False Positives Matter: Multidimensional Localization Evaluation and Training-Free Explainable Adversarial Patch Defense

Lihua Jing[1,2], **Rui Wang**[1,2*], **Jinwen Zhong**[1,2], **Runbo Li**[1,2], **Zixuan Zhu**[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{jinglihua, wangrui, zhongjinwen, lirunbo, zhuzixuan}@iie.ac.cn

Abstract

Adversarial patch attacks pose a significant threat to visual systems. While current patch purification-based defense methods enhance core metrics of visual perception models, they overlook the critical issue of false positive patches, severely compromising image usability. This paper reveals the inadequacy of existing evaluations for adversarial patch defenses, and pioneers a multidimensional adversarial patch localization evaluation framework, which comprehensively quantifies false positives, recall capability, and overall localization accuracy, providing a novel perspective for comparative analysis within the field. Furthermore, building upon the observation that false positives stem from a lack of semantic understanding, we propose a Semantic-Aware Training-free Explainable Defense method (SATED). SATED achieves zero-shot patch localization, false detection correction, and decision explanation by constructing a patch reasoning chain, while simultaneously performing integrated text-guided patch inpainting. Extensive experiments across digital and physical scenarios, detection and segmentation tasks, and diverse adversarial patches, demonstrate that our method significantly reduces false positives and doubles the overall patch localization accuracy, boosting both the generalizability and explainability of the defense.

1 Introduction

Adversarial patch attacks achieve a balance between attack effectiveness and physical feasibility by applying pixel perturbations to local regions of input images, posing a significant threat to real-world visual systems. With ongoing research advancements in recent years, adversarial patch attacks have expanded to multiple core tasks in computer vision (Brown et al. 2017; Thys et al. 2019; Nesti et al. 2022) with increasingly diverse patch types (Hu et al. 2021; Tan et al. 2021; Hu et al. 2022), enhancing both attack capabilities and stealthiness.

To counter the escalating threat of adversarial patch attacks, researchers have developed four main defense categories: adversarial training-based defenses (Gittings et al. 2020), model modification-based defenses (Yu et al. 2023), patch purification-based defenses (Liu et al. 2022), and certifiably robust defenses (Xiang et al. 2023). Among these, patch purification-based defenses stand out for their universal applicability. By localizing and removing adversarial patches from input images, they effectively secure model predictions across different model structures.

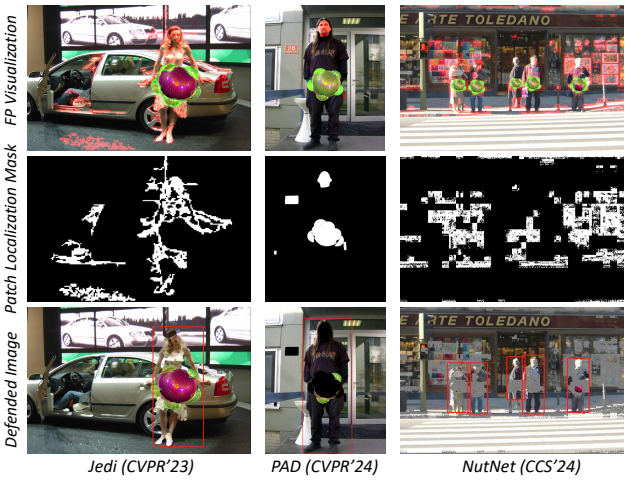

Figure 1: The false positives introduced by state-of-the-art defenses, though not affecting the detection of pedestrians, significantly degrade the overall usability of the images.

Existing patch purification defenses leverage various techniques to locate and eliminate adversarial patches, such as methods based on gradient anomalies (Naseer et al. 2019), high entropy (Tarchoun et al. 2023), external segmenters (Liu et al. 2022), autoencoder reconstruction (Lin et al. 2024), semantic independence and spatial heterogeneity analysis (Jing et al. 2024). While these methods significantly improve model performance on attacked images, insufficient attention has been paid to the efficacy of adversarial patch localization, especially false positives.

Figure 1 showcases examples of state-of-the-art defenses against Flower patch (Tan et al. 2021). The visualization of false positives (red regions) shows that while patch regions are detected, many other areas are also erroneously identified, particularly in complex scenes. Although eliminating these regions might not affect the detection of primary objects (e.g., pedestrians), it can have a detrimental impact on data usability. As shown, crucial information, such as human

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

faces, is incorrectly removed. Such an outcome is highly problematic and unacceptable for practical applications.

While current works often report no significant degradation in detection metrics on clean samples (Liu et al. 2022; Jing et al. 2024), this metric alone may not be sufficient. The erroneous removal of non-patch regions can severely affect image usability without causing a drop in detection metrics. Therefore, we believe that a comprehensive evaluation of patch localization performance is essential to determine whether a defense truly plays a sufficiently positive role.

To address this, we introduce multidimensional evaluation metrics designed to comprehensively assess patch misidentification, recall capability, and overall localization accuracy. Based on this new evaluation framework and an in-depth analysis of existing defenses, we found that most false positives arise from a fundamental lack of semantic understanding. For instance, the faces and shop windows in Figure 1 could be correctly preserved through semantic reasoning.

