Auditing Traffic-Sign Robustness via DDIM Inversion: Do Diffusion Latents Preserve Shadow Attacks?

Anonymous CVPR submission

Paper ID

Abstract

Traffic-sign classifiers must remain reliable under physically plausible perturbations such as cast shadows, which can be optimized into stealthy adversarial attacks while appearing visually indistinguishable from benign environmental effects. Pixel-space diagnostics often fail to separate adversarial shadows from incidental ones without degrading scene content. We propose deterministic DDIM inversion through a domain-adapted Stable Diffusion v1.5 model as an analysis interface for this problem. After fine-tuning on an in-distribution GTSRB subset under four shadow-aware curricula, DDIM reconstructions are visually faithful and behavior-preserving for non-adversarial inputs (accuracy deltas <3points; confidence shifts <0.01), while adversarial effects persist after reconstruction. A lightweight MLP trained on flattened inversion latents achieves ROC-AUC ≈ 0.96 *on the in-distribution split with meaningful transfer under class shift (GTSRB Yield) and dataset shift (LISA). These results position diffusion inversion as a practical auditing interface for shadow attacks, providing latent representations where adversarialness is measurable without pixel-space heuristics.*

1. Introduction

Traffic sign recognition is safety-critical in modern autonomous-driving pipelines, yet classifiers remain vulnerable to visually subtle environmental variations—especially cast shadows, which are pervasive, physically plausible, and can significantly degrade model confidence or induce misclassification [18]. Prior work has shown that shadow patterns can be *optimized* into effective physical-world adversarial perturbations while remaining visually similar to natural shadows [10, 18]. This motivates a concrete security question: how can we analyze and detect adversarial shadows in a way that generalizes across classifiers, sign categories, and datasets? Pixel-space diagnostics are poorly suited here because physically plausible shadows are structured by geometry and illumination, making it difficult to separate adversarial from benign shadows without heuristics that degrade scene content or alter classifier behavior.

(a) Stop, benign (b) Yield, benign (c) Stop, adversarial (d) Yield, adversarial

Figure 1. DDIM reconstructions maintain high visual similarity to inputs across benign and adversarial shadow conditions.

We propose using *diffusion inversion* as an analysis interface for this problem. We map traffic-sign images into the latent space of a domain-adapted Stable Diffusion v1.5 model via DDIM inversion [13, 15] and reconstruct them back to image space. This supports two complementary goals: (1) when reconstructions are behavior-preserving, the inversion latents provide a stable representation without altering the phenomenon under study; (2) if adversarial shadows imprint measurable signatures in latent space that benign shadows do not, lightweight detection becomes possible without pixel-level heuristics.

We evaluate on GTSRB and LISA [9, 16] across five input variants (Clean, Tri-Benign, Tri-Adv, Square-Benign, Square-Adv) using CNN [6] and ViT [3] classifiers. After domain adaptation, DDIM reconstructions are visually faithful and largely behavior-preserving for clean and benign-shadow inputs, while adversarial effects persist after reconstruction, confirming the pipeline does not act as a purifier. A lightweight MLP trained on flattened inversion latents achieves ROC-AUC ≈ 0.96 on the in-distribution split with meaningful transfer under both class and dataset shift. We make the following contributions:

- **Diffusion inversion as a behavior-preserving analysis interface.** We apply DDIM inversion and reconstruction to clean, benign-shadow, and adversarial-shadow traffic-sign images. After fine-tuning, reconstructions largely preserve

downstream classifier behavior on non-adversarial inputs while retaining adversarial effects.

- **Latent-space separability and detection.** Benign and adversarial shadows are separable in inversion latents, and a lightweight MLP detector achieves ROC-AUC ≈ 0.96 on the ID split with meaningful transfer to OOD subsets.
- **Out-of-distribution evaluation.** We evaluate reconstruction stability and latent separability under class shift (GTSRB Yield) and dataset shift (LISA Stop/Yield), quantifying generalization beyond the fine-tuning distribution.

2. Related Work

Diffusion inversion for anomaly detection and editing. Sakai *et al.* propose InvAD, which performs DDIM inversion and scores anomalies directly in latent space rather than in RGB, reducing sensitivity to noise-strength tuning [14]. Liu *et al.* extend diffusion-based anomaly detection to multi-view settings [7]. In the editing domain, Mokady *et al.* refine DDIM inversion via NULL-text optimization to improve reconstruction fidelity for real-image editing [11]. Our work shares the use of DDIM inversion but repurposes it as an *auditing interface* for physically plausible adversarial perturbations rather than as an anomaly detector or editing tool.

Shadow-based physical attacks on traffic-sign recognition. Zhong *et al.* demonstrate that realistic cast shadows can serve as stealthy adversarial perturbations for traffic-sign classifiers in both simulation and drive-by scenarios [18]. MohajerAnsari *et al.* extend this to temporally coherent shadow sequences optimized via a genetic algorithm with an attention-disruption objective [2, 10]. Related physical attack surfaces include LiDAR shadow manipulation [5] and RP2, which optimizes robust surface perturbations under varying viewing conditions [4]. Our focus differs: rather than crafting attacks, we audit and detect shadow-based perturbations through diffusion inversion latents.

