SleeperDiff: Generating Transform-Dependent Attacks via Diffusion Models

Gavin Gleinig
Texas A&M University
gavin.gleinig@tamu.edu

Vijay Murugan
Texas A&M University
vijay.muruganas@tamu.edu

Shristi Nadakatti
Texas A&M University
shristimn@tamu.edu

***Abstract*—Modern vision models remain highly sensitive to small changes in image presentation, making them vulnerable to adversarial examples that activate only under specific transformations such as scaling or blurring. Existing transform-dependent attacks typically rely on pixel-space perturbations, which can introduce visible artifacts and fail to generalize across transformation ranges. To address these limitations, we introduce *SleeperDiff*, a diffusion-based framework that embeds adversarial triggers directly within the latent space of a pre-trained generative model. Our method combines DDIM inversion with a dual-objective optimization that enforces both benign preservation and targeted misclassification under transformation, while Expectation over Transformations (EoT) ensures robustness across continuous transformation intervals. Through our experiments, we were able to demonstrate that SleeperDiff produces highly imperceptible adversarial examples, achieving attack success rates of 59%–72% across different transformations. These results highlight the potential of diffusion models for generating transformation-aware adversarial attacks with improved visual quality, robustness, and controllability. Our code is available at: https://github.com/gavingleinig/semantic-transformation-adv-attack**

***Index Terms*—adversarial examples, machine learning**

I. Introduction

We are currently witnessing a massive wave of AI investment and research interest, often characterized as the third golden summer of artificial intelligence. The capabilities of modern models are rapidly expanding; for instance, Large Language Models (LLMs) can now pass math olympiad tests and conduct autonomous research [1], [2]. Simultaneously, generative AI has captured the public imagination, allowing users to create content from simple prompts, while diffusion models generate natural-looking images that align closely with human perception. These advancements are primarily driven by the rise of improved model architectures like the transformer, the vast availability of training data on the internet, and the accessibility of fast, highly parallel computation hardware.

As this technology matures, it is increasingly applied in critical, real-world scenarios. For example, Waymo has deployed autonomous taxi services to the public in cities like Phoenix, San Francisco, Los Angeles, and Austin (soon coming to Dallas in 2026). However, the deployment of deep learning in safety-critical environments brings associated risks. Deep learning models are notoriously sensitive to how their input is presented, lacking the robustness required for reliability. Specifically, image models are susceptible to adversarial examples, which are samples fed into a network that are specifically crafted to trick the model into misclassifying them [5].

While conventional adversarial attacks utilize static perturbations to induce fixed mispredictions, recent research has introduced "transform-dependent" adversarial attacks [16]. In this paradigm, an input remains benign under normal conditions but triggers a targeted misclassification only when specific transformations such as zooming or blurring are applied. However, current methods often generate these attacks by manipulating the image pixel space directly. Despite mathematical constraints (such as the L_p-norm), pixel-based perturbations often result in high-frequency noise that remains perceptible to the human eye.

In this work, we propose a novel approach to transform-dependent attacks by leveraging the generative and discriminative power of diffusion models. Instead of direct manipulation in pixel space, we craft perturbations in the latent space of a pre-trained diffusion model. By combining the dynamic controllability of transform-dependent attacks with the semantic consistency of diffusion models, we aim to improve visual imperceptibility while maintaining attack effectiveness across different transformations.

II. Related Work

A. Adversarial Examples

Deep neural networks (DNNs) are highly vulnerable to adversarial attacks, where imperceptibly small perturbations are added to an input to induce misclassification. Traditional attacks, such as FGSM, PGD, and Carlini & Wagner, typically generate adversarial examples by adhering to a strict constraint $d(x_0, x_A) \leq \epsilon$ [6]–[8]. This distance metric d is usually an L_p-norm ($\|x_0 - x_A\|_p$), restricting perturbations to small changes in the raw pixel space. However, recent research argues that L_p-norm constraints are often insufficient for preserving visual quality; they can result in high-frequency noise that is perceptible to human observers [14].