Therefore, we propose SATED, a Semantic-Aware, Training-Free, and Explainable Defense. SATED significantly reduces misidentification and addresses two major challenges faced by existing defense: limited generalization and poor explainability. Unlike most prior methods that rely on patch priors or training with attack data (Liu et al. 2022; Lin et al. 2024), SATED is capable of handling various adversarial patches without training, fine-tuning, or parameter adjustment. It achieves high generalization by leveraging the powerful semantic understanding and cross-modal alignment of multimodal large language models (MLLM). By constructing a reasoning Chain of Thought (CoT), SATED can understand patch characteristics, perform global scans and semantic correlation analyses, provide detailed decision explanations, and generate prompts for patch inpainting.

Our contributions can be summarized as follows:

- We expose the inadequacy of existing evaluations for patch purification defenses and introduce a novel multidimensional evaluation framework that provides a holistic assessment of patch localization performance by measuring misidentification, recall, and localization accuracy. We believe this can offer a more comprehensive analytical perspective for the field.
- We present SATED, the first semantic-aware and explainable defense. SATED achieves zero-shot patch localization, false positives suppression, context-aware inpainting, and human-interpretable reasoning without any training or fine-tuning.
- We conduct extensive experiments on target object detection and semantic segmentation models in digital and physical attack scenarios, covering various adversarial patches and visual models. Our results demonstrate superior defense performance compared to state-of-the-art methods under multidimensional assessment.

2 Related Work

To counter the significant threat of adversarial patch attacks, researchers have developed several defense strategies to ensure model reliability. Based on their mechanisms, these methods can be categorized into four main types: 1) Adversarial Training (Gittings et al. 2020; Mao et al. 2024; Jia et al. 2023): Enhancing model robustness by injecting simulated adversarial patches during training. 2) Model Modification (Yu et al. 2021; Wang et al. 2023; Yu et al. 2023): Adjusting model architecture or internal feature processing to strengthen resistance to patch-induced anomalies. 3) Patch Purification (Liu et al. 2022; Kang et al. 2024; Ilina, Tereshonok, and Ziyadinov 2025): Preprocessing inputs to identify and remove patch regions before model inference. 4) Certifiably Robust Defenses (Xiang et al. 2021, 2023, 2024): Providing mathematically provable defense guarantees under specific threat models.

Of these, Patch Purification-based Defenses are advantageous for their model-agnostic nature, allowing for flexible deployment on pre-trained models. However, existing methods struggle with generalization and false positives, and they often suffer from an undesirable inpainting effect due to the disconnection between localization and inpainting processes. This work integrates patch reasoning, localization, and inpainting, effectively reducing false positives while ensuring generalization.

3 Patch Localization Evaluation Framework

To comprehensively assess the performance of adversarial patch localization, we establish a multi-dimensional localization evaluation framework, detailed in Table 1. Overcoming the limitations of single metrics, this framework addresses the common issue where a high recall rate can mask a high false positive rate in existing defenses. It provides a three-level evaluation: at the image, patch, and pixel levels.

3.1 False Positives

This dimension quantifies the occurrence of false positives in patch localization, measuring the extent to which normal regions are mistakenly identified as adversarial patches.

FP Pixel Ratio. This metric directly indicates the ratio of false positives within the localization results. By using the total number of predicted pixels as the denominator—rather than the total image pixels—this metric prevents the dilution of false positive rates in large images.

$$FPR_{pixel} = \frac{\sum_{i=1}^{N} FP_i}{\sum_{i=1}^{N} P_i}, \tag{1}$$

$$FP_i = |Pred_i \cap \overline{GT_i}|, P_i = |Pred_i|, \tag{2}$$

where $Pred_i$ represents the predicted patch mask of the i-th image, and GT_i represents its ground truth mask.

FP Image Ratio. To calculate false positive image statistics, we disregard minor false positive pixels that may arise from inaccuracies at the edges of the patch prediction mask. Specifically, we set a dynamic threshold: false positive pixels are ignored if their count is less than 10% of the true positive pixels.

$$FPR_{image} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(FP_i > 0.1 \times TP_i), \tag{3}$$

$$TP_i = |Pred_i \cap GT_i|. \tag{4}$$

Evaluation Dimension	Focus Area	Metric	Description
False Positives	Image usability damage level	FP Pixel Ratio	Ratio of false positive pixels to total predicted pixels
		FP Image Ratio	Proportion of images containing noticeable false positives
Recall Capability	Adversarial threat detection capability	Patch Recall	Ratio of recalled patches to total Ground Truth patches
		Recalled mIoA	Mean prediction coverage of successfully detected patches
Global Accuracy	Overall localization accuracy	mIoU	Mean IoU between predicted masks and GT masks
		mPrecision	Mean ratio of correctly predicted pixels to predicted pixels

Table 1: Our multidimensional patch localization evaluation framework, comprehensively measuring localization performance.

3.2 Recall Capability

This dimension assesses the ability to detect genuine adversarial patches, ensuring that defense methods can effectively identify threats.