Diffusion models for adversarial purification. Nie *et al.* propose DiffPure, which applies forward diffusion followed by reverse denoising to remove adversarial perturbations before classification [12]. Unlike purification approaches, we show that adversarial effects can *persist* through inversion–reconstruction and instead leverage inversion latents as detection signals.

3. Background and Threat Model

Task setup and notation. Let $x \in \mathbb{R}^{H\times W\times 3}$ denote an RGB traffic-sign image and $y \in \mathcal{Y}$ its ground-truth class. A classifier $f(\cdot)$ predicts $\hat{y} = \arg\max_i p(i \mid x)$. We report classification accuracy and true-class confidence $c(x) = p(y \mid x)$, tracking changes under both shadowing and diffusion reconstruction.

Physically plausible shadow model. We generate shadows using the ShadowAttack formulation [18], in which a shadow is parameterized by a polygon P (triangle or quadrilateral) and an attenuation factor $k \in (0, 1)$ that reduces luminance within $P \cap M$, where M is the sign mask. We instantiate two polygon families: 3-vertex (triangular) and 4-vertex (square-like) shadows.

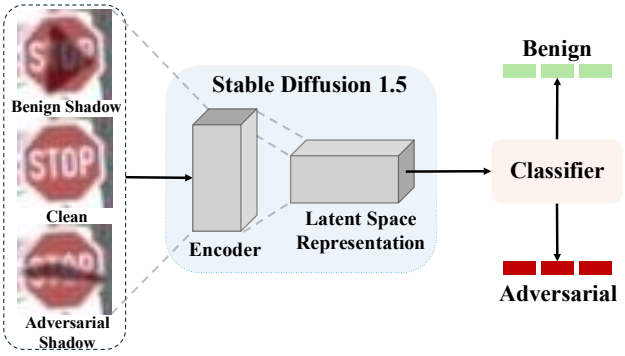

Figure 2. Pipeline overview: traffic-sign images are mapped to diffusion latent space via DDIM inversion; reconstructions are evaluated for classifier behavior preservation, and inversion latents are used for binary adversarial-shadow detection.

Benign vs. adversarial shadows. Adversarial shadows (*Tri-Adv* / *Square-Adv*) are produced by optimizing shadow parameters via query-based PSO with Expectation over Transformations (EOT) [1] to degrade classifier performance while preserving physical plausibility [18]. The attack objective is untargeted: reduce true-class confidence and/or induce misclassification. Benign shadows (*Tri-Benign* / *Square-Benign*) use randomly sampled polygon geometries and attenuation values without classifier-directed optimization.

Threat model. The attacker can cast a physically plausible shadow on a traffic sign, controlling polygon geometry and attenuation k subject to feasibility constraints (bounded attenuation, reasonable polygon size, luminance-only modification). The attacker is query-based: it observes output confidence scores but does not require model weights or gradients, consistent with ShadowAttack assumptions [18].

4. Data and Experimental Assets

Datasets and class selection. We use two public traffic-sign datasets: GTSRB [16] and LISA [9]. For diffusion fine-tuning, we select two GTSRB classes: *Stop* and *Speed Limit 30 km/h*. For OOD evaluation, we use *Yield* from GTSRB (class shift) and both *Stop* and *Yield* from LISA (dataset shift and dataset+class shift, respectively).

In-distribution (ID) protocol. The diffusion model is fine-tuned exclusively on GTSRB *Stop* and *Speed Limit 30 km/h* using the official train split (**GTSRB-ID-train**); the corresponding test split (**GTSRB-ID-test**) is reserved for held-out ID evaluation.

Out-of-distribution (OOD) evaluation. We define two

Subset	Dataset	Classes	Usage	Shift type
GTSRB-ID-train	GTSRB	Stop, SpeedLimit30	Diffusion fine-tune, Evaluation	In-distribution
GTSRB-ID-test	GTSRB	Stop, SpeedLimit30	Diffusion Evaluation	In-distribution
GTSRB-OOD-yield	GTSRB	Yield	Evaluation	Class shift
LISA-OOD	LISA	Stop, Yield	Evaluation	Dataset + class shift

Table 1. Data protocol used throughout the paper. Diffusion models are fine-tuned only on **GTSRB-ID-train** (Stop + SpeedLimit30). All other subsets are held out for evaluation to measure generalization under class and dataset shift.

Table 2. Diffusion fine-tuning curricula. Each row shows one SD v1.5 variant; dashes indicate shadow families withheld from training.

Model	Clean	Tri-B	Tri-A	Sq-B	Sq-A
SD-Pret.		—	—	—	—
FT-Clean		—	—	—	—
FT-Tri				—	—
FT-Sq		—	—		
FT-Both					

OOD subsets never used during fine-tuning: **(i) GTSRB-OOD-yield** (Yield signs from GTSRB; class shift) and **(ii) LISA-OOD** (Stop and Yield signs from LISA; dataset and dataset+class shift). Table 1 summarizes the full data protocol.

