# MULTIMODAL INFORMATION IS ALL YOU NEED FOR ADVERSARIAL PURIFICATION VIA DIFFUSION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Adversarial defense aims to find true semantic labels of adversarial examples, where diffusion-based adversarial purification as intriguing adversarial defense methods can restore data perturbed by unseen attacks to clean distribution without training classifiers. However, unimodal diffusion-based approaches rely on noise schedules to implicitly preserve labels, whereas recently proposed multimodal variants add textual control but require adversarial training and heavy distillation. Both approaches lack theoretical guarantees. In this work, we propose MultiDAP that uses multimodal diffusion models for adversarial purification. MultiDAP first learn prompts from clean text-image pair data for clean image generation, where context tokens are numerical instead of text templates such as "a photo of a ·" for rich contextual information and hence enhance adversarial robustness. Given learned prompts and adversarial examples, MultiDAP then purify inputs via minimizing regularized DDPM losses iteratively for only a few steps. Theoretical guarantees for two phases are also provided. In experiments, our proposed model achieve improvement of zero-shot adversarial defense performance over unimodal diffusion models and multimodal variants with text templates.

## 1 INTRODUCTION

Adversarial defense is fundamentally concerned with recovering the true semantic label information from adversarial examples which are perturbed by human imperceptible but carefully crafted noise for deep learning classifiers to predict incorrect labels (Goodfellow et al., 2014). Adversarial purification has recently emerged as a promising paradigm for adversarial defense (Shi et al., 2021; Yoon et al., 2021; Nie et al., 2022; Wang et al., 2022a; Bai et al., 2024; Lei et al., 2025). Unlike adversarial training (Croce and Hein, 2020; Laidlaw et al., 2021; Dolatabadi et al., 2022; Wang et al., 2023a), which explicitly trains classifiers on adversarial examples, adversarial purification methods employ generative models to remove adversarial perturbations before classification (Song et al., 2017; Nie et al., 2022). This strategy offers two key advantages. First, it does not need to retrain classifiers on generated attacks, thereby reducing computational overhead. Second, it provides stronger generalization to unseen adversarial attacks, as adversarial purification directly restores clean data distributions. However, early adversarial purification methods are based on generative adversarial networks (GANs) and energy-based models (EBMs) and fall behind adversarial training methods, because of their limited generative power (Nie et al., 2022).

Diffusion models have rapidly become the mainstream approach for adversarial purification due to their remarkable generative power and ability to approximate complex data distributions (Nie et al., 2022; Wang et al., 2022a; Chen et al., 2024a; Zhang et al., 2025; Bai et al., 2024). By progressively adding Gaussian noise and removing it, diffusion models can effectively reconstruct clean samples from corrupted or perturbed inputs, making them particularly well-suited for removing adversarial perturbations (Nie et al., 2022). However, most existing work relies on unimodal diffusion models which attempt to preserve semantic information implicitly by injecting Gaussian noise to a specific level in the forward process and then denoising the input. Hence unimodal approaches often struggle to fully obtain semantic label information, limiting defense against stronger or adaptive attacks.

To address this limitation, one step control purification has recently been proposed as the first multimodal diffusion model for adversarial purification. It leverages ControlNet to include additional modalities (e.g., textual prompts) for adversarial purification and thus preserves more semantic label

information than unimodal approaches (Lei et al., 2025). This importance of multi-modal information was first discovered from the fact that human can easily identify the true label of adversarial examples, in contrast to deep learning models on solely pixel spaces. The reason is that human cognition relies on semantic information from the context and is immune to distribution variations induced by adversarial attacks, whereas the deep learning models classify images via statistical distributional associations (Zhou et al., 2024) and are vulnerable to adversarial attacks. However, one step control purification still faces three critical challenges: (i) it lacks theoretical guarantees regarding purification effectiveness, and (ii) it still relies on adversarial training to learn robust cross-modal alignment, and more importantly, (iii) without knowledge distillation, the iterative reverse process of diffusion models will incur substantial computational overhead. These efficiency issues limit their practicality for real-time or large-scale adversarial defense.

In this paper, we propose Multimodal Diffusion for Adversarial Purification (MultiDAP) which leverages a single text-to-image diffusion model backbone. Unlike unimodal diffusion models for adversarial purification solely relying on image features, our approach conditions the diffusion model on textual prompts to infuse semantic information. In order to obtain the prompts which steer zero-shot adversarial purification, we design a paradigm to learn prompts for stable diffusion models from clean large-scale text-image pairs. This design not only leverages the powerful generative capacity of diffusion models but also capitalizes on the rich contextual representations encoded by the text encoder. With prompt-based conditioning, we introduce a more expressive feature space that distinguishes adversarial attacks from genuine content more effectively. Furthermore, we also propose to efficiently purify adversarial examples via prompt-guided likelihood maximization which only requires a few purification steps. Experiments demonstrate that MultiDAP achieves superior zero-shot adversarial defense performance compared to unimodal diffusion models on the CIFAR-10 dataset. These results highlight the dual contribution of our work: introducing a theoretically grounded framework for adversarial purification and delivering practical improvements for real-world deployment.

## 2 Related Work

**Unimodal Diffusion Models for Adversarial Defense.** Diffusion models have demonstrated remarkable performance in generative tasks, owing to their ability to progressively refine noisy data to high-quality output. This generative nature has been explored for robustness in various contexts, including adversarial purification (Nie et al., 2022), adversarial training (Wang et al., 2023b) and robust classification methods (Chen et al., 2024a). A notable application of diffusion models lies in purification-based defenses, where adversarially perturbed inputs are restored to their clean counterparts. Methods leveraging guided diffusion models have shown efficacy in removing perturbations while preserving the underlying data features, making them suitable for tasks like classification (Lee and Kim, 2023; Xiao et al., 2022; Bai et al., 2024; Yeh et al., 2024). Additionally, diffusion-based classifier have gained traction by integrating generative and discriminative modelling (Zimmermann et al., 2021; Clark and Jaini, 2023; Chen et al., 2024a). Their robustness to input perturbations and adversarial attack is attributed to optimal empirical score function (Chen et al., 2024b). While these approaches highlight the versatility of diffusion models in adversarial defense, they often face challenges in efficiency, effectiveness from unimodality and theoretical guarantees, motivating further investigations.