To address this, researchers have proposed "Unrestricted Adversarial Examples," a threat model where attacks are not limited by small norm bounds but rather by the retention of semantic meaning [9], [10]. While early methods explored spatial transformations or adversarial patches, these often result in obvious visual artifacts. Generative models have recently been employed to craft unrestricted attacks that are

more natural-looking [11], though achieving transferability with these methods remains a challenge.

B. Diffusion Models

Diffusion Denoising Probabilistic Models (DDPMs) have emerged as a powerful class of generative models, capable of producing higher-quality and more diverse images than GANs [12], [13]. These models operate via a forward process that gradually adds Gaussian noise to an image, and a reverse process where a neural network (typically a U-Net) predicts and removes this noise to reconstruct the data.

Crucially for adversarial attacks, the iterative denoising process inherent to diffusion models naturally filters out high-frequency noise, resulting in perturbations that align better with human perception. Furthermore, pre-trained diffusion models exhibit strong discriminative capabilities. Recent work, such as DiffAttack [14] and Diff-PGD [15], exploits these properties to generate imperceptible and transferable attacks by utilizing the diffusion process rather than simple additive noise.

C. Image Transformation in Adversarial Attacks

Image transformations in adversarial learning have primarily been studied in the context of robustness. The "Expectation over Transformations" (EOT) framework was originally introduced to synthesize physically robust adversarial examples [3]. By optimizing perturbations over a distribution of transformations (e.g., rotation, scaling), EOT ensures that an attack remains effective regardless of the input's specific presentation.

More recently, research has pivoted toward Transform-Dependent Adversarial Attacks. Unlike traditional robust attacks which seek invariance, these methods are designed to be benign under normal conditions but trigger a targeted misclassification only when specific transformations are applied [16]. EOT remains a critical component in this domain; rather than using EOT to achieve global invariance, transform-dependent methods adopt it to ensure local consistency. By smoothing the attack over a small distribution of transformation parameters, EOT ensures the adversarial trigger remains reliable despite minor variations in the applied transformation.

D. Adversarial Defense

Recent advancements in generative adversarial networks (GANs) for image generation and image editing has raised concerns of potential misuse of personal data for creating fake images. Significant research has been put in to discovering innovative ways of mitigating this manipulation, and various defense strategies have been developed to prevent adversarial attacks. However, the advent of publicly available models like Stable Diffusion and DALL·E 2 has enabled users to easily describe their desired image manipulation and achieves realistic results that outperform the GAN-based attacks.

Another line of work addresses data misuse during model training rather than after deployment, which includes unlearnable examples that add imperceptible backdoor signals to user data before uploading online to prevent models from exploiting this data during training. However, research demonstrated that these training-time protection methods can be circumvented by subsequent models that can avoid being fooled. This motivated the development of image immunization, which offers both backward compatibility with existing models and the potential for forward compatibility. Photoguard [17] is an approach to protect images by "immunizing" them by adding imperceptible adversarial perturbations that disrupt the diffusion model's operation, forcing it to generate unrealistic or irrelevant outputs when someone attempts to edit the protected image.

III. Method

Our approach synthesizes the transform-dependent capability of Transform-Dependent Adversarial Attacks with the imperceptible generative capabilities of Diffusion Models. Unlike standard adversarial attacks that add pixel-space noise, we optimize the latent representation of a pre-trained diffusion model. This allows us to generate a single adversarial example that is visually faithful to the original image but is misclassified into specific target labels when subjected to specific transformations.

A. Threat Model

We operate under a white-box assumption where the adversary has full access to the target classifier. Unlike traditional attacks constrained by pixel-level norms, our adversary optimizes within the continuous latent space, bounded only by the generative model's semantic priors to ensure naturalness. This is a targeted attack; the objective is to force the model to predict a specific, adversary-chosen label (y_{adv}) whenever the input undergoes transformations like scaling or blurring, while retaining the correct label (y_{benign}) otherwise.

B. DDIM Inversion and Initialization

To ensure the adversarial optimization begins from a semantically meaningful state, we utilize a deterministic Denoising Diffusion Implicit Models (DDIM) inversion [4]. Given a clean input image x_{clean}, DDIM Inversion reverses the deterministic sampling process to map the image back to its latent representation z_t at a specific timestep t.