Per-image shadow variants. For every base image x, we construct five variants: Clean, Tri-Benign, Tri-Adv, Square-Benign, and Square-Adv. Shadow generation follows the procedure described in Section 3. Table 3 provides a concise reference for all variants.

Preprocessing. For diffusion fine-tuning and DDIM inversion, all images are resized to 512×512. For classifier evaluation, images are resized and normalized to each backbone's native input pipeline.

Evaluation classifiers. We use two architectures—a CNN [6] and a ViT [3]—both trained on all 43 GTSRB classes. Both are evaluated on all ID and OOD subsets across all five shadow variants.

5. Method

Latent diffusion backbone. We build on Stable Diffusion v1.5 [13], which encodes an RGB image x into a latent $z \in \mathbb{R}^{4\times64\times64}$ via a VAE encoder and performs diffusion in this compressed space. Let $\epsilon_\theta(z_t, t)$ denote the U-Net noise predictor and $\bar{\alpha}_t$ the cumulative noise-schedule coefficients.

Fine-tuning curricula. Stable Diffusion v1.5 does not reliably reconstruct traffic-sign inputs without domain adaptation. We fine-tune five diffusion-model variants that differ only in shadow exposure during training: a clean-only baseline and four shadow-aware curricula. All fine-tuning uses exclusively GTSRB-ID-train (Stop and Speed Limit 30 km/h); OOD subsets are held out entirely. Table 2 summarizes all variants.

Training details. Images are resized to 512×512 and normalized to $[-1, 1]$. All variants are fine-tuned for 100 epochs with batch size 32, learning rate 1×10^{-5}, and bf16 mixed precision using AdamW [8] (weight decay 0.01).

DDIM inversion and reconstruction. Given image x, we encode it via the VAE to obtain z_0 and apply deterministic DDIM inversion to produce a latent trajectory $\{z_t\}_{t=0}^{T}$:

$$z_t = \sqrt{\tfrac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}\, z_{t-1} + \left(\sqrt{1-\bar{\alpha}_t} - \sqrt{\tfrac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}\sqrt{1-\bar{\alpha}_{t-1}}\right)\epsilon_\theta(z_{t-1}, t-1), \tag{1}$$

We refer to the terminal latent z_T as the *inversion latent*. Reconstruction reverses this process: deterministic DDIM denoising maps z_T back to $\hat{z}_0$, which is decoded to obtain $\hat{x}$. The scheduler and inference configuration are fixed across all variants.

Latent-space binary detector. To test whether inversion latents encode separable benign vs. adversarial signatures, we train a lightweight MLP on the flattened $z_T \in \mathbb{R}^{16384}$. Benign samples (Tri-Benign, Square-Benign) form the negative class; adversarial samples (Tri-Adv, Square-Adv) form the positive class; clean images are excluded. The MLP has three hidden layers ($1024 \to 256 \to 64$) with ReLU, dropout 0.3, and 2-way softmax output, trained for 20 epochs on an 80/20 split within GTSRB-ID-train. The trained detector is evaluated on OOD subsets without retraining.

Evaluation metrics. Reconstruction fidelity is measured by L1, LPIPS [17], and PSNR. Behavior preservation is assessed by comparing classifier accuracy and true-class confidence $c(x) = p(y \mid x)$ before and after reconstruction.

Variant	Geometry	Optimization	Description
Clean	None	None	Original image (no shadow)
Tri-Benign	Triangle (3 vertices)	None	Sampled feasible parameters
Tri-Adv	Triangle (3 vertices)	Query-based	Optimized to reduce true-class confidence / induce error
Square-Benign	Quad (4 vertices)	None	Sampled feasible parameters
Square-Adv	Quad (4 vertices)	Query-based	Optimized to reduce true-class confidence / induce error

Table 3. Per-image variants used in all experiments. Each base image is expanded into one clean sample and four physically plausible shadow variants (triangle vs. quadrilateral; benign vs. adversarial).

Category	Metric	What it measures
Recon.	L1 ($\|x - \hat{x}\|_1$)	Pixel-level reconstruction error
Recon.	LPIPS	Perceptual (feature-space) similarity
Recon.	PSNR	Signal-to-noise ratio of reconstruction
Behavior	Accuracy	Correct classification rate
Behavior	True-class conf. $c(x)$	Model certainty for correct class
Behavior	Conf. shift $\|c(x) - c(\hat{x})\|$	Behavior change after reconstruction
Behavior	ΔAcc / Δc	Before vs. after reconstruction deltas
Latent det.	Acc / P / R / F1	Binary detection quality
Latent det.	ROC-AUC	Threshold-independent separability
Latent det.	Log loss	Probabilistic calibration

Table 4. Metrics used to evaluate reconstruction fidelity, classifier behavior preservation, and separability of benign vs. adversarial shadows in diffusion inversion latents.