Patch Recall. To better reflect the threat identification ability, we use patch-level recall rather than pixel-level recall, especially since a single image may contain multiple patches. To prevent the issue of multiple predicted patch masks merging, which could create an overly large denominator and unfairly penalize the evaluation of small patches, we use Intersection over Area (IoA) instead of the standard Intersection over Union (IoU).

$$R_{patch} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{G_i}\mathbb{I}(IoA_{ij}\geq 0.5)}{\sum_{i=1}^{N} G_i}, \tag{5}$$

$$IoA_{ij} = \frac{|GT_{ij} \cap Pred_i|}{|GT_{ij}|}, \tag{6}$$

where $\mathbb{I}$ is the indicator function (taking a value of 1 when the condition is met), GT_{ij} denotes the mask of the j-th adversarial patch added to the i-th image, and G_i represents the number of adversarial patches added to the i-th image.

Recalled mIoA. For all successfully detected patches, we calculate their mean IoA values to assess the average coverage of the predictions on them.

$$IoA_{recalled} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{G_i} IOA_{ij}.\mathbb{I}(IOA_{ij}\geq 0.5)}{\sum_{i=1}^{N}\sum_{j=1}^{G_i}\mathbb{I}(IOA_{ij}\geq 0.5)}. \tag{7}$$

3.3 Global Accuracy

This dimension evaluates the spatial alignment between the overall localization results and the ground truth patches.

mIoU. As the most common metric for assessing spatial overlap, IoU considers both false negatives and false positives. We combine all ground truth patches within an image into a single mask to evaluate the overall localization performance, rather than the accuracy of individual patches.

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{|GT_i \cap Pred_i|}{|GT_i \cup Pred_i|}. \tag{8}$$

mPrecision. As a complement to mIoU, mPrecision focuses on the reliability of the predicted regions. A high mPrecision score is particularly favorable for subsequent patch elimination steps.

$$mPrecision = \frac{1}{N}\sum_{i=1}^{N}\frac{|GT_i \cap Pred_i|}{|Pred_i|}. \tag{9}$$

4 SATED Method

4.1 Semantic-Aware Patch Reasoning CoT

As mentioned in the Introduction, we consider semantic understanding crucial for defending against adversarial patches. Building on the advancements in multimodal large language models (MLLMs), which have demonstrated significant improvements in fine-grained semantic understanding and cross-modal alignment, it is now possible to locate specific image regions using semantic descriptions (Lai et al. 2024; Ren et al. 2024; Liu et al. 2025). Inspired by this, we reframe adversarial patch localization as a cross-modal reasoning task. Our method, illustrated in Figure 2, uses a reasoning CoT to discover adversarial patch regions.

Although MLLMs have rich general knowledge, localizing adversarial patches accurately requires domain-specific knowledge. Traditional methods for injecting domain knowledge, such as fine-tuning with attack data or in-context learning, often lead to generalization bottlenecks and struggle with novel attacks. To overcome this, we enable MLLMs to identify patches by describing their key characteristics in text. Our description focuses on three aspects: 1) Limited area; 2) Behavioral objective; 3) Semantic independence and stylistic differences.

To address misidentification, we introduce a false positive suppression mechanism–Semantic Correlation Analysis–into our reasoning chain, achieving dynamic logical inference. For each potential patch identified during a global scan, we prompt the model to analyze its semantic content and its relationship with the surrounding context. If a correlation is found, the model re-evaluates whether the region is truly adversarial.

Finally, we constrain the model's output to a structured format that includes: 1) A Decision Explanation, representing the model's thought process ($< think >$); 2) Bounding boxes and points for confirmed adversarial patches; 3) An inpainting prompt to remove the patches.

The entire process can be formalized as follows: First, construct the prompt P_{chain} that embodies the reasoning chain,

$$P_{chain} = P_{\text{char}} \oplus P_{\text{scan}} \oplus P_{\text{SCA}} \oplus P_{\text{output}}, \tag{10}$$

where P_{char}, P_{scan}, P_{SCA} and P_{output} represent the adversarial patch characteristics description, global scan, semantic correlation analysis, and output constraints, and $\oplus$ denotes the concatenation. For each input image I_i, jointly input it with P_{chain} into the MLLM to obtain the output triplet:

$$\langle thinking_i, SP_i, IP_i\rangle = M\left(I_i, P_{chain}\right), \tag{11}$$