Let Inverse$(\cdot)$ denote the inversion operation and $D(\cdot)$ denote the forward denoising (reconstruction) process. We initialize our adversarial latent z_t such that:

$$z_t = \text{Inverse}(x_{clean}, t) \tag{1}$$

Consistent with DiffAttack, we also optimize the unconditional text embeddings to enhance reconstruction quality.

C. Dual-Objective Optimization

We seek a latent z that satisfies two conflicting objectives: maintaining the benign classification y_{clean} under the identity transform, while triggering the adversarial target y_{adv} under any transformation θ within a target range Θ.

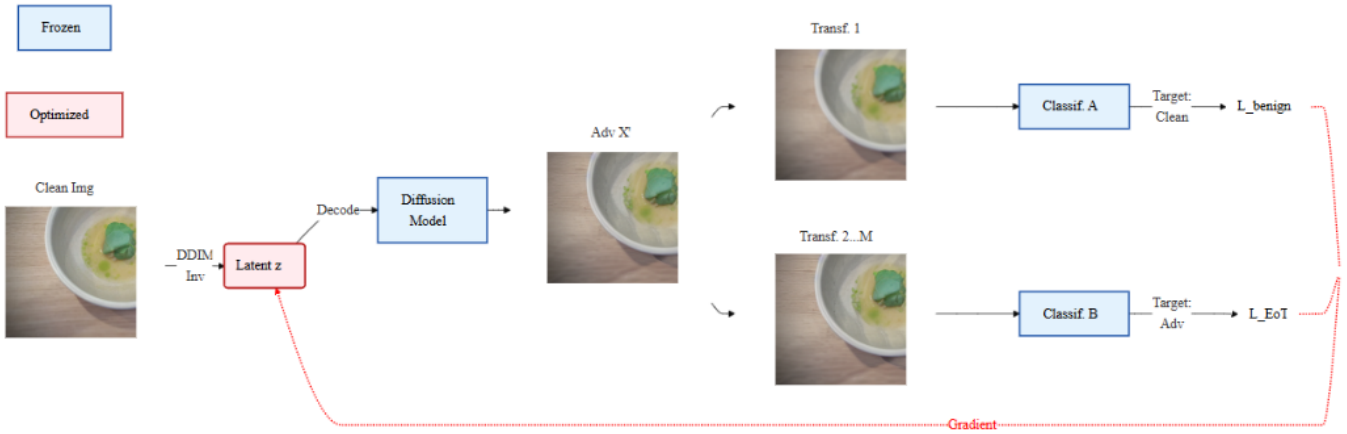

Fig. 1. **Overview of the proposed optimization framework.** The latent z is initialized via DDIM inversion of the clean image. The optimization splits into two branches: the benign branch (top) minimizes $\mathcal{L}_{clean}$ to preserve the original class under identity, while the adversarial branch (bottom) minimizes $\mathcal{L}_{EoT}$ to trigger the target class y_{adv} across a distribution of transformations. Gradients from both branches update the latent z through the frozen diffusion model.

$$\forall \theta \in \Theta : \quad \arg\max f(T(x_{\text{adv}}; \theta)) = y_{\text{adv}}, \qquad (2)$$
$$\arg\max f(T(x_{\text{adv}}; \theta_{\text{benign}})) = y_{\text{benign}}.$$

We formulate this as a joint optimization problem comprising two primary loss terms.

1) Benign Preservation: To ensure stealth, the image must retain its original classification when no transformation is applied. We minimize the loss of the reconstructed image $D(z_t)$ against the original label y_{benign} under the identity transform:

$$\mathcal{L}_{benign} = \ell\left(f(D(z_t)), y_{benign}\right) \qquad (3)$$

where $f(\cdot)$ is the classifier and $\ell(\cdot)$ is the loss function. We investigate both standard Cross-Entropy and the Carlini-Wagner margin loss, which optimizes the margin between the target class and the next most likely class to enhance robustness [8].