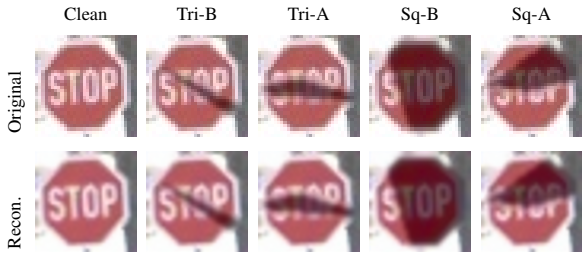

Figure 3. Original (top) and DDIM reconstruction (bottom) for all five shadow variants. Reconstructions preserve both benign and adversarial shadow content without smoothing or erasure.

Table 5. Latent-space binary classifier performance on inversion latents (ID split).

Acc	ROC-AUC	F1	Precision	Recall	LogLoss
0.8951	0.9591	0.8837	0.8902	0.8774	0.4754

We report the per-sample confidence shift:

$$L_{\text{conf}} = |c(x) - c(\hat{x})| \,. \tag{2}$$

For the latent detector we report accuracy, precision, recall, F1, ROC-AUC, and log loss. All metrics are summarized in Table 4. Diffusion fine-tuning, inversion, and reconstruction use the HuggingFace `diffusers` library with PyTorch.

6. Experiments

Our experiments address three questions: (1) does DDIM inversion–reconstruction preserve visual content across shadow variants? (2) does it preserve downstream classifier behavior? (3) are benign and adversarial shadows separable in inversion latents? All experiments follow the ID/OOD protocol of Section 4, using the five diffusion variants of Table 2 and both CNN and ViT classifiers across all subsets and shadow variants.

Experiment 1: Reconstruction fidelity. For each diffusion model, we compute L1, LPIPS, and PSNR between x and $\hat{x}$ on both ID and OOD subsets across all five shadow variants. This isolates three factors: shadow geometry (triangle vs. quadrilateral), adversarial optimization (benign vs. adversarial), and the effect of fine-tuning curriculum on reconstruction stability.

Experiment 2: Behavior preservation. For each sample, we run both classifiers on x and $\hat{x}$ and report accuracy and true-class confidence changes per subset and shadow variant.

Experiment 3: Latent-space separability. We train the MLP detector (Section 5) on flattened inversion latents within GTSRB-ID-train (80/20 split), then evaluate on OOD subsets without retraining to test transfer under class and dataset shift.

Controlled comparisons. Our analysis is structured around five axes: (i) SD-Pretrained vs. fine-tuned (domain adaptation effect); (ii) FT-Clean vs. shadow-aware variants (shadow exposure effect); (iii) FT-Tri vs. FT-Square (geometry-specific curricula); (iv) FT-Both vs. single-family variants (mixed vs. specialized training); and (v) benign vs. adversarial within the same geometry (shadow optimization effect).

7. Results and Analysis

Domain adaptation is necessary. Table 6 reports SD-Pretrained results. Without fine-tuning, reconstruction is poor across all variants: even clean ID images yield L1 = 11.58, LPIPS = 0.24, and PSNR = 24.57, with accuracy

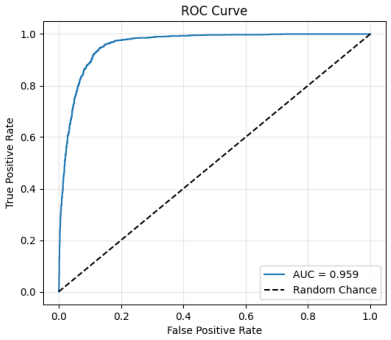

Figure 4. ROC curve for the latent-space detector (ID split).

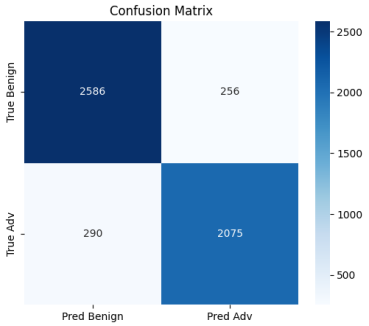

Figure 5. Confusion matrix for the latent-space binary detector (ID split), confirming high separability between benign and adversarial inversion latents.

drops up to 25 points after reconstruction. This degradation is not shadow-specific—domain mismatch destabilizes the pipeline regardless of input type.

Shadow-aware curricula improve fidelity. Fine-tuning closes most of the domain gap (Tables 7–9). FT-Clean reduces ID-train clean reconstruction to L1 = 3.48, LPIPS = 0.021, PSNR = 35.31, but degrades on shadowed variants. Adding shadow families during training reduces this gap: FT-Tri and FT-Square achieve L1 below 3.30 and LPIPS below 0.018 across all variants on ID-train. Table 10 shows that FT-Both attains the best overall fidelity (LPIPS = 0.015–0.016, PSNR = 36.88–37.47), confirming that broader shadow coverage during training improves reconstruction without hurting clean-input performance.