**Multimodal Approaches in Adversarial Defense.** Recent advances in vision-language models (VLMs) have demonstrated potentials of multimodal information in improving robustness. Models such as CLIP (Radford et al., 2021) learn joint embeddings of images and text, enabling strong cross-modal alignment that provides richer semantic priors than unimodal vision models. This multimodal alignment has inspired adversarial finetuning (Schlarmann et al., 2024), adversarial prompt tuning (Zhang et al., 2024; Li et al., 2024; Sheng et al., 2025), and multimodal defenses leveraging vision-language pretraining (Wang et al., 2025). These approaches hold clear advantages: auxiliary modalities such as text can act as high-level semantic constraints, guiding models toward correct semantic label predictions. Nevertheless, current multimodal robustness methods face critical limitations. Many rely on adversarial training to establish robust cross-modal alignment, and recently proposed first multimodal diffusion model for adversarial purification: one step control purification in particular suffer from substantial computational overhead due to multi-step denoising if knowl-

edge distillation is not used (Lei et al., 2025). Furthermore, theoretical guarantees remain largely absent, leaving their robustness difficult to formally assess. These challenges motivate the need for approaches that combine multimodal semantic priors with efficiency and provable purification guarantees—precisely the focus of our proposed Multimodal Diffusion for Adversarial Purification (MultiDAP).

# 3 Multimodal Diffusion Models for Adversarial Purification

## 3.1 Problem Setup

Let $x \in \mathcal{X}$ denote a clean input with label $y$, and $x^{adv} = x + \delta$ be an adversarial example generated under perturbation constraint $\|\delta\|_p \leq \epsilon$. Here the perturbation $\delta$ is constrained under an $\ell_p$-norm threat model, with $p \in \{2, \infty\}$ being the most common cases. The $\ell_\infty$ attack bounds the maximum per-pixel distortion, ensuring imperceptibility, while the $\ell_2$ attack restricts the overall perturbation energy.

Adversarial purification aims to transform an adversarial input $x^{adv}$ back to a sample $x^{pur}$ that lies close to the clean data manifold, such that $f(x^{pur}) = y$. Recent works have demonstrated that diffusion models are particularly well suited for this task, due to their strong generative ability to approximate complex data distributions (Nie et al., 2022).

A diffusion model (Song et al., 2020) defines a forward noising process that gradually perturbs clean data $x_0$ into Gaussian noise through a sequence of latent variables $\{x_t\}_{t=0}^{T}$:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\Big)\,,$$

where $\{\beta_t\}$ is a variance schedule (Ho et al., 2020). This process ensures that as $t \to T$, the sample $x_T$ approaches pure noise, as $T$ is large enough. The reverse denoising process is parameterized by a neural network $\epsilon_\theta$, which predicts the added noise and iteratively reconstructs clean data:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))\,.$$

The mean term $\mu_\theta(x_t, t)$ is computed from this noise prediction via a closed-form reparameterization: $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t)\right)$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The covariance $\Sigma_\theta(x_t, t)$ is typically fixed by the variance schedule $\{\beta_t\}$, though some variants allow it to be partially learned for improved sample quality (Nichol and Dhariwal, 2021). Together, $\mu_\theta$ and $\Sigma_\theta$ define the Gaussian reverse step, while $\epsilon_\theta$ remains the core predicted quantity that drives the denoising trajectory.

For adversarial purification, the intuition is to inject the adversarial input $x^{adv}$ into the forward process at a chosen noise level $t$, so that adversarial perturbations are drowned out by Gaussian noise (Nie et al., 2022). Then, the reverse process denoises $x_t$ step by step, ideally converging to a purified sample $x^{pur}$ close to the clean distribution. Formally, the purification mapping can be written as:

$$x^{pur} \sim P(x^{adv}) = p_\theta(x^{pur} \mid x_t^{adv}, t), \quad \text{with } x_t^{adv} \sim q(x_t \mid x^{adv}).$$

Here $x_t^{adv}$ denotes the adversarial input injected into the forward noising process at step $t$, and $x^{pur}$ is the purified output after reverse diffusion. This framework has achieved strong empirical robustness across various benchmarks (Nie et al., 2022; Wang et al., 2022b; Chen et al., 2024a).

However, existing diffusion-based purification suffers from two main drawbacks: (i) the denoising process is essentially *unimodal*, since it is conditioned only on Gaussian noise schedules without leveraging explicit semantic cues (text prompts), which limits its ability to preserve class-consistent information; and (ii) the multi-step reverse process is computationally expensive, making such defenses inefficient for real-time deployment. These limitations motivate our proposed Multimodal Diffusion for Adversarial Purification with explicit prompt guidance and improved efficiency.

## 3.2 Stable Diffusion with Prompt Learning

While diffusion-based purification can remove adversarial perturbations, prior defenses typically rely on *small, unimodal* diffusion models to approximate the data distribution (Nie et al., 2022). Limited representational capacity often leads to suboptimal likelihood estimates and unstable denoising trajectories, where semantic information may not be faithfully preserved.

To address this limitation, we adopt Stable Diffusion—a large-scale latent diffusion model (LDM)—as our backbone (Rombach et al., 2022). Pretrained on massive image–text corpora, Stable Diffusion provides substantially stronger modeling power and a richer, more informative likelihood landscape than small diffusion models. Moreover, operating in a compact latent space enables high-resolution synthesis with improved efficiency compared to pixel-space diffusion (Rombach et al., 2022; Dhariwal and Nichol, 2021). This stronger backbone lets our purifier start denoising from a more faithful approximation of the clean data manifold, reducing reliance on long reverse diffusion chains and mitigating semantic drift.

Formally, Stable Diffusion operates in a latent space defined by a variational autoencoder (VAE). Given an input image $x$, the encoder maps it into a compact latent representation $z = \mathcal{E}_{\text{VAE}}(x)$, where $\mathcal{E}_{\text{VAE}}$ denote the encoder. Given a class label or textual description $y$, we obtain a prompt embedding through a text encoder $e_p = \mathcal{E}_{\text{text}}(p)$, where $p$ is the input text, such as 'a photo of a cat'. In our approach, these embeddings serve as semantic conditions that guide the purification process. The denoising network $\epsilon_\theta(x_t, t, e_p)$ is implemented as a U-Net with cross-attention, which predicts the noise at each timestep. The denoising network $\epsilon_\theta(x_t, t, e_p)$ is trained with the standard denoising diffusion probabilistic model (DDPM) objective (Ho et al., 2020), which treats noise prediction as score matching:

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{x_0, e_p, t, \epsilon \sim \mathcal{N}(0,I)} \left[ \left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\ t,\ e_p\right) \right\|_2^2 \right],$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denote the variance schedule. This loss enforces the network to accurately predict the added Gaussian noise at each timestep, which is equivalent to maximizing a variational lower bound on the conditional data likelihood. In our case, the conditioning $e_p$ provides semantic priors that explicitly align the denoising trajectory with the true class, thereby enhancing the stability and fidelity of purification.

