_2^2.$$

Now take conditional expectation given $x^\star$. Since

$$\xi_\gamma = \xi_c + \gamma(\xi_f - \xi_c),$$

and both ξ_f and ξ_c have zero conditional mean, we have

$$\mathbb{E}[\xi_\gamma \mid x^\star] = \mathbb{E}[\xi_c \mid x^\star] + \gamma\left(\mathbb{E}[\xi_f \mid x^\star] - \mathbb{E}[\xi_c \mid x^\star]\right) = 0.$$

Therefore, the cross term vanishes:

$$\mathbb{E}[2b_\gamma(x^\star)^\top \xi_\gamma \mid x^\star] = 2b_\gamma(x^\star)^\top \mathbb{E}[\xi_\gamma \mid x^\star] = 0.$$

Hence,

$$\mathbb{E}[\|u_\gamma - x^\star\|_2^2 \mid x^\star] = \|b_\gamma(x^\star)\|_2^2 + \mathbb{E}[\|\xi_\gamma\|_2^2 \mid x^\star].$$

Substituting the definitions of b_γ and ξ_γ, we obtain

$$\mathbb{E}[\|u_\gamma - x^\star\|_2^2 \mid x^\star] = \|b_c + \gamma(b_f - b_c)\|_2^2 + \mathbb{E}[\|\xi_c + \gamma(\xi_f - \xi_c)\|_2^2 \mid x^\star],$$

where we suppress the explicit dependence of b_f and b_c on $x^\star$ for readability.

This proves the proposition. □

A.3 TARO with Adversary-Aware Optimization

TARO can be extended with prior optimization-based purification objectives. In this section, we describe TARO+AA, a hybrid variant that integrates TARO with the adversarial perturbation likelihood introduced by AAOpt (Wesego & Rooshenas, 2025). This gives the whole MAP optimization objective that adds the adversarial likelihood term on top of the multi-scale TARO prior. This allows the optimization to exploit two complementary sources of information: the pretrained clean-image diffusion model supplies scale-consistent image structure, while the perturbation model supplies an adversary-aware likelihood over residual artifacts.

To achieve this, we add the $\lambda \ell_{pert}$ loss described in Wesego & Rooshenas (2025) to the original prior loss ℓ we have in TARO. As in TARO, K effective diffusion timescales are used. At each optimization step, TARO+AA evaluates the denoiser at these K scales, and uses the resulting multi-scale estimate target to define the diffusion prior loss. The perturbation loss is added similar to AAOpt from the residual $\boldsymbol{\delta} = \mathbf{x_i} - \mathbf{x}_{\text{adv}}$.

A.4 Proper Evaluation of Purification with Optimization

Optimization-based purification requires a stricter evaluation protocol than standard reverse-sampling diffusion defenses because the purifier is itself an inner optimization procedure. Let $f : \mathbb{R}^d \to \mathbb{R}^K$ be the classifier and let $P : \mathbb{R}^d \to \mathbb{R}^d$ be an optimization-based purifier. Given a clean image $\mathbf{x}$ and perturbation $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_p \leq \epsilon$, the adversarial input is $\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}$. A correct white-box adversary should optimize the complete purifier–classifier composition:

$$\max_{\|\boldsymbol{\delta}\|_p \leq \epsilon} \mathcal{L}_{\text{cls}} \left(f(P(\mathbf{x} + \boldsymbol{\delta})), y \right). \tag{4}$$

Thus, the relevant attack gradient is

$$\mathbf{g}_{\text{adv}} = \nabla_{\mathbf{x}_{\text{adv}}} \mathcal{L}_{\text{cls}} \left(f(P(\mathbf{x}_{\text{adv}})), y \right), \tag{5}$$

rather than the classifier gradient after purification.