Behavior preservation on non-adversarial inputs. Across fine-tuned models, accuracy and confidence deltas for Clean, Tri-Benign, and Square-Benign variants are consistently small: accuracy shifts within ± 1.7 points and confidence shifts below 0.01 on ID-test. Benign shadow variants follow the same pattern with accuracy deltas under 3 points in nearly all cases.

Adversarial effects persist after reconstruction. For Tri-Adv and Square-Adv, classifier accuracy after reconstruction remains low across all fine-tuned models and both architectures, with true-class confidence staying around 0.09–0.13. This confirms the pipeline is not acting as a purifier. Table 3 provides qualitative confirmation: reconstructions retain the adversarial shadow pattern without visible artifact removal.

OOD generalization. OOD reconstruction degrades gradually relative to ID, with LPIPS increasing to 0.026–0.076 on GTSRB-OOD-yield and LISA-OOD versus 0.016–0.021 on ID-train. PSNR remains above 31 dB for all fine-tuned models, and behavior shifts stay moderate. We observe that the CNN performs better on ID subsets while the ViT generalizes better on OOD subsets, suggesting architectural inductive biases interact with the distribution shifts considered here.

Latent-space separability. Table 5 reports binary detector performance on ID inversion latents. The MLP achieves ROC-AUC = 0.959 with balanced precision (0.890) and recall (0.877), indicating that benign and adversarial signatures are separable in latent space without pixel-space processing (Figures 5 and 4). Transfer to OOD subsets shows meaningful but reduced performance, indicating latent signatures persist under shift while leaving room for improvement with broader training coverage.

8. Discussion

Diffusion inversion as an analysis interface. Once domain-adapted, DDIM inversion and reconstruction behave as an approximate identity mapping for clean and benign-shadow inputs, while adversarial effects persist. This property is useful for security analysis: the pipeline preserves the phenomenon under study rather than denoising it away. Latent separability further suggests that adversarial optimization leaves consistent signatures beyond what is apparent in pixels, motivating latent-space diagnostics as a complement to pixel-space approaches for physically constrained perturbations such as cast shadows [18].

Practical deployment. Our results suggest two concrete uses. First, *auditing*: inversion latents provide an additional signal to distinguish incidental shadows from adversarially optimized ones when a classifier exhibits atypical confidence drops. Second, *runtime monitoring*: a lightweight latent detector can flag suspicious inputs, enabling system-level responses such as temporal consistency checks or switching to a conservative operating mode. In practice, inversion would serve as a side-channel monitor: the primary classifier runs on the raw frame while the inversion latent is computed in parallel and passed to the detector. If flagged, the system reduces reliance on the sign channel and requests additional evidence across frames. This decouples detection from classification without assuming that any single preprocessing step can sanitize adversarial inputs.

Computational cost. Full DDIM inversion is more expen-

Table 6. Reconstruction and behavior-preservation results for **SD-Pretrained** (no domain fine-tuning).

		Reconstruction			CNN		ViT	
Dataset subset	Variant	↓ L1	↓ LPIPS	↑ PSNR	↑ Acc (Δ)	Conf (Δ)	↑ Acc (Δ)	Conf (Δ)
GTSRB-ID-train (Stop, SpeedLimit30)	Clean	11.58	0.2358	24.57	88.9 (9.1)	0.7837 (0.0935)	76.4 (15.2)	0.6692 (0.1494)
	Tri-Benign	11.65	0.2348	24.66	78.8 (20.8)	0.6597 (0.1787)	60.9 (20.5)	0.5261 (0.1923)
	Tri-Adv	11.16	0.2251	25.10	13.6 (8.3)	0.1018 (0.0109)	14.9 (11.2)	0.1256 (0.0988)
	Square-Benign	11.57	0.2330	24.77	75.7 (24.0)	0.6226 (0.1998)	55.9 (21.4)	0.4780 (0.2028)
	Square-Adv	11.11	0.2286	25.13	12.1 (6.7)	0.0917 (0.0014)	14.2 (12.7)	0.1208 (0.1109)
	Average	11.41	0.2315	24.85	53.8 (7.8)	0.4519 (0.0925)	44.5 (16.2)	0.3839 (0.1508)
GTSRB-ID-test (Stop, SpeedLimit30)	Clean	11.74	0.2369	24.44	90.4 (8.8)	0.8022 (0.0868)	82.2 (10.2)	0.7173 (0.1166)
	Tri-Benign	12.05	0.2466	24.26	79.4 (20.3)	0.6642 (0.1748)	64.3 (15.6)	0.5534 (0.1619)
	Tri-Adv	12.22	0.2477	24.39	10.6 (7.8)	0.0817 (0.0045)	13.7 (8.9)	0.1201 (0.0738)
	Square-Benign	12.11	0.2482	24.29	74.5 (25.4)	0.6098 (0.2155)	59.7 (18.7)	0.5069 (0.1863)
	Square-Adv	12.17	0.2478	24.49	8.8 (6.1)	0.0662 (0.0093)	13.4 (10.1)	0.1093 (0.0900)
	Average	12.06	0.2454	24.37	52.7 (8.1)	0.4448 (0.0964)	46.7 (12.7)	0.4014 (0.1257)
LISA-OOD (Stop, Yield)	Avg	8.97	0.2193	28.30	62.7	0.4293 (0.0211)	67.9	0.5546 (0.0176)
GTSRB-OOD-yield (Yield)	Avg	10.66	0.2457	25.59	74.5	0.6187 (0.0337)	72.6	0.6438 (0.0943)