**Prompt Learning Objective.** A central challenge for purification-based defenses lies in the accuracy of likelihood estimation during denoising. Although Stable Diffusion provides a strong backbone, its conditioning typically depends on fixed or manually designed text prompts, which may be generic and fail to provide task-specific guidance. Such limitations are particularly critical for adversarial purification, where the model must recover the clean data distribution from inputs corrupted by imperceptible but adversarial perturbations. To overcome this issue, we propose a *prompt learning* module that explicitly optimizes prompt embeddings from clean data, allowing the model to acquire semantic priors that are robust to adversarial noise.

In general, a prompt $p$ can be represented as a concatenation of $M$ learnable context tokens,

$$p_{\text{context}} = [v_1, v_2, \ldots, v_M],$$

where each $v_m \in \mathbb{R}^d$ has the same dimensionality as the text encoder's word embeddings (e.g., $d = 512$ for CLIP). In prior works, such context tokens are often combined with a class-specific token (e.g., the word "cat"), yielding a class-dependent prompt $p = [p_{\text{context}}, p_{\text{class}}]$ that provides label-conditioned guidance. By contrast, our objective is to design a *class-agnostic prompt* that captures global semantic priors without relying on class labels. This choice is crucial for adversarial purification, since the ground-truth label of an adversarial input is typically unknown at inference time. We therefore optimize a shared prompt vector $p$ directly from clean data, such that it enhances the unconditional likelihood estimation of the diffusion model.

Our prompt learning module is optimized by reusing the standard DDPM noise-prediction loss, with the key difference that only the prompt parameters $p$ are updated while the diffusion backbone $\theta$ remains frozen:

$$\mathcal{L}_{\text{prompt}}(p) = \mathbb{E}_{x_0, t, \epsilon}\left[ \left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\ t,\ p\right) \right\|_2^2 \right].$$

Optimizing this loss is equivalent to maximizing a variational lower bound on the likelihood $p_\theta(x_0 \mid p)$. Thus, the learned prompt $p^*$ serves as a universal semantic prior that stabilizes the denoising trajectory and improves the fidelity of adversarial purification without requiring class labels or adversarial training.

Similar to prompt learning in CLIP (Zhou et al., 2022), we optimize the learnable context tokens using a gradient-based method, such as Adam (Kingma, 2014)). In each training iteration, we sample

**Algorithm 1:** Prompt Learning on Stable Diffusion (class-agnostic)

**Input:** Frozen diffusion backbone $\theta$ (VAE, U-Net $\epsilon_\theta$), clean images $\{x^{(b)}\}$, steps $T$, optimizer (Adam), prompt length $M$, iters $N$.
**Output:** Learned class-agnostic prompt $p^\star = [v_1, \ldots, v_M]$.

Initialize learnable tokens $p = [v_1, \ldots, v_M]$ (random or text-init).
**for** $n = 1, \ldots, N$ **do**
  Sample a mini-batch $\{x^{(b)}\}$; sample $t \sim \mathrm{Unif}(\{1, \ldots, T\})$ and $\epsilon \sim \mathcal{N}(0, I)$.
  $x_t \leftarrow \sqrt{\bar{\alpha}_t}\, x^{(b)} + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$
  $\mathcal{L}_{\text{prompt}} \leftarrow \left\| \epsilon - \epsilon_\theta(x_t, t, p) \right\|_2^2$ (average over batch).
  Update $p \leftarrow \mathrm{Adam}\big(p, \nabla_p \mathcal{L}_{\text{prompt}}\big)$ while keeping $\theta$ frozen.
**return** $p^\star \leftarrow p$.

a mini-batch of clean data points $x^{(b)}$, apply the forward diffusion process to obtain noisy latents $x_t^{(b)}$ with Gaussian noise $\epsilon$, and evaluate the prompt loss $\mathcal{L}_{\text{prompt}}$ with the current tokens $p = [v_1, \ldots, v_M]$. The gradients are then backpropagated through the denoising network $\epsilon_\theta(x_t^{(b)}, t, p)$ to update the prompt parameters. This process is repeated until convergence, yielding a shared prompt vector $p^*$ that minimizes the denoising objective. The detailed optimization procedure is summarized in Algorithm 1. Compared to full model fine-tuning, optimizing only a small set of prompt parameters significantly reduces trainable variables, which mitigates overfitting and keeps computational cost manageable, while still providing strong semantic guidance for purification.

**Theoretical Guarantee: Prompt Learning Improves Likelihood.** We show that optimizing the class-agnostic prompt $p$ with the DDPM objective monotonically increases a variational lower bound (ELBO) of the unconditional data likelihood under a fixed diffusion model $\theta$.

**Theorem 1** (Prompt learning improves the likelihood lower bound)**.** *Let $x_0 \sim p_{data}$ denote clean latents, and let $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. Fix the diffusion backbone parameters $\theta$, and optimize only the prompt $p$ using the DDPM objective. Then the optimal prompt $p^\star = \arg\min_p \mathcal{L}_{prompt}(p)$ maximizes the evidence lower bound (ELBO) on the data likelihood $p_\theta(x_0 \mid p)$,*

$$\log \underline{p}_\theta(x_0 \mid p^\star) \;\geq\; \log \underline{p}_\theta(x_0 \mid p), \quad \forall\, p,$$

*where $\log \underline{p}_\theta$ denotes the variational lower bound.*

Moreover, $\nabla_p \mathcal{L}_{\mathrm{VLB}}(p)$ and $\nabla_p \mathcal{L}_{\mathrm{prompt}}(p)$ are colinear since the weights $w_t$ are positive. Thus updating $p$ along $-\nabla_p \mathcal{L}_{\mathrm{prompt}}$ strictly decreases $\mathcal{L}_{\mathrm{VLB}}$ for sufficiently small step size, thereby monotonically increasing the likelihood.

**Corollary 1** (Score matching view)**.** *The objective $\mathcal{L}_{prompt}$ is equivalent to minimizing a weighted Fisher divergence between the conditional score $\nabla_{x_t} \log p_\theta(x_t \mid p)$ and the forward diffusion score $\nabla_{x_t} \log q(x_t \mid z_0)$. Hence optimizing $p$ aligns the model score with the true score, which directly improves the* data likelihood $\log p_\theta(x_0 \mid p^\star) \;\geq\; \log p_\theta(x_0 \mid p), \forall\, p$.