2) Robust Triggering via EoT: A naive attack optimized for a single transformation parameter (e.g., scale $s = 0.5$) is often brittle, failing if the actual transformation varies slightly (e.g., $s = 0.51$). To create a robust trigger, we employ the Expectation over Transformation (EoT) framework [3]. We define a distribution of transformation parameters $\mathcal{T}$ where $\theta \sim p(\theta)$.

Since the exact expectation is intractable, we approximate it via Monte Carlo sampling. At each optimization step, we draw a batch of M parameters using a center-point jittering strategy to ensure coverage of the attack range. The robust adversarial loss is defined as:

$$\mathcal{L}_{\text{EoT}} \approx \frac{1}{M} \sum_{i=1}^{M} \ell\big(f\big(T(D(z); \theta_i)\big), y_{\text{adv}}\big)\,, \qquad (4)$$
$$\theta_i \overset{\text{iid}}{\sim} p(\theta).$$

This stochastic optimization provides two critical benefits. First, it facilitates margin smoothing; by minimizing the loss across a neighborhood of parameters rather than a single point, we implicitly smooth the decision boundary, making the attack tolerant to quantization and measurement noise. Second, it ensures coverage, preventing the optimizer from exploiting narrow, brittle failure modes at specific parameters (overfitting) and forcing the generation of perturbations that generalize across the continuous range.

To enable end-to-end gradient propagation, the transformation functions $T(x; \theta)$ must be both deterministic and differentiable. We implement three specific transformations. For **scaling**, We resize the image tensor using differentiable bilinear interpolation. For a scaling factor S, the output dimensions are scaled by S. For **blurring**, we simulate an optical defocus using Gaussian blurring. The transform applies a Gaussian kernel with a fixed size of 5×5, parameterized by the standard deviation σ. For **gamma**, we apply non-linear brightness adjustment.

D. Content Preservation and Regularization

To maintain visual fidelity and prevent the optimization from producing unnatural artifacts, we incorporate two regularization terms inspired by DiffAttack [14].

1) Structure Preservation: To preserve the geometric layout of the original image, we enforce similarity between the self-attention maps of the optimized latent and the fixed original latent. Let S_t and $S_{t(fix)}$ denote the self-attention maps of the optimizing latent z_t and the fixed initial latent, respectively. The loss is calculated as:

$$\mathcal{L}_{structure} = \|S_t - S_{t(fix)}\|_2^2 \qquad (5)$$

2) Implicit Transferability: While our primary focus is white-box attack, we retain the transferability loss to distract the internal attention of the diffusion model, hypothesized to facilitate the generation of more robust semantic perturbations.

This objective minimizes the variance of the averaged cross-attention maps between the image latents and the text condition C:

$$\mathcal{L}_{transfer} = \text{Var}\left(\text{AvgPool}\left(\text{CrossAttn}(z_t, C)\right)\right) \quad (6)$$

E. Overall Objective

The final adversarial latent z_t^* is obtained by minimizing the weighted sum of the stealth, attack, and regularization objectives:

$$z_t^* = \min_{z_t}\Big(\lambda_b\mathcal{L}_{benign} + \lambda_r\mathcal{L}_{EoT} + \lambda_s\mathcal{L}_{struct} + \lambda_t\mathcal{L}_{trans}\Big) \quad (7)$$

where λ terms are hyperparameters controlling the trade-off between benign accuracy, attack robustness, and visual quality.

The optimization is performed iteratively in the latent space of the frozen diffusion model. We initialize the latent vector z_0 using the deterministic DDIM inversion of the clean input x_{clean}. The optimization proceeds for K iterations; at each step k, we decode the current latent z_k via the differentiable decoder $\mathcal{D}$ to reconstruct the intermediate image. We then approximate the Expectation over Transformation (EoT) by sampling a batch of M transformation parameters $\{\theta_i\}_{i=1}^{M}$ centered around the fixed anchor points defined in our attack range. Gradients from the weighted total objective are back-propagated through the differentiable transformation functions and the decoder to update z_k. We utilize the Adam optimizer to minimize the loss, yielding the final adversarial latent z^*, which decodes to the robust adversarial example x_{adv}.