Table 7. Reconstruction and behavior-preservation results for **FT-Clean** (fine-tuned on ID Clean only).

		Reconstruction			CNN		ViT	
Dataset subset	Variant	↓ L1	↓ LPIPS	↑ PSNR	↑ Acc (Δ)	Conf (Δ)	↑ Acc (Δ)	Conf (Δ)
GTSRB-ID-train (Stop, SpeedLimit30)	Clean	3.48	0.0208	35.31	97.3 (0.7)	0.8753 (0.0019)	91.8 (0.2)	0.8215 (0.0029)
	Tri-Benign	4.17	0.0271	33.71	95.1 (4.5)	0.8160 (0.0224)	80.6 (0.9)	0.7111 (0.0073)
	Tri-Adv	4.74	0.0332	32.74	13.7 (8.4)	0.1262 (0.0353)	25.7 (0.4)	0.2228 (0.0016)
	Square-Benign	4.40	0.0292	33.32	93.8 (6.0)	0.7933 (0.0291)	76.0 (1.3)	0.6697 (0.0111)
	Square-Adv	4.88	0.0350	32.64	12.7 (7.2)	0.1169 (0.0238)	25.0 (1.9)	0.2157 (0.0160)
	Average	4.33	0.0291	33.54	62.5 (0.9)	0.5455 (0.0011)	59.8 (0.9)	0.5282 (0.0066)
GTSRB-ID-test (Stop, SpeedLimit30)	Clean	3.94	0.0250	34.03	98.9 (0.3)	0.8901 (0.0011)	92.6 (0.3)	0.8405 (0.0066)
	Tri-Benign	4.16	0.0263	33.70	96.9 (2.8)	0.8261 (0.0129)	81.0 (1.1)	0.7181 (0.0028)
	Tri-Adv	4.30	0.0274	33.51	12.2 (9.3)	0.1149 (0.0377)	23.3 (0.8)	0.2048 (0.0109)
	Square-Benign	4.29	0.0276	33.51	95.5 (4.4)	0.8043 (0.0210)	78.4 (0.0)	0.6872 (0.0060)
	Square-Adv	4.37	0.0279	33.51	9.8 (7.1)	0.0951 (0.0196)	24.1 (0.6)	0.2007 (0.0014)
	Average	4.21	0.0268	33.65	62.7 (1.8)	0.5461 (0.0049)	59.9 (0.6)	0.5303 (0.0031)
LISA-OOD (Stop, Yield)	Avg	5.71	0.0763	31.87	66.3	0.4546 (0.0041)	72.2	0.5886 (0.0164)
GTSRB-OOD-yield (Yield)	Avg	5.15	0.0478	31.84	67.2	0.5746 (0.0104)	79.2	0.7071 (0.0310)

sive than a single classifier forward pass. Practical mitigations include selective invocation triggered by confidence anomalies, fewer inversion steps for a coarser but faster signal, distillation into a compact encoder, and offline forensic use for post-hoc incident analysis.

9. Limitations

Narrow fine-tuning distribution. Our diffusion models are fine-tuned on only two GTSRB classes. The observed OOD degradation indicates that broader coverage across sign categories, geographic styles, and capture conditions is needed for robust generalization. Similarly, the MLP detector may overfit to the specific shadow generator and fine-tuning curriculum used; cross-generator transfer and integration with temporal or multi-sensor cues remain untested.

Synthetic shadows only. All shadows are synthetically generated using a physically motivated parameterization, but we do not verify that every configuration is achievable under real-world illumination, camera motion, or sensor noise. A more complete evaluation would include real capture loops with physical signs across diverse environments.

Attack model coverage. We study only polygonal cast shadows optimized via query-based PSO. Other physically feasible perturbations (e.g., patches, projected light, adverse weather) may behave differently under diffusion inversion, and broader threat model coverage is needed before treating diffusion latents as a general detection substrate.

Table 8. Reconstruction and behavior-preservation results for **FT-Tri** (fine-tuned on ID Clean + triangular shadows).