### 3.3 Prompt-Guided Likelihood Maximization for Purification

Given an adversarial input $x^{adv}$, we purify it by maximizing the model likelihood under the learned class-agnostic prompt $p^\star$ while using the pretrained diffusion backbone $\theta$. We obtain a noisy image by the forward diffusion

$$x_{t^\star}^{adv} \;=\; \sqrt{\bar{\alpha}_{t^\star}}\, x_0^{adv} \;+\; \sqrt{1 - \bar{\alpha}_{t^\star}}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

Our goal is to recover an $x_0$ that maximizes the posterior (or the conditional likelihood surrogate)

$$x_0^{\text{pur}} \;\in\; \arg\max_{x_0} \; \log p_\theta(x_{t^\star}^{adv} \mid x_0, p^\star) + \log p(x_0), \tag{1}$$

where $p(x_0)$ is the prior. Maximizing (1) is intractable directly, so we instead minimize the purification "simple loss" with respect to the image variable $x_0$, while conditioning on the learned prompt $p^\star$:

$$x_0^{\text{pur}} \;\in\; \arg\min_{x_0} \underbrace{\mathbb{E}_{t,\epsilon} \left\| \epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon,\; t,\; p^\star\big) \right\|_2^2}_{=:\, \mathcal{L}_{\text{DDPM}}(x_0; p^\star)} + \lambda\, \mathcal{R}(x_0, x^{adv}), \tag{2}$$

**Algorithm 2:** Purification via Regularized DDPM-Loss Minimization in Pixel Space

**Input:** Adversarial image $x^{adv}$, learned prompt $p^\star$, frozen $\theta$, steps $T_1$ $T_2$, Purification steps $N$ (e.g., 5), step size $\eta$, optional regularizer weight $\lambda$.

**Output:** Purified image $x^{\text{pur}}$.

$x_0^{(0)} \leftarrow x^{adv}$ **for** $n = 0, \dots, N-1$ **do**

  Sample $t \sim \text{Unif}(\{T_1, \dots, T_2\})$ and $\epsilon \sim \mathcal{N}(0, I)$.

  $x_t \leftarrow \sqrt{\bar{\alpha}_t}\, x_0^{(n)} + \sqrt{1-\bar{\alpha}_t}\, \epsilon$

  $\mathcal{L}_{\text{pur}} \leftarrow \left\|\epsilon - \epsilon_\theta(x_t, t, p^\star)\right\|_2^2 \; + \; \lambda\, \mathcal{R}\Big(x_0^{(n)}, x^{adv}\Big)$

  $g \leftarrow \nabla_{x_0} \mathcal{L}_{\text{pur}}$

  $x_0^{(n+1)} \leftarrow \Pi_{[0,1]}\big(x_0^{(n)} - \eta\, g\big)$

$x^{\text{pur}} \leftarrow x_0^{(N)}$

**return** $x^{\text{pur}}$.

where $t$ is sampled uniformly from $\{1, \dots, T\}$, $\epsilon \sim \mathcal{N}(0, I)$, and $\mathcal{R}$ is an optional proximity or naturalness regularizer (e.g., $\mathcal{R}(x_0, x^{adv}) = \|x_0 - x^{adv}\|_2^2$ or a TV prior). By Theorem 1, minimizing $\mathcal{L}_{\text{DDPM}}$ w.r.t. $p$ tightens the ELBO; when optimizing w.r.t. $x_0$, Eq. (2) serves as a surrogate that increases the conditional likelihood under $p^\star$.

**Gradient and Update.** Let $x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$. The gradient of Eq. (2) is

$$\nabla_{x_0} \mathcal{L}_{\text{Pur}} = \nabla_{x_0} \mathbb{E}_{t,\epsilon}\Big[\sqrt{\bar{\alpha}_t}\, \nabla_{x_t} \left\|\epsilon - \epsilon_\theta(x_t, t, p^\star)\right\|_2^2\Big] \; + \; \lambda\, \nabla_{x_0} \mathcal{R}(x_0, x^{adv}),$$

where we define $\mathcal{L}_{\text{Pur}} = \mathcal{L}_{\text{DDPM}} \; + \; \lambda\, \nabla_{x_0} \mathcal{R}(x_0, x^{adv})$ and we perform a few iterations of gradient descent with box constraints:

$$x_0^{(k+1)} = \Pi_{[0,1]}\Big(x_0^{(k)} - \eta\, \nabla_{x_0} \mathcal{L}_{\text{Pur}}(x_0^{(k)}; p^\star)\Big), \qquad x_0^{(0)} = x^{adv},$$

where $\Pi_{[0,1]}$ clips pixels to the valid range and $\eta$ is the step size. In practice, we estimate the expectations with a single $(t, \epsilon)$ per iteration and use 5–10 steps; the prompt guidance $p^\star$ stabilizes the descent by injecting high-level semantics, yielding fast and faithful purification in pixel space. Besides, we adopt the proximity regularizer $\mathcal{R}(x_0, x^{adv}) = \|x_0 - x^{adv}\|_2^2$ with a fixed weight $\lambda = 0.9$, which encourages purified outputs to remain close to the original adversarial inputs while removing perturbations. The overall purification process is summarized in Algorithm 2.

**Theory: Stochastic One–Sample Purification and Few–Step Ascent** We analyze the pixel–space purification objective

$$\mathcal{L}_{\text{pur}} = \mathcal{L}_{\text{DDPM}}(x_0; p^\star) + \lambda \mathcal{R}(x_0, x^{adv}). \tag{3}$$

Let $x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ and denote $\ell(x_0; t, \epsilon) = \left\|\epsilon - \epsilon_\theta(z_t, t, p^\star)\right\|_2^2$. We update $x_0$ by projected SGD with one sampled $(t, \epsilon)$ per step:

$$x_0^{(k+1)} = \Pi_{[0,1]}\left(x_0^{(k)} - \eta\, g(x_0^{(k)}; t_k, \epsilon_k)\right), \quad g(x_0; t, \epsilon) =: \nabla_{x_0}\big(\ell(x_0; t, \epsilon) + \lambda \mathcal{R}(x_0, x^{adv})\big).$$

**Assumptions.** (A1) *Smoothness:* $\mathcal{L}_{\text{Pur}}(x_0; p^\star)$ is $L$–smooth on $[0,1]^d$. (A2) *Bounded variance:* $\mathbb{E}\big[\|g(x_0; t, \epsilon) - \nabla \mathcal{L}_{\text{Pur}}(x_0; p^\star)\|_2^2\big] \le \sigma^2$. (A3) *Regularizer:* $\mathcal{R}$ is convex and $L_R$–smooth (e.g., $\|x_0 - x^{adv}\|_2^2$ or TV with smooth surrogate).