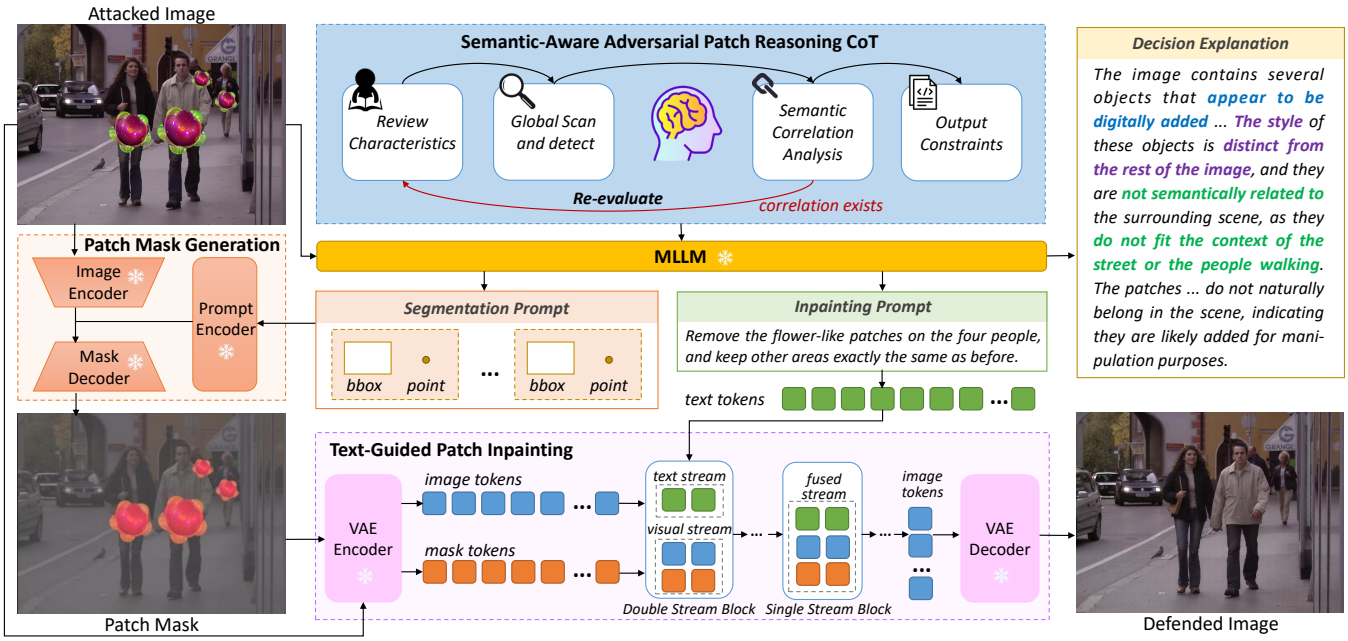

Figure 2: Overview of our proposed SATED. Zero-shot elimination of adversarial patches is achieved through a unified reasoning-segmentation-inpainting process. All model weights are frozen.

Where M represents the MLLM (such as Qwen2.5-VL (Bai et al. 2025)), IP_i represents the inpainting prompt, and SP_i represents the segmentation prompt, composed of several bounding box and point pairs.

$$SP_i = \left[\langle bbox_{i1}, point_{i1}\rangle, ..., \langle bbox_{im}, point_{im}\rangle\right]. \quad (12)$$

4.2 Patch Mask Generation

While the semantic understanding capabilities of MLLMs assist in locating adversarial patches, the varied sizes and shapes of these patches result in the need for variable coordinate lengths for precise localization, making direct mask output challenging. Therefore, during the patch reasoning phase, our model outputs only bounding boxes and points. A pre-trained Segment Anything Model (SAM) (Ravi et al. 2024) is then used in a subsequent step to generate the accurate segmentation mask.

$$Mask_i = S(I_i, SP_i), \quad (13)$$

where S represents the Segment Anything Model.

4.3 Text-Guided Patch Inpainting

After precisely localizing adversarial patches, another crucial step is to remove the patches from the image. Existing defenses utilize methods such as filling with a fixed color (Liu et al. 2022; Jing et al. 2024) and inpainting based on coherence transport (Tarchoun et al. 2023) to eliminate adversarial patches. However, the visual discontinuity and disharmony can still degrade model performance, affecting the effectiveness of defense.

To address this issue, we introduce the pretrained Diffusion Transformer (DiT) Model (Labs et al. 2025), turning patch removal into an in-context image editing problem. For each input image I_i, after obtaining the predicted mask of adversarial patches $Mask_i$ through previous steps, we input I_i and $Mask_i$ into a VAE Encoder to obtain encoded token sequences T_{image_i} and T_{mask_i}, respectively:

$$T_{image_i} = E(I_i), T_{mask_i} = E(Mask_i). \quad (14)$$

Simultaneously, the text instruction IP_i obtained from Patch Reasoning is encoded into a token sequence T_{text_i}. Next, the token sequences are concatenated and passed through double stream and single stream blocks, undergoing cross-attention to obtain the merged sequence $\left[T_{image_i}{}'; T_{mask_i}{}'; T_{text_i}{}'\right]$. Subsequently, the mask tokens and text tokens are discarded, and $T_{image_i}{}'$ is input into a VAE Decoder to obtain the final defended image:

$$I_{defended_i} = D(T_{image_i}{}'). \quad (15)$$

5 Experiments

5.1 Experimental Setups

Target Models.Being task-agnostic and model-agnostic, our defense is evaluated across different visual perception tasks and model architectures. We focus on the more complex and physically relevant tasks of object detection and semantic segmentation. For object detection, we use representative models pre-trained on the MS COCO dataset (Lin et al. 2014), including two-stage model Faster R-CNN (Ren et al. 2015), one-stage models YOLOv5 and YOLOv8, and Transformer-based model DETR (Carion et al. 2020). For