		Reconstruction			CNN		ViT	
Dataset subset	Variant	↓ L1	↓ LPIPS	↑ PSNR	↑ Acc (Δ)	Conf (Δ)	↑ Acc (Δ)	Conf (Δ)
GTSRB-ID-train (Stop, SpeedLimit30)	Clean	2.92	0.0157	37.03	97.9 (0.1)	0.8771 (0.0001)	91.3 (0.4)	0.8202 (0.0016)
	Tri-Benign	3.04	0.0163	36.77	98.0 (1.6)	0.8338 (0.0046)	81.3 (0.2)	0.7177 (0.0007)
	Tri-Adv	3.14	0.0168	36.55	10.0 (4.7)	0.1055 (0.0146)	26.2 (0.1)	0.2239 (0.0005)
	Square-Benign	3.15	0.0168	36.42	97.4 (2.4)	0.8155 (0.0069)	76.9 (0.5)	0.6793 (0.0015)
	Square-Adv	3.27	0.0173	36.27	10.3 (4.8)	0.1027 (0.0096)	25.7 (1.2)	0.2217 (0.0100)
	Average	3.10	0.0166	36.61	62.7 (1.1)	0.5469 (0.0025)	60.3 (0.4)	0.5326 (0.0022)
GTSRB-ID-test (Stop, SpeedLimit30)	Clean	3.72	0.0195	34.81	98.9 (0.3)	0.8916 (0.0026)	94.0 (1.7)	0.8448 (0.0109)
	Tri-Benign	3.81	0.0203	34.67	97.3 (2.4)	0.8359 (0.0031)	81.6 (1.7)	0.7229 (0.0076)
	Tri-Adv	3.88	0.0210	34.58	9.8 (7.0)	0.1023 (0.0251)	23.4 (0.9)	0.2039 (0.0100)
	Square-Benign	3.85	0.0205	34.59	97.2 (2.8)	0.8194 (0.0059)	78.7 (0.3)	0.6925 (0.0007)
	Square-Adv	3.96	0.0215	34.44	9.9 (7.2)	0.0959 (0.0204)	24.6 (1.1)	0.2035 (0.0042)
	Average	3.84	0.0206	34.62	62.6 (1.7)	0.5490 (0.0078)	60.5 (1.1)	0.5335 (0.0064)
LISA-OOD (Stop, Yield)	Avg	4.50	0.0312	33.61	67.7	0.4589 (0.0084)	72.6	0.5903 (0.0180)
GTSRB-OOD-yield (Yield)	Avg	4.55	0.0270	33.56	71.7	0.5977 (0.0127)	82.8	0.7390 (0.0009)

Table 9. Reconstruction and behavior-preservation results for **FT-Square** (fine-tuned on ID Clean + quadrilateral shadows).

		Reconstruction			CNN		ViT	
Dataset subset	Variant	↓ L1	↓ LPIPS	↑ PSNR	↑ Acc (Δ)	Conf (Δ)	↑ Acc (Δ)	Conf (Δ)
GTSRB-ID-train (Stop, SpeedLimit30)	Clean	2.89	0.0156	37.11	97.9 (0.1)	0.8771 (0.0001)	91.1 (0.5)	0.8199 (0.0013)
	Tri-Benign	3.02	0.0163	36.75	98.1 (1.5)	0.8344 (0.0040)	81.2 (0.2)	0.7180 (0.0004)
	Tri-Adv	3.12	0.0168	36.54	11.0 (5.6)	0.1094 (0.0185)	26.6 (0.5)	0.2265 (0.0021)
	Square-Benign	3.06	0.0163	36.72	97.5 (2.2)	0.8164 (0.0060)	76.9 (0.5)	0.6791 (0.0017)
	Square-Adv	3.18	0.0169	36.51	9.7 (4.2)	0.1013 (0.0082)	25.9 (1.0)	0.2218 (0.0099)
	Average	3.05	0.0164	36.72	62.8 (1.2)	0.5477 (0.0033)	60.3 (0.3)	0.5331 (0.0017)
GTSRB-ID-test (Stop, SpeedLimit30)	Clean	3.70	0.0195	34.86	98.9 (0.3)	0.8918 (0.0028)	92.9 (0.6)	0.8427 (0.0088)
	Tri-Benign	3.78	0.0203	34.72	97.6 (2.2)	0.8367 (0.0023)	81.2 (1.4)	0.7228 (0.0075)
	Tri-Adv	3.85	0.0212	34.62	9.8 (6.9)	0.0992 (0.0220)	23.2 (0.7)	0.2025 (0.0086)
	Square-Benign	3.83	0.0206	34.63	97.3 (2.7)	0.8204 (0.0049)	78.9 (0.6)	0.6924 (0.0008)
	Square-Adv	3.93	0.0217	34.51	9.3 (6.6)	0.0910 (0.0155)	23.9 (0.4)	0.2005 (0.0012)
	Average	3.82	0.0207	34.67	62.6 (1.7)	0.5478 (0.0066)	60.0 (0.7)	0.5322 (0.0051)
LISA-OOD (Stop, Yield)	Avg	4.46	0.0316	33.66	67.4	0.4582 (0.0078)	72.5	0.5901 (0.0179)
GTSRB-OOD-yield (Yield)	Avg	4.46	0.0261	33.76	71.2	0.5949 (0.0099)	82.7	0.7401 (0.0020)