**Lemma 1** (Unbiased one–sample gradient with bounded variance)**.** *With the reparameterization $z_t(x_0, \epsilon)$, the stochastic gradient is unbiased:*

$$\mathbb{E}_{t,\epsilon}\big[\, g(x_0; t, \epsilon)\,\big] = \nabla_{x_0} \mathcal{L}_{Pur}(x_0; p^\star),$$

*and satisfies Assumption (A2).*

*Proof Sketch.* Differentiate under the expectation using reparameterization; the Jacobian $\partial x_t / \partial x_0 = \sqrt{\bar{\alpha}_t} I$ is deterministic. Linearity of expectation and the uniform sampling of $t$ give the unbiasedness. Bounded variance follows from (A1) and standard Lipschitz/activation bounds on $\epsilon_\theta$. □

**Theorem 2** (Expected descent of the purification loss)**.** *Under (A1)–(A3), let $\eta \le 1/(2L)$. Then the projected SGD iterate satisfies*

$$\mathbb{E}\Big[\mathcal{L}_{Pur}(x_0^{(k+1)}; p^\star)\Big] \le \mathbb{E}\Big[\mathcal{L}_{Pur}(x_0^{(k)}; p^\star)\Big] - \frac{\eta}{2}\,\mathbb{E}\Big[\|\nabla \mathcal{L}_{Pur}(x_0^{(k)}; p^\star)\|_2^2\Big] + \frac{\eta^2 L}{2}\,\sigma^2.$$

*Consequently, after $K$ steps,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\Big[\|\nabla \mathcal{L}_{DDPM}(x_0^{(k)}; p^\star)\|_2^2\Big] \le \frac{2(\mathcal{L}_{DDPM}(x_0^{(0)}; p^\star) - \mathcal{L}_{inf})}{\eta K} + \eta L \sigma^2,$$

*where $\mathcal{L}_{inf}$ is the infimum over $[0,1]^d$.*

*Proof Sketch.* Apply the standard smoothness (descent) lemma to the projected step and take expectations. Use Lemma 1 to replace $\mathbb{E}[g]$ with the true gradient and bound the variance term by $\sigma^2$. Summing the per–step inequality yields the average–gradient bound. □

**Why Single $(t, \epsilon)$ and Few Steps Suffice.** The one–sample estimator is unbiased but noisy; this *stochasticity* serves as exploration that helps escape local pixel–level artifacts, while Theorem 2 guarantees expected descent provided $\eta$ is small. Moreover, the bound shows $O(1/K)$ decay of the average gradient norm up to a variance floor $\eta L \sigma^2$, so a *small fixed* number of steps ($N = 5$–$10$) already yields a measurable reduction of the loss/ELBO gap—matching our practice.

# 4 Experiments

## 4.1 Experimental Design

This section presents an extensive empirical study to validate the effectiveness of the proposed Multimodal Diffusion for Adversarial Purification (MultiDAP). We describe experimental setups (datasets and implementation), report quantitative and qualitative results under adversarial attacks, and provide ablations dissecting the contributions of different design choices.

**Datasets and Model Architectures.** We primarily evaluate our method on CIFAR-10, a standard benchmark comprising 60,000 $32 \times 32 \times 3$ images from 10 classes and upsample them where each image is resized to $256 \times 256$ to match the input resolution of Stable Diffusion. We use the 50,000 clean training images to optimize the class-agnostic prompt embeddings, and evaluate purification performance on a held-out subset of 512 adversarial test images sampled from the standard 10,000-image test split. We use miniSD-diffusers (Lambda Labs, 2022) as a frozen stable diffusion generative backbone for purification. By default we adopt a class-agnostic prompt, parameterized as a set of $M$ learnable context tokens that are shared across all images and classes. For ablation studies, we also evaluate hand-crafted class-agnostic prompts (e.g., "a photo of"), which provide weaker guidance compared to our learned prompt but highlight the effectiveness of explicitly optimizing context tokens. The classifiers are WideResNet70-16 which are pretrained on clean datasets.

**Implementation Details.** We initialize miniSD-diffuser and freeze all network parameters except for the learnable prompt embeddings. During prompt learning, we optimize $M = 16$ context tokens of dimension $d = 768$ using the Adam optimizer where the learning rate is $2 \times 10^{-4}$ and the batch size 64. Training is performed on the 50,000 clean CIFAR-10 training images for 10 epochs. We incorporate an expectation of the noise $\epsilon$ and timestep $t$ while training, ensuring consistency with the diffusion objective. During purification, we inject adversarial examples into the forward diffusion process with a single randomly sampled timestep $t \sim \text{Unif}([T_1, T_2])$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, where we set $T_1 = 400$ and $T_2 = 600$. We then run 5 iterations of gradient descent on $x_0$ using the Regularized DDPM loss (Eq. (3)), which provides an efficient surrogate for likelihood maximization. The step size $\eta$ is set to 0.2, and we clip pixel values into $[0, 1]$ after each update. Unless otherwise specified, we adopt the class-agnostic prompt $p^\star$ learned in Sec. 3.2 as the conditioning signal.

Table 1: Clean and robust accuracy (%) on CIFAR-10. Robust results under AutoAttack are reported for $\ell_\infty$ ($\epsilon = 8/255$) and $\ell_2$ ($\epsilon = 0.5$). The last column reports $\ell_\infty$ PGD-20 with 10 random restarts (step size $\alpha = \epsilon/4$). For DiffPure, $t_1 = 0.125$ and $t_2 = 0.1$ denote the time scales used. Our method (MultiDAP) uses a class-agnostic prompt and only 5 purification steps. We also report an ablation using the fixed template prompt "a photo of a", which is widely used in CLIP-based zero-shot classification. **Bold** denotes the best, underline the second best, and shading the third best.

| Method | Architecture | Clean Acc | AA ($\ell_\infty$) | AA ($\ell_2$) | PGD-20 |
|---|---|---|---|---|---|
| AT-DDPM-$\ell_\infty$ | WRN70-16 | 88.87 | 63.28 | 64.65 | 55.31 |
| AT-DDPM-$\ell_2$ | WRN70-16 | 93.16 | 49.41 | 81.05 | 51.47 |
| AT-EDM-$\ell_\infty$ | WRN70-16 | 93.36 | 70.90 | 69.73 | **72.96** |
| AT-EDM-$\ell_2$ | WRN70-16 | **95.90** | 53.32 | **84.77** | 55.34 |
| DiffPure ($t_1$) | UNet+WRN70-16 | 87.50 | 40.62 | 75.59 | 47.89 |
| DiffPure ($t_2$) | UNet+WRN70-16 | 90.97 | 44.53 | 72.65 | 51.89 |
| LM | UNet+WRN70-16 | 87.89 | 71.68 | 75.00 | 65.22 |
| MultiDAP ("a photo of a ·") | UNet+WRN70-16 | 93.80 | 70.29 | 74.53 | 65.00 |
| MultiDAP (prompt learning) | UNet+WRN70-16 | 94.12 | **72.38** | 76.05 | 68.21 |