Detector	Defense	Clean	OBJ	CLS	P1	P2	P3	P4	P5	P6	Flower	Ivysaur	Texture
Faster R-CNN	Undefended	*97.1*	*57.9*	*70.6*	*66.4*	*74.9*	*56.4*	*74.9*	*67.0*	*51.6*	*79.7*	*87.2*	*77.7*
	LGS (WACV'19)	96.4	57.2	83.8	67.5	74.7	57.6	80.0	71.2	61.9	83.0	91.5	84.3
	SAC (CVPR'22)	**97.1**	81.8	86.2	68.1	74.9	56.3	79.7	67.5	52.2	80.6	87.5	83.0
	Jedi (CVPR'23)	96.6	60.1	70.8	66.7	74.8	58.0	75.1	67.1	52.6	79.7	87.1	87.4
	PAD (CVPR'24)	97.0	88.1	89.6	72.1	87.8	85.9	88.8	89.2	84.1	87.0	92.8	91.3
	NutNet (CCS'24)	97.0	72.1	67.8	64.6	72.8	54.5	69.1	64.8	51.3	72.8	88.4	83.5
	SATED (Ours)	97.0	**91.6**	**93.5**	**90.7**	**92.0**	**90.7**	**90.1**	**92.3**	**90.3**	**95.0**	**95.4**	**93.8**
YOLOv8n	Undefended	*96.4*	*56.7*	*75.3*	*68.5*	*50.9*	*51.8*	*65.8*	*51.6*	*50.1*	*94.2*	*87.9*	*84.9*
	LGS (WACV'19)	96.6	47.5	82.6	68.1	51.4	53.1	79.4	62.4	64.3	93.7	91.6	89.5
	SAC (CVPR'22)	96.4	81.9	87.0	69.9	51.0	51.8	78.2	53.5	51.0	94.1	87.9	89.3
	Jedi (CVPR'23)	96.4	59.0	75.4	68.6	51.1	52.3	65.6	51.7	50.5	93.9	88.1	89.4
	PAD (CVPR'24)	96.4	87.5	87.7	70.7	74.8	78.7	85.4	81.5	77.3	95.0	93.9	92.7
	NutNet (CCS'24)	96.8	78.3	73.7	65.7	50.9	50.5	70.8	54.2	51.0	86.2	89.7	87.6
	SATED (Ours)	**96.9**	**91.4**	**94.1**	**90.6**	**92.6**	**91.8**	**91.6**	**92.3**	**90.9**	**96.7**	**95.6**	**94.1**
DETR	Undefended	*94.8*	*76.3*	*84.2*	*79.9*	*80.7*	*71.4*	*82.0*	*74.2*	*77.9*	*90.8*	*86.7*	*83.1*
	LGS (WACV'19)	**95.0**	73.7	88.5	81.1	83.1	73.1	82.5	80.1	84.0	91.2	89.1	84.8
	SAC (CVPR'22)	94.8	83.8	88.1	80.9	80.7	71.4	83.5	74.6	78.2	90.4	86.9	84.3
	Jedi (CVPR'23)	94.8	77.7	84.5	79.9	80.9	71.9	82.2	74.4	78.1	90.7	87.1	82.9
	PAD (CVPR'24)	94.9	87.9	89.9	80.7	88.9	86.9	88.3	88.8	86.5	91.7	91.4	85.5
	NutNet (CCS'24)	90.8	73.7	77.8	76.3	79.0	70.8	77.6	74.3	76.3	77.8	81.1	77.0
	SATED (Ours)	94.5	**89.9**	**91.8**	**89.4**	**90.0**	**89.9**	**89.3**	**90.1**	**90.1**	**93.6**	**93.5**	**88.1**

Table 2: Detection mAP(%) after defenses against different adversarial patches. The best performance is **bolded**.

semantic segmentation, we use the BiseNet model (Yu et al. 2018), which has been targeted by previous attack studies.

Datasets. We evaluate our method on both digital and physical scenarios. For digital attacks, we use the widely-adopted INRIA Person dataset (Dalal and Triggs 2005) and CityScapes dataset (Cordts et al. 2016). For physical attacks, we use APRICOT dataset and custom-recorded videos.

Performed Patch Attacks. To validate SATED's generalization ability, we use a total of 14 adversarial patches generated by different attack methods (Thys et al. 2019; Hu et al. 2021; Tan et al. 2021; Hu et al. 2022; Huang et al. 2023; Nesti et al. 2022). These patches cover a wide range of sizes, shapes, styles, locations, and quantities.

Adaptive Attack. We design and verify an adaptive attack targeting our SATED, with almost no attack effect.

Compared Patch Defenses. We conduct a comprehensive comparison of SATED with five state-of-the-art defense methods: LGS (Naseer et al. 2019), SAC (Liu et al. 2022), Jedi (Tarchoun et al. 2023), PAD (Jing et al. 2024), and NutNet (Lin et al. 2024).

Implementation Details. We utilize Qwen2.5-VL (Bai et al. 2025) fine-tuned by Seg-Zero (Liu et al. 2025) as our MLLM, SAM2-Large (Ravi et al. 2024) as our segmentation model, and a quantized version of FLUX.1-fill (Labs et al. 2025) as the DiT model. Our experiments were conducted with 2 NVIDIA RTX 3090 GPUs. The seed for the DiT model was fixed to 0, and the final results were averaged over three runs.