10. Conclusion

We studied physically plausible shadow attacks on traffic-sign recognition through the lens of diffusion inversion. Fine-tuning Stable Diffusion v1.5 on a narrow GTSRB subset makes DDIM inversion and reconstruction largely behavior-preserving for clean and benign-shadow inputs, while adversarial effects persist rather than being removed. Inversion latents further expose discriminative structure that a lightweight MLP exploits to separate benign from adversarial shadows (ROC-AUC ≈ 0.96 on ID data), with gradual degradation under class and dataset shift. These results position diffusion inversion in the monitoring and auditing layer of a perception stack: it provides a complementary signal—latent adversarial signatures—that pixel-space classifiers cannot easily recover on their own. Future work should expand fine-tuning coverage, characterize sensitivity to inversion hyperparameters, and study transfer across shadow generators and other physically feasible perturbation types.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

Table 10. Reconstruction and behavior-preservation results for **FT-Combined** (fine-tuned on ID Clean + triangular + quadrilateral shadows).

		Reconstruction			CNN		ViT	
Dataset (Signs)	Subset	↓ L1	↓ LPIPS	↑ PSNR	↑ Acc (Δ)	Conf (Δ)	↑ Acc (Δ)	Conf (Δ)
GTSRB Train (Stop, Speed)	Clean	**2.78**	**0.0147**	**37.47**	97.9 (0.1)	**0.8770** (0.0002)	**91.5** (0.2)	**0.8204** (0.0018)
	BS	2.94	0.0153	37.11	97.4 (2.4)	0.8158 (0.0066)	77.1 (0.3)	0.6795 (0.0013)
	BT	2.87	0.0150	37.27	**98.0** (1.6)	0.8343 (0.0041)	81.2 (0.2)	0.7177 (0.0007)
	AS	3.05	0.0158	36.88	9.6 (4.1)	0.1001 (0.0070)	25.8 (1.1)	0.2216 (0.0101)
	AT	2.97	0.0155	37.06	9.4 (4.1)	0.1021 (0.0112)	26.3 (0.2)	0.2239 (0.0005)
	Average	2.92	0.0153	37.16	62.5 (0.8)	0.5459 (0.0015)	60.4 (0.3)	0.5326 (0.0022)
GTSRB Test (Stop, Speed)	Clean	**3.65**	**0.0186**	**35.05**	**98.9** (0.3)	**0.8927** (0.0037)	**94.0** (1.7)	**0.8455** (0.0116)
	BS	3.79	0.0194	34.82	97.5 (2.5)	0.8209 (0.0044)	79.2 (0.9)	0.6958 (0.0026)
	BT	3.75	0.0193	34.88	97.7 (2.0)	0.8364 (0.0026)	81.6 (1.7)	0.7256 (0.0103)
	AS	3.91	0.0205	34.64	9.2 (6.5)	0.0886 (0.0131)	24.6 (1.1)	0.2025 (0.0032)
	AT	3.84	0.0202	34.75	9.5 (6.6)	0.0949 (0.0177)	23.4 (0.9)	0.2009 (0.0070)
	Average	3.79	0.0196	34.83	62.6 (1.7)	0.5467 (0.0055)	60.6 (1.3)	0.5341 (0.0069)
LISA (Stop, Yield)	LISA Avg	4.55	0.0307	33.60	67.7	0.4600 (0.0095)	72.9	0.5894 (0.0172)
Yield (Yield)	Yield Avg	4.47	0.0243	33.88	69.3	0.5860 (0.0009)	82.6	0.7376 (0.0005)

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 2

[5] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. Invisible but detected: Physical adversarial shadow attack and defense on lidar object detection. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*, Seattle, WA, USA, 2025. 2

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002. 1, 3

[7] Chieh Liu, Yu-Min Chu, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning diffusion models for multi-view anomaly detection. In *Computer Vision – ECCV 2024, 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXIII*, pages 328–345. Springer, 2024. 2

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[9] Andreas Møgelmose. Lisa traffic sign dataset, 2012. 1, 2

[10] Pedram MohajerAnsari, Amir Salarpour, David Fernandez, Cigdem Kokenoz, Bing Li, and Mert D Pesé. Attention-aware temporal adversarial shadows on traffic sign sequences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3591–3599, 2025. 1, 2

[11] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2023. 2

[12] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 2

[13] 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*, pages 10684–10695, 2022. 1, 3

[14] Shunsuke Sakai, Xiangteng He, Chunzhi Gu, Leonid Sigal, and Tatsuhito Hasegawa. Invad: Inversion-based reconstruction-free anomaly detection with diffusion models. *arXiv preprint arXiv:2504.05662*, 2025. 2

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[16] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks (IJCNN)*, 2011. 1, 2

[17] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[18] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15345–15354, 2022. 1, 2, 5