For test-time optimization purifiers, $P(\mathbf{x}_{\text{adv}})$ is not a single forward pass. The purifier initializes $\mathbf{z}_0 = \mathbf{x}_{\text{adv}}$ and performs T inner updates using a guidance objective $\mathcal{L}_{\text{guide}}$:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{guide}}(\mathbf{z}_t; \phi), \qquad t = 0, \ldots, T-1, \tag{6}$$

where ϕ denotes the parameters of the guidance model, such as a diffusion denoiser, score model, CLIP model, or adversarial perturbation model. The final purified image is $\mathbf{x}^\star = P(\mathbf{x}_{\text{adv}}) = \mathbf{z}_T$.

A correct attack must differentiate through the full dependence of $\mathbf{z}_T$ on $\mathbf{z}_0 = \mathbf{x}_{\text{adv}}$. Applying the chain rule gives

$$\frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{x}_{\text{adv}}} = \frac{\partial \mathcal{L}_{\text{cls}}}{\partial f} \frac{\partial f}{\partial \mathbf{z}_T} \frac{\partial \mathbf{z}_T}{\partial \mathbf{z}_0}. \tag{7}$$

The final factor is the trajectory Jacobian of the inner purification process. For the gradient-descent update above, each local Jacobian satisfies

$$\frac{\partial \mathbf{z}_{t+1}}{\partial \mathbf{z}_t} = \mathbf{I} - \eta \nabla^2_{\mathbf{z}_t} \mathcal{L}_{\text{guide}}(\mathbf{z}_t; \phi). \tag{8}$$

Therefore, differentiating through an optimization-based purifier requires preserving second-order information from the inner guidance gradient. More generally, if the purifier uses an optimizer update

$$\mathbf{z}_{t+1} = U_t(\mathbf{z}_t, \mathbf{g}_t), \qquad \mathbf{g}_t = \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{guide}}(\mathbf{z}_t; \phi), \tag{9}$$

then the chain rule gives

$$\frac{\partial \mathbf{z}_{t+1}}{\partial \mathbf{z}_t} = \frac{\partial U_t}{\partial \mathbf{z}_t} + \frac{\partial U_t}{\partial \mathbf{g}_t} \nabla^2_{\mathbf{z}_t} \mathcal{L}_{\text{guide}}(\mathbf{z}_t; \phi). \tag{10}$$

Thus, when the inner guidance gradient is treated as a constant, the attack omits the Hessian-dependent term and no longer corresponds to a true white-box attack on the optimization-based purifier.

This failure mode is easy to introduce in PyTorch. If the guidance gradient is computed inside the purification loop with the default setting,

$$\texttt{torch.autograd.grad}\left(\mathcal{L}_{\text{guide}}, \mathbf{z}_t\right), \tag{11}$$

then `create_graph=False` by default, and the returned gradient is detached from the graph needed by the outer attack. The inner update still changes the purified image, but the outer adversary cannot differentiate through how this update changes as $\mathbf{x}_{\text{adv}}$ changes. In effect, the attacker sees an incomplete surrogate of the purifier. A graph-consistent unrolled evaluation must instead preserve the higher-order path:

$$\texttt{torch.autograd.grad}\left(\mathcal{L}_{\text{guide}}, \mathbf{z}_t, \texttt{create_graph=True}, \texttt{retain_graph=True}\right). \tag{12}$$

Here, `create_graph=True` is the essential flag because it keeps the gradient computation differentiable; `retain_graph=True` is needed when the same graph is reused for additional backward passes.

This graph-break issue is distinct from stochastic gradient masking. Diffusion purifiers often contain stochastic denoising steps, so expectation over transformation (EOT) is required to average gradients over noise realizations. However, EOT alone does not repair a detached inner optimization graph. A

defense may use many EOT samples and still be mis-evaluated if the attacker cannot differentiate through the optimization trajectory that produced each purified sample. Therefore, optimization-based purification requires both stochasticity-aware evaluation and graph-consistent differentiation.

Our evaluation follows this principle. PGD-EOT attacks the full purifier–classifier composition, while full reverse-optimization AutoAttack-EOT in Appendix 4.3.1 provides an additional check in which the attack differentiates through the complete optimization trajectory. The corrected CLIPure Zhang et al. (2025) case study below illustrates why this requirement is essential.