5.2 Defense Effectiveness for Detection Task

For object detection task, we use mean Average Precision (mAP) at IoU 0.5 as the primary metric, consistent with existing defense works. The detection performance of different defenses against various adversarial patches on Faster R-CNN, YOLOv8, and DETR is presented in Table 2.

Table 2 shows that SATED demonstrates excellent generalization against a wide variety of adversarial patches, irrespective of their style or appearance. Our method boosts the average mAP to over 90% across all tested detectors, significantly improving the usability of attacked data. Furthermore, SATED exhibits high defense stability. Even for patches where other defenses show limited improvement (e.g., P1), SATED consistently elevates the mAP to over 90%, outperforming state-of-the-art methods by more than 20% on YOLOv8.

5.3 Patch Localization Performance

In addition to detection metrics, assessing patch localization performance is crucial as it can reveal issues not reflected in detection metrics. We use our newly developed Patch Localization Evaluation Framework to comprehensively evaluate the localization results of various defenses, with the results detailed in Table 3.

NutNet, while achieving a high Patch Recall against the Ivysaur and Flower attacks, suffers from a severe false positive problem. Its False Positive Image Ratio reaches 99.65%, indicating that nearly every image contains noticeable false positives. SAC, having not been trained on natural-looking adversarial patches, exhibits the lowest Patch Recall but fewer false positives. Jedi also shows a relatively low Patch Recall with a slightly higher false positive image ratio than SAC. However, Jedi has a high false positive pixel ratio, indicating that incorrect pixels constitute the majority of the predicted mask. PAD demonstrates a high Patch Recall but faces significant false positive issues, which greatly impact image usability.

In contrast, our SATED achieves high Patch Recall with the lowest false positive rate. The overall patch localization accuracy is more than doubled compared to state-of-the-art methods. Additionally, Figure 3 shows detailed distribution

Attack	Defense	False Positives		Recall Capability		Global Accuracy	
		FP Pixel Ratio↓	FP Image Ratio↓	Patch Recall↑	Recalled mIOA↑	mIOU↑	mPrecision↑
Ivysaur	SAC	7.03	6.25	0.43	73.26	0.34	19.39
	Jedi	75.42	13.89	7.14	77.87	2.16	3.57
	PAD	56.86	88.19	56.49	92.05	38.28	45.73
	NutNet	69.07	99.65	**65.23**	75.12	30.90	41.08
	SATED (Ours)	**3.70**	**4.86**	64.72	**92.39**	**77.19**	**86.35**
Flower	SAC	3.44	5.90	4.11	76.02	3.26	57.93
	Jedi	80.06	14.58	5.19	79.41	1.64	3.25
	PAD	61.53	87.50	55.19	79.43	27.89	40.22
	NutNet	63.61	99.65	**85.42**	83.15	37.74	44.79
	SATED (Ours)	**1.35**	**1.04**	84.82	**86.54**	**83.77**	**96.57**
P1	SAC	3.72	2.78	1.33	61.57	2.25	22.16
	Jedi	59.40	2.08	0.27	51.18	0.33	1.23
	PAD	58.76	93.06	70.97	75.48	32.48	44.05
	NutNet	89.94	95.14	7.72	81.97	4.69	16.25
	SATED (Ours)	**0.48**	**1.04**	**81.36**	**91.95**	**89.70**	**98.20**

Table 3: Comparison of Patch Localization Performance. SATED achieves low false positives, high recall, and high precision.

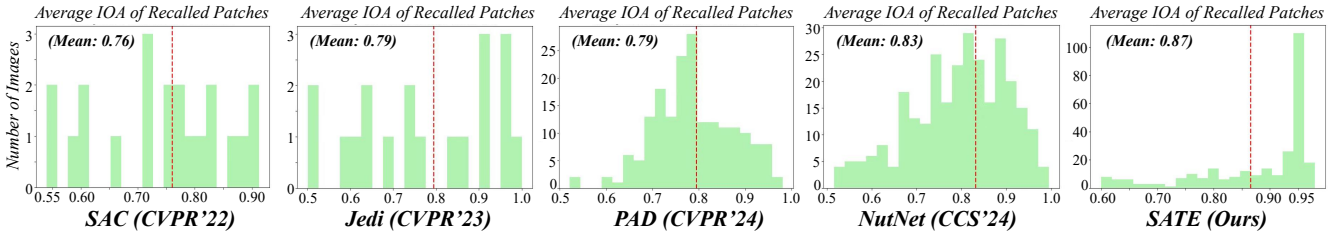

Figure 3: Comparison of Recalled mIoA distributions for different defenses.

Figure 4: Visualization examples for explainability and inpainting effects.

of the Recalled IoA for different defenses. While the average values may not differ significantly, there are notable variations in actual localization accuracy. For the majority of images, SATED achieves an average localization coverage of over 95%, significantly outperforming other defenses.

5.4 Explainability and Inpainting Demonstration

In addition to its strong patch localization and defense performance, SATED also demonstrates clear explainability, which existing defenses lack. Figure 4 shows two examples where SATED accurately localizes adversarial patches of varying sizes, shapes, styles, and quantities. It also outputs a detailed textual explanation of its decision-making process, highlighting the crucial role of semantic analysis. The defended images also demonstrate the excellent results of our patch inpainting method. Compared to existing defenses, SATED's inpainting results appear more natural, preserving the semantic coherence of the images.