IV. Experiments

To evaluate our proposed method, we conducted experiments across three dimensions. First, we assess the fundamental feasibility of the attack, measuring white-box success rates, perceptual imperceptibility, and cross-model transferability (Section IV-A). Next, we investigate the resilience of our method against active adversarial defenses, specifically testing its ability to bypass image immunization techniques (Section IV-B). Finally, we analyze the stability of our Expectation over Transformation (EoT) framework, comparing the impact of different loss functions on robustness across continuous parameter ranges (Section IV-C).

We conducted evaluations on a Google Colab setup with an A100 GPU. The average generation time was approximately 2.5 minutes per adversarial example.

For comparative analysis, results for baseline methods (e.g., Tan et al. and DiffAttack) were sourced directly from their respective original publications.

A. Performance and Perceptual Quality

We conduct an initial evaluation to assess the feasibility of embedding transform-dependent triggers directly into the latent space of a diffusion model. This experiment focused on a small-scale test set ($N = 100$) from the ImageNet-Compatible dataset.

1) Experimental Setup: The adversarial optimization was performed using Inception-v3 as the white box / surrogate model. We utilized the pre-trained stable-diffusion-v1-5 as the frozen diffusion backbone. The optimization process ran for 30 iterations with a guidance scale of 2.5. We split the diffusion scheduler into $t = 20$ steps and optimized starting from step 15.

To balance the objectives of stealth and attack success, we used the following loss weights: $\lambda_{benign} = \lambda_{EoT} = 10$ (using Cross-Entropy), $\lambda_{transfer} = 10000$, and $\lambda_{structure} = 100$. We evaluated three distinct transformations: Scaling ($0.5\times$), Gamma Correction ($\gamma = 0.5$), and Gaussian Blurring ($\sigma = 1.5$)

TABLE I
White-box Performance. Benign Consistency (Clean) and Attack Success (Adv) on the source model (Inception).

Method	Scaling		Gamma		Blurring	
	Clean	Adv	Clean	Adv	Clean	Adv
Ours	97	5	100	59	100	72
Tan et al.	-	61.9	-	75.4	-	93.1

2) White-Box Performance: Table I summarizes the performance on the source Inception model. We observed a significant disparity in attack success across the different transformation types. The Blurring and Gamma transformations achieved moderate Attack Success Rates (ASR) of 72.00% and 59.00%, respectively, while maintaining perfect benign consistency (100.00%). This shows that the latent optimization successfully embedded triggers for these transformations while preserving the original class identity under normal viewing conditions.

However, the Scaling attack proved difficult to optimize in the latent space, achieving a negligible ASR of 5.00% despite a high benign accuracy of 97.00%. Compared to prior pixel-space methods like Tan et al. [16], which achieved 61.9% ASR on scaling, our latent-based approach struggled to find robust features that persist after significant downsampling.

3) Black-Box Transferability: We evaluated the cross-model transferability of the generated examples against 14 models, including CNN, Tranformer, and MLP architectures. As detailed in Table IV-A3, the perturbations exhibited negligible transferability. The ASR remained at 0.00% for nearly all target models across all three transformations. Minor exceptions were observed only in adversarial ensembles (e.g., Ens4 Adv Inc v3), which reached a maximal ASR of 2.00% under the blurring transformation. This suggests that the perturbations are highly overfitted to the decision boundary of the source Inception model and do not generalize to the learned features of distinct architectures.

4) Perceptual Quality: To quantify the stealthiness of our attacks, we computed the Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID). The results are presented in Table III.

Our method achieved an average LPIPS score of 0.132, which is comparable to state-of-the-art diffusion-based attacks

TABLE II
PERFORMANCE COMPARISON OF TARGET MODELS UNDER SCALING, GAMMA, AND BLUR TRANSFORMATIONS.