Table 2: Ablation studies on CIFAR-10 under AutoAttack ($\ell_\infty$-norm $\epsilon = 8/255$, $\mathrm{n\_iter} = 100$). Left: effect of purification steps $N$ with fixed $[T_1, T_2] = (400, 600)$ and $\eta = 0.2$. Right: effect of timestep range $[T_1, T_2]$ with fixed $N = 5$ and $\eta = 0.2$.

| $N$ | $\eta$ | Clean Acc | Robust Acc |
|---|---|---|---|
| 1 | 0.2 | **95.01** | 62.31 |
| 2 | 0.2 | 94.57 | 65.24 |
| 5 | 0.2 | 94.12 | **72.38** |
| 10 | 0.2 | 91.31 | 71.88 |
| 20 | 0.2 | 92.88 | 65.12 |
| 30 | 0.2 | 88.75 | 58.44 |
| 50 | 0.2 | 85.94 | 53.75 |

| $[T_1, T_2]$ | $N$ | $\eta$ | Clean Acc | Robust Acc |
|---|---|---|---|---|
| $[100, 200]$ | 5 | 0.2 | **96.12** | 58.44 |
| $[200, 400]$ | 5 | 0.2 | 95.31 | 65.32 |
| $[400, 600]$ | 5 | 0.2 | 94.12 | **72.38** |
| $[600, 800]$ | 5 | 0.2 | 92.25 | 66.41 |
| $[800, 999]$ | 5 | 0.2 | 90.14 | 60.27 |

**Adversarial Attacks.** We evaluate robustness against two widely used adversarial attacks for diffusion-based adversarial purification on CIFAR-10. The first is PGD-20 (Madry, 2017), a multi-step iterative $\ell_\infty$-bounded attack with random restarts. The second is stronger adversarial attacks: AutoAttack (Croce and Hein, 2020), a parameter-free ensemble of diverse attacks. Because of the stochasticity in MultiDAP, we use rand version of AutoAttacks. Following Chen et al. (2024a)'s settings, we have $\mathrm{n\_iter} = 100$ for AutoAttack. Unless otherwise specified, the maximum perturbation budget is set to $\epsilon = 8/255$ for $\ell_\infty$ threat models and $\epsilon = 0.5$ for $\ell_2$ threat models.

**Baselines.** We compare our method with two representative purification-based defenses. The first is DiffPure (Nie et al., 2022), which performs iterative reverse diffusion to remove adversarial perturbations. The second is Likelihood Maximization (LM) (Chen et al., 2024b) which formulates purification as direct maximization of the unimodal diffusion model likelihood. These methods provide strong and conceptually related baselines for evaluating the effectiveness of our proposed MultiDAP. Other baseline methods are adversarial training methods which rely on unimodal diffusion models to generate adversarial examples or argument data.

### 4.2 Robustness against Adversarial Attacks

Table 1 summarizes the robustness results of our method compared with DiffPure and Likelihood Maximization under PGD-20 and AutoAttack. Our proposed MultiDAP achieves substantially better performance than Likelihood Maximization across both $\ell_\infty$ and $\ell_2$ threat models, highlighting the benefit of leveraging a learned prompt prior. Compared with DiffPure, MultiDAP attains slightly higher robust accuracy, but requires only 5 purification steps, whereas DiffPure typically involves dozens of Langevin iterations. This efficiency gap makes MultiDAP significantly more practical in scenarios where the inference speed is critical, while still preserving competitive robustness.

While prompt engineering relies on manually crafted textual templates (e.g., "a photo of a ·"), our approach learns task-adaptive prompt embeddings that encode richer semantic priors. This learned representation provides stronger and more flexible conditioning, enabling the diffusion model to better suppress adversarial noise and reconstruct class-consistent images even under severe perturbations. Compared to hand-designed prompts, prompt learning thus offers a systematic and scalable way to inject semantic guidance into the purification process.

From a theoretical perspective, the robustness of MultiDAP can be explained by the stochastic nature of purification combined with the learned semantic prior. Instead of explicitly computing the expectation over all timesteps and noises, which would be computationally prohibitive, we approximate it with a single random $(t, \epsilon)$ per iteration. This stochastic approximation, when combined with semantic guidance from prompts, provides both efficiency and regularization, preventing overfitting to specific perturbations. In practice, this explains why only 5–10 purification steps are sufficient to achieve competitive robustness, which is consistent with our ablation results showing that excessive steps may even degrade performance (see Table 2).

### 4.3 Ablation Studies

**Number of Purification Steps.** We first study the effect of the number of purification steps $N$ while keeping the timestep range fixed at $[400, 600]$ and step size $\eta = 0.2$. As shown in Table 2 (left), increasing $N$ from 1 to 5 steadily improves robustness, with the best performance observed at $N = 5$. Using 10 steps achieves comparable accuracy, but going beyond 10 steps (e.g., 20, 30, 50) leads to diminishing or even negative returns. In particular, excessive purification may overfit the injected noise, causing a degradation in both clean and robust accuracy. This observation highlights that a small number of purification steps (around 5–10) is sufficient for effective adversarial denoising, striking a good balance between robustness and efficiency.

**Choice of Timestep Range.** We next analyze the effect of varying the timestep interval $[T_1, T_2]$ during purification, while fixing the purification steps $N = 5$ and step size $\eta = 0.2$. As shown in Table 2(right), choosing too small a range (e.g., $[100, 200]$) fails to inject sufficient noise, causing the purification process to underperform in terms of robustness. Conversely, excessively large ranges (e.g., $[800, 999]$) introduce too much noise, which harms clean accuracy due to over-smoothed reconstructions. The best performance is observed in the mid-range (e.g., $[400, 600]$), which strikes a balance between removing adversarial perturbations and preserving semantic fidelity. This finding highlights that an appropriate noise level is crucial for effective adversarial denoising.

## 5 Conclusions and Discussions

We proposed Multimodal Diffusion for Adversarial Purification (MultiDAP), a novel adversarial defense that leverages text-to-image diffusion models guided by learnable prompts. Experiments on CIFAR-10 demonstrate that MultiDAP achieves competitive robustness against strong white-box attacks, outperforming likelihood maximization baselines while being significantly more efficient than DiffPure. Our ablation studies further highlight the importance of prompt learning and careful timestep design, showing that semantic priors play a crucial role in improving purification quality.

**Limitations and Future Work.** Despite promising results, MultiDAP currently depends on Stable Diffusion backbones, which are computationally heavier than standard feed-forward models. Future work may explore lightweight diffusion architectures or truncated sampling to further improve efficiency. Moreover, extending MultiDAP to large-scale datasets (e.g., ImageNet) and domain-shifted benchmarks would provide stronger evidence of generalization. Another exciting direction is to incorporate richer or multi-modal prompts (e.g., textual descriptions beyond class names) to enhance semantic guidance during purification.