5.5 Ablation Study

Ablation of Semantic Correlation Analysis module. To validate the role of the Semantic Correlation Analysis (SCA) module in reducing false positives within our patch reasoning CoT, we conduct an ablation study. Table 4 presents the false positive metrics for different patch attacks. It is evident that with the inclusion of SCA, both the False Positive Pixel Ratio and False Positive Image Ratio significantly decrease, demonstrating its effectiveness.

	SATED	Ivysaur	Shaymin	P2	P3
FP Pixel Ratio↓	w/o SCA	4.42	6.73	2.13	3.25
	w/ SCA	**3.70**	**2.51**	**1.52**	**1.22**
FP Image Ratio↓	w/o SCA	5.21	3.82	2.08	1.74
	w/ SCA	**4.86**	**2.43**	**2.08**	**0.35**

Table 4: Impact of SCA module on false positive ratio.

Defense	OBJ	P1	P2	P3	T-SEA
SATED-Black	89.90	83.80	79.30	79.60	77.20
SATED-NS	88.80	79.60	87.10	84.60	76.50
SATED-Ours	**91.40**	**90.60**	**92.60**	**91.80**	**80.80**

Table 5: Impact of inpainting method on mAP (%).

Attack patch size		300x600		150x300	
Metric		mIoU	mAcc	mIoU	mAcc
Defense	Undefended	40.63	52.73	55.41	65.82
	LGS	44.36	54.73	56.46	66.75
	SAC	50.1	60.45	58.49	72.31
	Jedi	43.84	57.35	51.72	65.01
	PAD	50.19	63.98	55.51	69.68
	NutNet	43.89	58.71	52.26	66.70
	SATED (Ours)	**52.83**	**65.35**	**62.01**	**72.32**

Table 6: Segmentation mIoU and mAcc (%) after defenses.

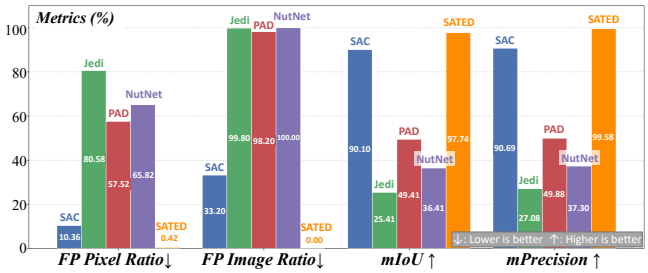

Figure 5: Patch localization performance on CityScapes.

Ablation of Inpainting method. To validate the effectiveness of our text-guided patch inpainting, we perform an ablation study on the inpainting method used after patch localization. We replace our inpainting with filling all black pixels (Liu et al. 2022; Jing et al. 2024) and the Navier-Stokes algorithm (Bertalmio, Bertozzi, and Sapiro 2001), and report the results in Table 5. Across different patch attacks, the defended images obtained using our inpainting method achieve significantly higher mAP scores.

5.6 Defense Effectiveness for Segmentation Task

To validate SATED's effectiveness for the semantic segmentation task, we conduct adversarial patch attacks on BiseNet using the CityScapes dataset and compare the key metrics mIoU and mAcc of the semantic segmentation models. As shown in Table 6, SATED achieves the best defense performance against patches of different sizes.

We also compare the patch localization performance of the different defenses. Since these attacks generate localized noise patches that are relatively easy to locate, all defense

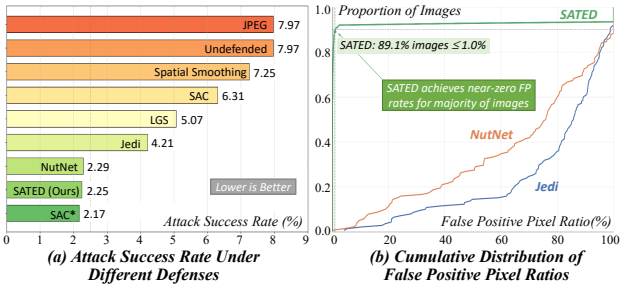

Figure 6: Comparison of performance on APRICOT.

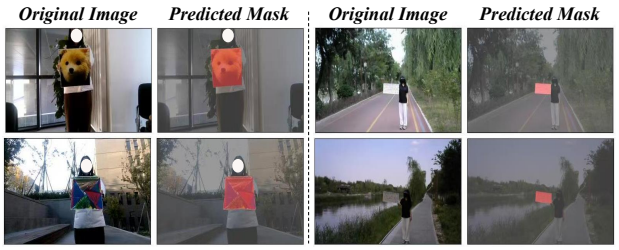

Figure 7: Patch localization results on captured videos.

methods achieved a 100% Patch Recall. Therefore, we focus on comparing the false positive ratios and overall localization accuracy, as detailed in Figure 5. It shows that SATED achieves the highest localization accuracy with its FP Pixel Ratio and FP Image Ratio being essentially zero.