	Scaling		Gamma		Blur	
Target Model	Clean	Adv	Clean	Adv	Clean	Adv
ResNet	89.00	0.00	84.00	0.00	89.00	0.00
VGG	79.00	0.00	85.00	0.00	76.00	0.00
Mobile	79.00	0.00	74.00	0.00	71.00	0.00
Inception	**97.00**	**5.00**	**100.00**	**59.00**	**100.00**	**72.00**
ConvNext	93.00	0.00	96.00	0.00	92.00	0.00
ViT	87.00	0.00	89.00	0.00	87.00	0.00
Swin	87.00	0.00	94.00	0.00	95.00	0.00
DeiT-B	86.00	0.00	89.00	0.00	93.00	0.00
DeiT-S	85.00	0.00	87.00	0.00	92.00	0.00
Mixer-B	72.00	0.00	75.00	0.00	74.00	0.00
Mixer-L	67.00	0.00	73.00	0.00	68.00	0.00
Adv Inception v3	79.00	0.00	75.00	0.00	77.00	2.00
Ens3 Adv Inc v3	62.00	0.00	66.00	0.00	60.00	0.00
Ens4 Adv Inc v3	71.00	0.00	67.00	0.00	64.00	1.00
Ens Adv Inc Res v2	88.00	0.00	85.00	0.00	82.00	1.00

TABLE III
QUALITY COMPARISON. COMPARISON OF OUR AVERAGE PERCEPTUAL QUALITY AGAINST SIMILAR ATTACKS.

Method	**LPIPS**	**FID**
Clean	-	57.8
Ours	0.132	206.6*
DI-FGSM	0.131	67.1
ReColorAdv	0.158	75.1
DiffAttack	0.126	62.3

Fig. 2. **Results from the adversarial defense** The first figure shows the original image, the second shows an attacked non-immunized image, the third shows an immunized image with a mask, and the last one shows an immunized image with editing made to look like the benign image

such as DiffAttack (0.126) and outperforms color-based attacks like ReColorAdv (0.158). This confirms that optimizing the latent space preserves high perceptual fidelity. See 3 for a visualization. While the FID scores were large (> 200), we attribute this to the unreliability of FID when calculated on small sample sizes ($N = 100$) rather than a true degradation in image quality.

B. Performance against Adversarial Defenses

Various experiments were conducted to evaluate prevention from diffusion based attacks, here PhotoGuard is used to evaluate whether the attack can evade the defense mechanism.

1) Experimental Setup: The evaluation was conducted on three conditions, Non-immunized image(baseline), Immunized Image (defense applied), and Immunized Image after editing (defense applied with modifications). A gradio interface is deployed and used to allow the immunization of images to selectively mask regions to preserve specific components. A modification on parameters, such as prompts and seeds, can be applied to generate different edited versions of the immunized images.

2) Results and Analysis: The defense validation revealed that the PhotoGuard successfully neutralized most attacks on both non-immunized and fully immunized images. The only attack success out of the sample size of 30 images, on an immunized-edited image. This demonstrates that while the defense mechanism is generally effective, it can still be compromised while editing the image to make it as close as possible to the benign image.

This validation showed that the attack preserves masked components during the adversarial generation process, safeguarding against diffusion-based attacks.

TABLE IV
DEFENSE VALIDATION RESULTS

Condition	**Images**	**Defense Rate (%)**
Non-immunized	10	0.0
Immunized (before editing)	10	100.0
Immunized (after editing)	10	90.0

C. Expectation and Robustness of Transformations

In this section, we evaluate the robustness of our method against variations in transformation parameters. Unlike deterministic attacks that often fail under slight parameter deviations (e.g., resizing or resampling noise), our EoT-based optimization is designed to maintain attack success across a continuous range.

1) Experimental Setup: We evaluated the method on two transformation families, **Scaling** and **Blurring**, using a standardized optimization protocol of $K = 50$ iterations. To approximate the transformation distribution effectively, we utilized a jitter radius of $r = 0.1$ around three fixed center points: $\{0.5, 0.65, 0.8\}$. This sampling strategy creates an effective coverage range of $\Theta_{scale} \in [0.4, 0.9]$. The evaluation was performed on a subset of 10 randomly selected images from the examaple data set. To assess the impact of the objective function on robustness, we compared the performance of two distinct attack loss formulations: standard Cross-Entropy (CE) and the Carlini-Wagner (CW) margin loss.

Fig. 3. **Results from the Scaling Attack over the tranformation range 0.5x-0.8x** The first column shows the original images from the sample dataset, while Columns 2–6 display images with increasing scaling factors. The soup bowl adversarial image achieved 100% attack success, the car image achieved 80%, and the knives image achieved 0%.