## Ethics Statement

This research employs computational approaches exclusively with publicly accessible datasets, avoiding any human subject involvement or confidential data handling. We adhere to ICLR's ethical

guidelines without competing interests, emphasizing responsible deployment of our contributions while ensuring transparent reporting to support reproducible research practices.

## Reproducibility Statement

Our experiments utilize publicly available datasets with detailed experimental configurations provided throughout the paper. To facilitate reproducibility, we commit to releasing the complete source code upon paper acceptance, enabling the research community to validate and build upon our findings.

## References

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervision. *arXiv preprint arXiv:2101.09387*, 2021.

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

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16805–16827, 2022.

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

Mingyuan Bai, Wei Huang, Tenghui Li, Andong Wang, Junbin Gao, César Federico Caiafa, and Qibin Zhao. Diffusion models demand contrastive guidance for adversarial purification to advance. 2024.

Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24331–24340, 2025.

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

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

Hadi M Dolatabadi, Sarah Erfani, and Christopher Leckie. $\ell_\infty$-robustness and beyond: Unleashing efficient adversarial training. In *European Conference on Computer Vision*, pages 467–483, 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*, pages 36246–36263. PMLR, 2023a.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

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*, ICML'24. JMLR.org, 2024a.

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.

Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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*, pages 36246–36263. PMLR, 2023b.

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

Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models towards adversarial robustness. *arXiv preprint arXiv:2211.00322*, 2022.

Cheng-Han Yeh, Kuanchun Yu, and Chun-Shien Lu. Test-time adversarial defense with opposite adversarial path and high attack time cost. *arXiv preprint arXiv:2410.16805*, 2024.

Roland S Zimmermann, Lukas Schott, Yang Song, Benjamin A Dunn, and David A Klindt. Score-based generative classifiers. *arXiv preprint arXiv:2110.00473*, 2021.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36, 2023.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 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.

Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision*, pages 56–72. Springer, 2024.

Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24408–24419, 2024.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29958–29967, 2025.

Zeyu Wang, Cihang Xie, Brian Bartoldson, and Bhavya Kailkhura. Double visual defense: Adversarial pre-training and instruction tuning for improving vision-language model robustness. *arXiv preprint arXiv:2501.09446*, 2025.

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.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9):2337–2348, September 2022. ISSN 1573-1405.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lambda Labs. miniSD-diffusers: A Text-to-Image Model based on Stable Diffusion, 2022. URL https://huggingface.co/lambdalabs/miniSD-diffusers.

Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

## USE OF LARGE LANGUAGE MODELS (LLMS)

LLMs served as supplementary instruments for textual improvement and code standardization throughout this research. Their application was limited to enhancing readability, correcting grammatical inconsistencies, and maintaining uniform presentation standards for pseudocode blocks and LaTeX expressions. The core intellectual contributions—including research conception, methodological frameworks, analytical reasoning, and experimental findings—remain entirely human-generated.

## A COMPLETE PROOFS

### A.1 PROOF OF THEOREM 1

*Proof of Theorem 1.* For DDPM (or latent DDPM), the negative ELBO can be written (Ho et al., 2020; Nichol and Dhariwal, 2021) as

$$-\log \underline{p}_\theta(z_0 \mid p) = C(\theta) + \sum_{t=1}^{T} \underbrace{\mathbb{E}\big[\mathrm{KL}\big(q(z_{t-1} \mid z_t, z_0) \,\|\, p_\theta(z_{t-1} \mid z_t, p)\big)\big]}_{=:\,\mathcal{R}_t(p)} + \underbrace{\mathbb{E}\big[\mathrm{KL}\big(q(z_T) \,\|\, p(z_T)\big)\big]}_{\text{independent of } p}, \tag{4}$$

where $C(\theta)$ is independent of $p$ and only $\mathcal{R}_t(p)$ depends on the prompt.

Using the standard mean parameterization,

$$\mu_\theta(x_t, t, p) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t, p)\Big), \quad \Sigma_\theta(x_t, t) = \sigma_t^2 I,$$

one can show that $\mathcal{R}_t(p)$ is equivalent to a weighted noise-prediction MSE:

$$\mathcal{R}_t(p) = w_t\, \mathbb{E}_{x_0,t,\epsilon}\Big[\big\|\epsilon - \epsilon_\theta(x_t, t, p)\big\|_2^2\Big] + \text{const}, \quad w_t = \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} > 0.$$

Substituting into Eq. (4) yields

$$\mathcal{L}_{\mathrm{VLB}}(p) =: -\log \underline{p}_\theta(x_0 \mid p) = C'(\theta) + \sum_{t=1}^{T} w_t\, \mathbb{E}_{x_0,t,\epsilon}\Big[\big\|\epsilon - \epsilon_\theta(x_t, t, p)\big\|_2^2\Big],$$

which differs from $\mathcal{L}_{\mathrm{prompt}}(p)$ only by positive weights and a constant. Therefore,

$$\arg\min_p \mathcal{L}_{\mathrm{prompt}}(p) = \arg\min_p \mathcal{L}_{\mathrm{VLB}}(p).$$

Let $p^\star$ be the minimizer; then $\mathcal{L}_{\mathrm{VLB}}(p^\star) \le \mathcal{L}_{\mathrm{VLB}}(p)$ for all $p$, equivalently $\log \underline{p}_\theta(x_0 \mid p^\star) \ge \log \underline{p}_\theta(x_0 \mid p)$. □

### A.2 PROOF OF LEMMA 1

*Proof of Lemma 1.* We first recall the purification objective:

$$\mathcal{L}_{\text{pur}}(x_0; p^\star) = \mathbb{E}_{t,\epsilon}\big[\ell(x_0; t, \epsilon)\big] + \lambda \mathcal{R}(x_0, x^{adv}),$$

where $t \sim \mathrm{Unif}(\{1, \dots, T\})$ and $\epsilon \sim \mathcal{N}(0, I)$.

By linearity of expectation and interchangeability of expectation and gradient under standard regularity conditions:

$$\nabla_{x_0} \mathcal{L}_{\text{pur}}(x_0; p^\star) = \nabla_{x_0}\, \mathbb{E}_{t,\epsilon}\big[\ell(x_0; t, \epsilon)\big] + \lambda \nabla_{x_0} \mathcal{R}(x_0, x^{adv}).$$

On the other hand, the stochastic gradient $g(x_0; t, \epsilon)$ is defined as

$$g(x_0; t, \epsilon) = \nabla_{x_0} \ell(x_0; t, \epsilon) + \lambda \nabla_{x_0} \mathcal{R}(x_0, x^{adv}).$$

Taking expectation over $(t, \epsilon)$ gives:

$$\mathbb{E}_{t,\epsilon}[g(x_0; t, \epsilon)] = \mathbb{E}_{t,\epsilon}[\nabla_{x_0} \ell(x_0; t, \epsilon)] + \lambda \nabla_{x_0} \mathcal{R}(x_0, x^{adv}),$$

which equals $\nabla_{x_0}\mathcal{L}_{\text{pur}}(x_0;p^\star)$. Thus the estimator is unbiased.