5.7 Defense Against Physical Attacks

Evaluation on APRICOT. Following previous works, we use Faster R-CNN to evaluate the Attack Success Rate (ASR) after defense on APRICOT. As shown in Figure 6(a), SATED reduces the ASR to 2.25% without any training, second only to the specifically trained SAC. We also conduct a detailed false positive analysis, plotting the Cumulative Distribution Function (CDF) for the FP Pixel Ratio, as shown in Figure 6(b). SATED shows a significant advantage, with its curve rising almost vertically near 0%, indicating that the vast majority of images have zero false positive pixels.

Evaluation on captured videos. We evaluated SATED on captured physical videos. For both object detection and semantic segmentation, we printed various adversarial patches and recorded physical attack videos in different scenarios. Figure 7 shows several examples of our localization results, demonstrating SATED can accurately locate different types of adversarial patches across various physical scenes.

6 Conclusion

In this work, we address a critical, yet overlooked problem in current defense works: false positive patch localization. To tackle this, we propose a comprehensive patch localization evaluation framework and a novel defense method that demonstrates superior localization accuracy, defense generalization, and explainability across various scenarios.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China Under Grants No.62176253.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bertalmio, M.; Bertozzi, A. L.; and Sapiro, G. 2001. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, I–I. IEEE.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, 886–893.

Gittings, T.; et al. 2020. Vax-a-net: Training-time defence against adversarial patch attacks. In *Proceedings of the Asian Conference on Computer Vision*.

Hu, Y.-C.-T.; Kung, B.-H.; Tan, D. S.; Chen, J.-C.; Hua, K.-L.; and Cheng, W.-H. 2021. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7848–7857.

Hu, Z.; Huang, S.; Zhu, X.; Sun, F.; Zhang, B.; and Hu, X. 2022. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13307–13316.

Huang, H.; Chen, Z.; Chen, H.; Wang, Y.; and Zhang, K. 2023. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20514–20523.

Ilina, O.; Tereshonok, M.; and Ziyadinov, V. 2025. Increasing Neural-Based Pedestrian Detectors' Robustness to Adversarial Patch Attacks Using Anomaly Localization. *Journal of Imaging*, 11(1): 26.

Jia, Y.; Poskitt, C. M.; Zhang, P.; Wang, J.; Sun, J.; and Chattopadhyay, S. 2023. Boosting adversarial training in safety-critical systems through boundary data selection. *IEEE Robotics and Automation Letters*, 8(12): 8350–8357.

Jing, L.; Wang, R.; Ren, W.; Dong, X.; and Zou, C. 2024. PAD: Patch-Agnostic Defense against Adversarial Patch Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24472–24481.

Kang, C.; Dong, Y.; Wang, Z.; Ruan, S.; Chen, Y.; Su, H.; and Wei, X. 2024. Diffender: Diffusion-based adversarial defense against patch attacks. In *European Conference on Computer Vision*, 130–147. Springer.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context.

Lin, Z.; Zhao, Y.; Chen, K.; and He, J. 2024. I don't know you, but I can catch you: Real-time defense against diverse adversarial patches for object detectors. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 3823–3837.

Liu, J.; Levine, A.; Lau, C. P.; Chellappa, R.; and Feizi, S. 2022. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14973–14982.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Mao, Z.; Chen, S.; Miao, Z.; Li, H.; Xia, B.; Cai, J.; Yuan, W.; and You, X. 2024. Enhancing robustness of person detection: A universal defense filter against adversarial patch attacks. *Computers & Security*, 146: 104066.

Naseer, M.; et al. 2019. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1300–1307. IEEE.

Nesti, F.; Rossolini, G.; Nair, S.; Biondi, A.; and Buttazzo, G. 2022. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2280–2289.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. 28.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.

Tan, J.; Ji, N.; Xie, H.; and Xiang, X. 2021. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *Proceedings of the 29th ACM international conference on multimedia*, 5307–5315.

Tarchoun, B.; Ben Khalifa, A.; Mahjoub, M. A.; Abu-Ghazaleh, N.; and Alouani, I. 2023. Jedi: Entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4087–4095.

Thys, S.; et al. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Wang, Z.; Wang, B.; Zhang, C.; and Liu, Y. 2023. Defense against adversarial patch attacks for aerial image semantic segmentation by robust feature extraction. *Remote Sensing*, 15(6): 1690.

Xiang, C.; Bhagoji, A. N.; Sehwag, V.; and Mittal, P. 2021. {PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security 21)*, 2237–2254.

Xiang, C.; Valtchanov, A.; Mahloujifar, S.; and Mittal, P. 2023. Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In *2023 IEEE Symposium on Security and Privacy (SP)*, 1329–1347. IEEE.

Xiang, C.; Wu, T.; Dai, S.; Petit, J.; Jana, S.; and Mittal, P. 2024. {PatchCURE}: Improving Certifiable Robustness, Model Utility, and Computation Efficiency of Adversarial Patch Defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, 3675–3692.

Yu, C.; Chen, J.; Wang, Y.; Xue, Y.; and Ma, H. 2023. Improving adversarial robustness against universal patch attacks through feature norm suppressing. *IEEE Transactions on Neural Networks and Learning Systems*.

Yu, C.; Chen, J.; Xue, Y.; Liu, Y.; Wan, W.; Bao, J.; and Ma, H. 2021. Defending against universal adversarial patches by clipping feature norms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16434–16442.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.