A.5 Corrected Evaluation Case Study: CLIPure

We revisit CLIPure Zhang et al. (2025) as a case study in graph-consistent evaluation. CLIPure performs purification through test-time optimization using either a cosine-similarity objective in CLIP space or a diffusion-based objective. Since the purified image is produced by an inner optimization loop, a white-box attack must differentiate through that loop. If the inner guidance gradient is computed without preserving higher-order dependencies, the purification update becomes partially detached from the adversarial input, and the attack receives an incomplete gradient.

This produces accidental gradient masking. The defense appears robust not because the purified image is invariant to adversarial perturbations, but because the attacker is unable to observe how the purification trajectory changes as $\mathbf{x}_{\mathrm{adv}}$ changes. Restoring the computational graph through the inner loop exposes the true bilevel structure of the attack.

Table 9 compares the originally reported CLIPure robust accuracies with corrected graph-consistent evaluation on CIFAR-10 under AutoAttack with ℓ_∞ perturbations and $\epsilon = 8/255$. Under corrected evaluation, CLIPure-Cos drops from 91.1% to 0.0%, and CLIPure-Diff drops from 88.0% to 0.0%.

Table 9: Reported versus corrected robust accuracy for CLIPure variants on CIFAR-10 under AutoAttack (ℓ_∞, $\epsilon = 8/255$). Corrected evaluation preserves the computational graph through the inner optimization loop.

Method	Reported Robust Acc.	Corrected Robust Acc.
CLIPure-Cos	91.1%	0.0%
CLIPure-Diff	88.0%	0.0%

These results show that optimization-based purification can be severely mis-evaluated when higher-order gradient flow is not preserved. They do not imply that test-time optimization defenses are inherently ineffective; rather, they show that robustness must be established under graph-consistent attacks. In the next section, we therefore report AutoAttack-EOT with full reverse optimization for TARO and the strongest optimization baselines, providing a stricter adaptive evaluation beyond the one-step gradient approximation used in the main PGD-EOT experiments.

A.6 Imagenette Results

In addition to the CIFAR datasets, we evaluate TARO on Imagenette, a higher-resolution subset of ImageNet. Table 10 shows that TARO remains effective on a more complex image distribution. TARO-2 achieves the highest robust accuracy, while maintaining clean accuracy close to the undefended classifier and improving substantially over Score-Opt-O. These results indicate that temporal rectification is not limited to low-resolution CIFAR images and can scale to more challenging image distributions without severe clean-accuracy degradation.

A.7 Computation Ablation

Because TARO-K uses K denoising experts per optimization step, we compare TARO-2 and TARO-3 against TARO-1 under increased computational budgets. In Table 11, we increase the number of defense EOT for TARO-1 optimization by factors of two and three to match the function-evaluation cost of TARO-2 and TARO-3. Increasing the number of defense EOT gives only marginal gains. TARO-2 and TARO-3 remain stronger under the matched defense-EOT setting, indicating that its improvements are due to the temporal multi-scale prior rather than simply additional computation.

Table 10: PGD-EOT attack on Imagenette (ℓ_∞, $\epsilon = 8/255$), WRN-28-10.

Method	Clean	Robust
Default	82.9	0.0
Score-Opt-O	76.4$\pm$1.2	69.5$\pm$1.7
AAOpt	80.4$\pm$1.4	66.4$\pm$1.9
TARO-2	81.6$\pm$1.4	**72.3**$\pm$2.2

Table 11: Robustness of TARO variants with additional compute on CIFAR-10 (ℓ_∞, $\epsilon = 8/255$), WRN-28-10.