TABLE V
ROBUSTNESS EVALUATION ACROSS TRANSFORMATION RANGES

Transformation	Benign preserved	CE avg. attack	CW avg. attack
Scaling	100%	54.0%	64.0%
Blurring	CE: 90% CW: 50%	46.0%	64.0%

2) Results and Analysis: To quantify performance, we track the **Attack Success Rate (ASR)**, defined as the percentage of test parameter points within Θ where the image is successfully classified as the target, and **Benign Preservation**, which represents the percentage of images retaining their original classification at the benign parameter point ($\theta_{\text{benign}} = 1.0$ for scaling).

As shown in Table V, the choice of loss function significantly impacts robustness. For scaling transformations, the CW loss achieved a higher average attack success rate (64.0%) compared to Cross-Entropy (54.0%). This confirms that maximizing the margin around the decision boundary, a key feature of CW loss is beneficial when optimizing against a distribution of transformations rather than a single point. Some of the visual results for the scaling are shown in Figure 5.

While CW loss improves attack robustness, it introduces a trade-off in benign preservation for certain transformations. In the blurring experiments, using CW loss caused a significant drop in benign accuracy (down to (50%)) compared to scaling (which preserved (100%)). In contrast, CE loss maintained much higher benign accuracy under the blurring attack, preserving up to 90%. This suggests that stronger gradients toward the adversarial target can occasionally override the benign preservation constraint unless the weight λ_{benign} is dynamically adjusted. Overall, the results demonstrate that our EoT-based latent optimization successfully prevents overfitting to specific anchor points, as the high ASR indicates that the generated perturbations generalize well to unseen parameters within the interpolation range.

V. ABLATION STUDY: CHOICE OF LOSS FUNCTION

The ablation study investigates the impact of loss function choice on attack performance, comparing Carlini-Wagner (CW) loss against Cross-Entropy (CE) loss. With this analysis, we determine the transferability and answer the fundamental design question of which function would produce the most effective attacks.

A. Experimental Setup

The study uses a subset of the Food-11 dataset, focusing on bread images to evaluate classification accuracy under different loss functions. The subset consists of 100 images that have been used to generate adversarial examples under both loss functions. We evaluate Cross-Entropy loss, which was used by the authors of PhotoGuard, and Carlini-Wagner loss, which our preliminary analysis indicated showed potential for improved performance. For each loss function, we follow an identical experimental pipeline to ensure fair comparison. Images undergo diffusion-based attacks in their latent space and are optimized using either CE or CW loss functions. The attack success is measured based on the percentage of images that successfully fool the classifier.

B. Key Findings

The CW loss significantly outperformed CE loss across multiple transformation types. In the sample data tested against 100 images, CW loss successfully attacked 39 images, while CE loss failed to generate successful attacks on any images. For scaling transformations, CW loss achieved a 64% average attack success rate compared to 54% for CE loss. Similarly, for blurring transformations, CW loss maintained a 64% success

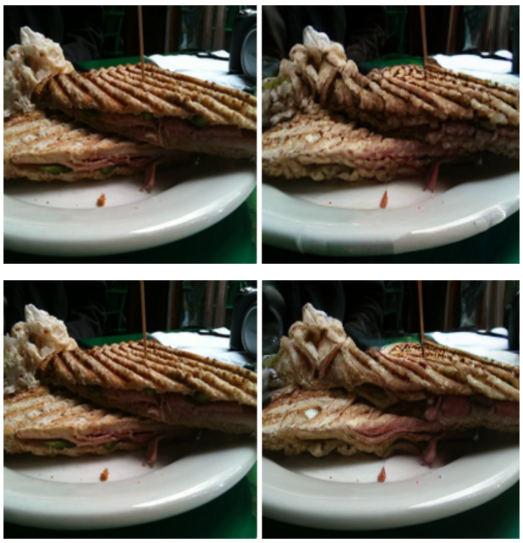

Fig. 4. **Results from the comparison of CW loss with CE loss.** The left column represents the benign image, the second image represents the successful CW loss attack, and the fourth image represents CE loss, which was a failure.

rate while CE loss achieved only 46%, demonstrating CW loss's superiority in both scenarios.