By Assumption (A2), we assume that the variance of the stochastic gradient is bounded:

$$\mathbb{E}_{t,\epsilon}\big[\|g(x_0;t,\epsilon)-\nabla_{x_0}\mathcal{L}_{\text{pur}}(x_0;p^\star)\|_2^2\big]\le\sigma^2.$$

This holds because $\ell(x_0;t,\epsilon)$ is quadratic in $\epsilon$ and $\epsilon\sim\mathcal{N}(0,I)$ has bounded second moment, while $\mathcal{R}$ is convex and $L_{\mathcal{R}}$–smooth (Assumption A3). Therefore the stochastic gradient inherits bounded variance.

Together, we conclude that $g(x_0;t,\epsilon)$ is an unbiased stochastic gradient estimator of $\nabla_{x_0}\mathcal{L}_{\text{pur}}(x_0;p^\star)$ with bounded variance, completing the proof. □

### A.3 Proof of Theorem 2

*Proof of Theorem 2.* Recall the purification objective

$$\mathcal{L}_{\text{pur}}(x_0;p^\star)=\mathbb{E}_{t,\epsilon}\big[\ell(x_0;t,\epsilon)\big]+\lambda\mathcal{R}(x_0,x^{adv}),\qquad \ell(x_0;t,\epsilon)=\big\|\epsilon-\epsilon_\theta(z_t,t,p^\star)\big\|_2^2,$$

and the projected update on the pixel cube $\mathcal{C}=[0,1]^d$:

$$x_0^{(k+1)}=\Pi_{\mathcal{C}}\big(x_0^{(k)}-\eta\,g(x_0^{(k)};t_k,\epsilon_k)\big),\qquad g(x_0;t,\epsilon)=\nabla_{x_0}\ell(x_0;t,\epsilon)+\lambda\nabla_{x_0}\mathcal{R}(x_0,x^{adv}).$$

Since $\mathcal{L}_{\text{pur}}$ is $L$-smooth on $\mathcal{C}$ (Assumption A1), the standard smoothness inequality gives

$$\mathcal{L}_{\text{pur}}(x_0^{(k+1)};p^\star)\le\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)+\big\langle\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star),\,x_0^{(k+1)}-x_0^{(k)}\big\rangle+\frac{L}{2}\big\|x_0^{(k+1)}-x_0^{(k)}\big\|_2^2.\quad(5)$$

The projection onto a closed convex set is nonexpansive and satisfies $\|x_0^{(k+1)}-x_0^{(k)}\|_2\le\eta\,\|g(x_0^{(k)};t_k,\epsilon_k)\|_2$. Hence, from Eq. (5),

$$\mathcal{L}_{\text{pur}}(x_0^{(k+1)};p^\star)\le\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)-\eta\,\big\langle\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star),\,g(x_0^{(k)};t_k,\epsilon_k)\big\rangle+\frac{L\eta^2}{2}\big\|g(x_0^{(k)};t_k,\epsilon_k)\big\|_2^2.$$

Conditioning on $x_0^{(k)}$ and using Lemma 1,

$$\mathbb{E}_{t_k,\epsilon_k}\big[g(x_0^{(k)};t_k,\epsilon_k)\,\big|\,x_0^{(k)}\big]=\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star),$$
$$\mathbb{E}_{t_k,\epsilon_k}\big[\|g(x_0^{(k)};t_k,\epsilon_k)\|_2^2\,\big|\,x_0^{(k)}\big]\le\|\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\|_2^2+\sigma^2.$$

Using the standard variance decomposition bound $\mathbb{E}\|g\|_2^2\le 2\|\nabla\mathcal{L}_{\text{pur}}\|_2^2+2\sigma^2$ (or equivalently absorbing constants into $\sigma^2$) and taking full expectation yields

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}_{\text{pur}}(x_0^{(k+1)};p^\star)\big]&\le\mathbb{E}\big[\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\big]-\eta\,\mathbb{E}\big[\|\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\|_2^2\big]+\frac{L\eta^2}{2}\,\mathbb{E}\big[\|g(x_0^{(k)};t_k,\epsilon_k)\|_2^2\big]\\&\le\mathbb{E}\big[\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\big]-\Big(\eta-\tfrac{L\eta^2}{2}\Big)\,\mathbb{E}\big[\|\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\|_2^2\big]+\frac{L\eta^2}{2}\,\sigma^2.\end{aligned}$$

If $\eta\le 1/(2L)$, then $\eta-\frac{L\eta^2}{2}\ge\eta/2$. Hence we obtain the claimed one–step descent:

$$\mathbb{E}\Big[\mathcal{L}_{\text{pur}}(x_0^{(k+1)};p^\star)\Big]\le\mathbb{E}\Big[\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\Big]-\frac{\eta}{2}\,\mathbb{E}\Big[\|\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\|_2^2\Big]+\frac{\eta^2L}{2}\,\sigma^2.$$

Summing the inequality over $k=0,\dots,K-1$ and using the lower bound $\mathcal{L}_{\text{pur}}(x_0;p^\star)\ge\mathcal{L}_{\text{inf}}:=\inf_{x\in[0,1]^d}\mathcal{L}_{\text{pur}}(x;p^\star)$, we get

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\|\nabla\mathcal{L}_{\text{pur}}(x_0^{(k)};p^\star)\|_2^2\Big]\le\frac{2\big(\mathcal{L}_{\text{pur}}(x_0^{(0)};p^\star)-\mathcal{L}_{\text{inf}}\big)}{\eta K}+\eta L\,\sigma^2.$$

Since $\mathcal{L}_{\text{pur}}=\mathcal{L}_{\text{DDPM}}+\lambda\mathcal{R}$ and $\mathcal{R}$ is convex and smooth (Assumption A3), the same derivation applies when focusing on the data-fidelity part. In particular, using $\|\nabla\mathcal{L}_{\text{DDPM}}(x)\|_2^2\le\|\nabla\mathcal{L}_{\text{pur}}(x)\|_2^2$ (up to a constant absorbed into $\sigma^2$) yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\big\|\nabla\mathcal{L}_{\text{DDPM}}(x_0^{(k)};p^\star)\big\|_2^2\Big]\le\frac{2\big(\mathcal{L}_{\text{DDPM}}(x_0^{(0)};p^\star)-\mathcal{L}_{\text{inf}}\big)}{\eta K}+\eta L\,\sigma^2,$$

which is the stated bound. □