Model	Clean	Robust
TARO-1	91.66$\pm$0.6	86.28$\pm$0.8
TARO-1-dEOT$\times$2	91.44$\pm$0.8	86.21$\pm$0.8
TARO-1-dEoT$\times$3	91.52$\pm$0.7	86.36$\pm$0.8
TARO-2	93.78$\pm$1.0	87.26$\pm$1.5
TARO-3	94.02$\pm$0.9	**87.41**$\pm$1.1

A.8 Time-Scale Scheduling Ablation

We further evaluate the generality of TARO across different time-scale scheduling choices. The core mechanism of TARO is multi-scale temporal aggregation: the purifier combines complementary denoising experts evaluated at different noise regimes. This mechanism should not depend on a single hand-designed schedule for selecting the noisier expert. In the main experiments, we use the square-root schedule $\tau = \sqrt{2t}$, which scales the variance-like diffusion parameter by a factor of 2. To test whether TARO remains effective under an alternative parameterization, we also evaluate a linear schedule that directly scales the diffusion time as $\tau = 2t$ for the noisier expert. Table 12 shows that TARO remains robust under both schedules, with the linear schedule achieving comparable clean accuracy and slightly higher robust accuracy. These results indicate that TARO's gains arise from temporal multi-scale rectification rather than a brittle dependence on a particular time-scaling rule.

Table 12: Time-scale scheduling ablation on CIFAR-10 under PGD-EOT-20 (ℓ_∞, $\epsilon = 8/255$), WRN-28-10. The square-root schedule uses $\tau = \sqrt{2t}$ for the noisy expert, while the linear schedule uses $\tau = 2t$.

Time-Scale Schedule	Clean Accuracy	Robust Accuracy
Square-root scaling: $\tau = \sqrt{2t}$	93.78$\pm$1.0	87.26$\pm$1.5
Linear scaling: $\tau = 2t$	94.09$\pm$1.1	87.76$\pm$1.3

A.9 Applying Multi-Scale Prior To Other Methods

To examine whether the benefits of the multi-scale prior extend beyond TARO, we incorporate it into AAOpt. We denote the variants using two and three diffusion noise scales as AAOpt-2 and AAOpt-3, respectively. Table 13 reports results on CIFAR-10 with WRN-28-10 under PGD-EOT-20 attacks with ℓ_∞ budget $\epsilon = 8/255$. Robust accuracy increases from 84.76% for AAOpt to 85.74% for AAOpt-2 and 86.32% for AAOpt-3, while clean accuracy improves from 92.18% to 92.57% and 93.55%, respectively. Thus, AAOpt-2 and AAOpt-3 improve clean and robust accuracy over the baseline. These results suggest that aggregating information across diffusion noise scales can also benefit other baselines, supporting the complementary roles of multi-scale prior modeling with other methods.

Table 13: Multi Time-scale Prior on AAOpt on CIFAR-10 under PGD-EOT-20 (ℓ_∞, $\epsilon = 8/255$), WRN-28-10.

Method	Clean Accuracy	Robust Accuracy
AAOpt	92.18	84.76
AAOpt-2	92.57	85.74
AAOpt-3	93.55	86.32

A.10 Robustness Against Common Corruptions

We evaluate robustness to natural distribution shifts using CIFAR-10-C. Unlike adversarial attacks, these corruptions are not optimized against the classifier or purifier, but they test whether the multi-scale prior also improves robustness to common image degradations. Table 14 reports accuracy on five corruption types using WRN-70-16. TARO-2 achieves the highest mean accuracy, outperforming Score-Opt-O and AAOpt. The improvement is especially strong on Gaussian noise, where TARO-2 reaches 92.0% accuracy. The pure TARO-2 result suggest that the unconstrained multi-scale consensus is particularly effective for non-adversarial corruptions where an adversarial residual likelihood is unnecessary.

Table 14: Robustness against common corruptions on CIFAR-10-C (WRN-70-16). The **Mean** column averages across the five corruption types.