C. Implications

The ablation study confirms the intuition that CW loss is better suited for adversarial attack generation than CE loss. CW loss improved the transferability of adversarial attacks, successfully attacking all unseen test cases, while CE loss failed on all of them. Cross-entropy loss optimizes for correct classification during model training, whereas CW loss is specifically designed to fool pre-trained models by maximizing the margin between target and true class logits. This fundamental difference makes CW loss the preferred choice for generating effective adversarial attacks with higher success rates.

D. Limitations

While the new loss function demonstrates superior performance, a more comprehensive study is needed to determine whether CW loss consistently outperforms CE loss across a broader range of classes and a larger dataset. Such an extensive evaluation would require substantial computational resources and overhead.

VI. Future Work

While SleeperDiff shows potential, there are several areas where this project could be expanded. First, we would validate our results on a much larger dataset, as our primary evaluation was limited to 100 images, with subsequent experiments using only 30 images. We also want to further optimize the hyperparameters to better understand the specific trade-offs between attack success, image quality, and computational cost.

In terms of capabilities, future work can test more complex transformations beyond scaling and blurring. Related work [16] implement JPEG compression and diffusion purification transforms. It would also be valuable to investigate if this attack could become *physically realizable*: testing whether SleeperDiff can fool real-world sensors rather than just digital classifiers.

Finally, further investigation can dig into why the transferability of this attack was so poor relative to the current works of black box attacks. We also aim to compare our method's imperceptibility against a wider range of techniques to better benchmark just how stealthy these latent-space perturbations really are.

VII. Conclusion

In this work, we presented a diffusion-based framework for generating transform-dependent adversarial examples that remain benign under normal viewing conditions while reliably triggering misclassifications under specific image transformations. Our experiments show that latent-space optimization preserves high perceptual quality and, when combined with Expectation over Transformations (EoT), produces perturbations that generalize across continuous transformation ranges rather than overfitting to single parameters. Among the loss functions evaluated, the CW margin loss consistently achieved higher attack success and improved robustness compared to Cross-Entropy. We further observed that the Photoguard defense neutralized most adversarial examples, though one immunized-and-edited case remained vulnerable, suggesting targeted weaknesses in current defenses. While the approach demonstrates strong potential, it also comes with limitations, including high computational cost due to latent optimization and multi-transform EoT sampling, a large number of tunable hyperparameters, a small evaluation dataset, and limited black-box transferability. Addressing these challenges through improved optimization strategies, stronger transfer learning techniques, and larger-scale experimentation offers promising directions for future research.

References

[1] Google DeepMind, "AI achieves silver-medal standard solving International Mathematical Olympiad problems," DeepMind Blog, July 2024.

[2] Sakana AI, "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery," Sakana AI Blog, August 2024.

[3] A. Athalye, N. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[4] J. Song, Y. Teng, X. Song, and S. Ermon, "Denoising Diffusion Implicit Models," *arXiv preprint arXiv:2010.02502*, 2020.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[8] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[9] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv preprint arXiv:1801.02612*, 2018.

[10] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "Colorfool: Semantic adversarial colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1151–1160.

[11] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[13] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[14] J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, and Z. Shi, "Diffusion models for imperceptible and transferable adversarial attack," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] H. Xue, A. Araujo, B. Hu, and Y. Chen, "Diffusion-based adversarial sample generation for improved stealthiness and controllability," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2894–2921, 2023.

[16] Y. Tan, Z. Cai, and M. S. Asif, "Transform-Dependent Adversarial Attacks," *arXiv preprint arXiv:2406.08443*, 2025.

[17] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, A. Madry "Raising the Cost of Malicious AI-Powered Image Editing,"*arXiv:2302.06588*, 2023.

VIII. Appendix: More Examples

Fig. 5. **Results from the Blurring Attack over the tranformation range 0.40x-3.1x** The first column shows the original images from the sample dataset, while Columns 2–6 display images with increasing blurring factors. The soup bowl and car adversarial images achieved 100% attack success, whereas the monastery image achieved 0%.