Method	Gaussian	Elastic	JPEG	Snow	Brightness	**Mean**
Default	45.6	82.9	75.8	83.5	93.8	76.3
Score-Opt-O	79.8	78.5	86.1	82.7	89.6	83.3
AAOpt	85.3	84.8	89.5	85.3	90.5	87.1
TARO-2	**92.0**	84.3	**90.5**	**87.3**	**91.3**	**89.1**

A.11 Computational Resources and Runtime

Our experiments were conducted on NVIDIA A100 GPUs. Table 15 reports the wall-clock time required to complete a 5-step purification trajectory on CIFAR-10 with batch size 128. TARO introduces a modest increase in runtime relative to single-scale baselines because it evaluates multiple denoising experts per optimization step. However, this additional cost remains practical and is substantially smaller than the cost of extending a single-scale optimization trajectory by a factor of 2.

Table 15: Runtime of purification optimization on CIFAR-10 using an NVIDIA A100 GPU.

Model	Batch Size	Time
Score-Opt-O	128	1.16 sec.
Score-Opt-N	128	1.26 sec.
AAOpt	128	1.05 sec.
TARO-2	128	1.38 sec.

A.12 Experiment Details

We use 512 samples for each trial unless otherwise stated and report mean and standard deviation over trials. The DiffBreak LF evaluation is run on 64 samples averaged over 3 trials due to the cost of sequential EOT and 100 attack steps. For the main PGD-EOT experiments, we use 20 EOT samples for both attack and defense across TARO and directly reimplemented baselines. This differs from the original AAOpt evaluation, which uses 20 EOT samples for attack but a larger number of defense EOT samples.

For TARO-K, we use stream factors $[t, 2]$ for $K = 2$ and $[t, 1, 2]$ for $K = 3$. Tables 16, 17 summarize the main hyperparameters used in our experiments. Additional implementation details are included in the supplementary code.

Table 16: Hyperparameters used for TARO-2.

Dataset	Learning Rate	Num of Iterations	Guidance Strength γ	Time Schedule
CIFAR-10	0.02	20	1.4	$[0.25, \ldots, 0]$
CIFAR-100	0.1	3	1.4	$[0.25, \ldots, 0]$

Table 17: Hyperparameters used for TARO-3.

Dataset	Learning Rate	Num of Iterations	Guidance Strength γ	Time Schedule
CIFAR-10	0.02	20	1.1	$[0.25, \ldots, 0]$
CIFAR-100	0.1	3	1.1	$[0.25, \ldots, 0]$

A.13 Limitations and Discussion

TARO improves diffusion-based test-time purification by aggregating denoising estimates across multiple temporal regimes, but it has several limitations. First, TARO is an empirical defense and does not provide certified robustness. The temporal PoE and affine aggregation interpretations provide intuition for why multi-scale denoising can improve purification, but the local Gaussian view is only a conditional explanation: it relies on assumptions such as approximate local Gaussianity, positive definiteness of the signed precision, and suitable ordering between fine and coarse experts.

Second, TARO increases inference cost relative to single-scale purification methods because each optimization step evaluates multiple denoising experts. Our runtime ablations suggest that this additional cost is moderate. Nevertheless, TARO remains more expensive than a standard classifier forward pass or a single denoising trajectory.

Third, as with all diffusion purification methods, TARO's performance depends on the quality of the pretrained diffusion prior, the downstream classifier, and the choice of optimization hyperparameters. In particular, the correction strength γ, the learing rate, the number of iterations and the time schedule need to be carefully selected based on validation dataset.

Finally, TARO is designed and evaluated as a test-time image purification method for classification. Extending the temporal multi-expert prior to other modalities, detection, segmentation, inverse problems, or certified defenses remains an important direction for future work.

Societal impact. TARO contributes to the development of more reliable machine learning systems by improving robustness against adversarial manipulation at inference time, which is important for safety-critical applications where models may be exposed to malicious inputs. The proposed proper and reliable evaluation protocol also has a positive societal impact by discouraging misleading robustness claims and promoting more reliable security evaluation. At the same time, stronger purification defenses may increase the computational cost and energy use of deployed systems, especially when repeated stochastic evaluations are required. We view TARO as part of a broader effort toward transparent, reproducible, and carefully evaluated defense rather than as a complete solution to adversarial risk.