AngleRoCL: Angle-Robust Concept Learning for Physically View-Invariant T2I Adversarial Patches

Wenjun Ji[1,2] **Yuxiang Fu**[1,2] **Luyang Ying**[1,2] **Deng-Ping Fan**[1,2]
Yuyi Wang[3] **Ming-Ming Cheng**[1,2] **Ivor Tsang**[4] **Qing Guo**[1,2]*
[1]NKIARI, Shenzhen Futian [2]VCIP, CS, Nankai University [3]CRRC Zhuzhou Institute
[4]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR)
wenjj@mail.nankai.edu.cn tsingqguo@ieee.org

Abstract

Cutting-edge works have demonstrated that text-to-image (T2I) diffusion models can generate adversarial patches that mislead state-of-the-art object detectors in the physical world, revealing detectors' vulnerabilities and risks. However, these methods neglect the T2I patches' attack effectiveness when observed from different views in the physical world (*i.e.*, angle robustness of the T2I adversarial patches). In this paper, we study the angle robustness of T2I adversarial patches comprehensively, revealing their angle-robust issues, demonstrating that texts affect the angle robustness of generated patches significantly, and task-specific linguistic instructions fail to enhance the angle robustness. Motivated by the studies, we introduce Angle-Robust Concept Learning (AngleRoCL), a simple and flexible approach that learns a generalizable concept (*i.e.*, text embeddings in implementation) representing the capability of generating angle-robust patches. The learned concept can be incorporated into textual prompts and guides T2I models to generate patches with their attack effectiveness inherently resistant to viewpoint variations. Through extensive simulation and physical-world experiments on five SOTA detectors across multiple views, we demonstrate that AngleRoCL significantly enhances the angle robustness of T2I adversarial patches compared to baseline methods. Our patches maintain high attack success rates even under challenging viewing conditions, with over 50% average relative improvement in attack effectiveness across multiple angles. This research advances the understanding of physically angle-robust patches and provides insights into the relationship between textual concepts and physical properties in T2I-generated contents. We released our code in https://github.com/tsingqguo/anglerocl.

1 Introduction

Recent advances in deep learning have revealed that text-to-image (T2I) diffusion models can generate adversarial patches capable of misleading state-of-the-art object detectors in physical environments [48, 58]. Unlike traditional optimization-based adversarial patch generation methods [4, 23, 62] that rely on gradient-based pixel manipulations and often struggle to transfer to physical settings and detection models due to environmental factors like printer's color shifting, T2I generation approaches offer significant advantages (*e.g.*, low-cost, model-agnostic, and transferable) in the physical deployment [48, 58]. These patches exploit vulnerabilities in detection systems, creating significant security concerns for critical applications such as autonomous driving and surveillance. However, existing approaches to generating T2I adversarial patches have largely overlooked a crucial real-world challenge: maintaining attack effectiveness when patches are viewed from different angles

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

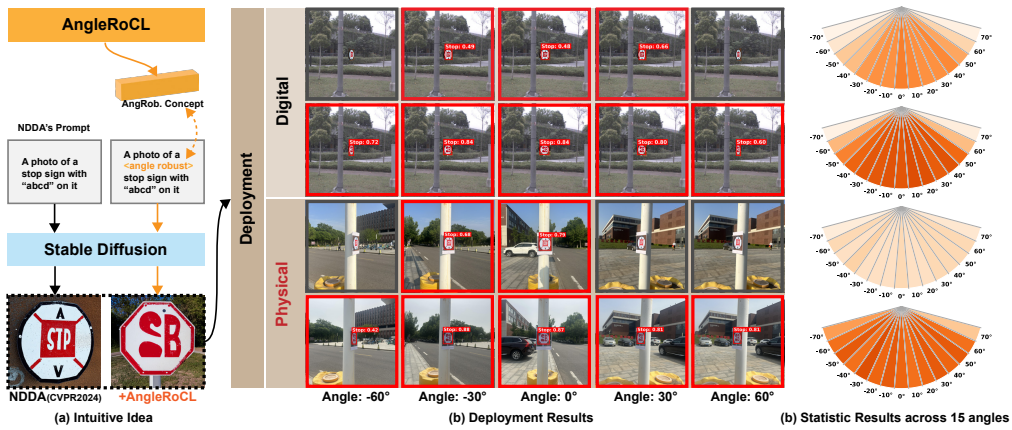

Figure 1: (a) Intuitive idea of AngleRoCL: incorporating the learned angle-robust concept into textual prompts to generate patches that maintain attack effectiveness across multiple viewing angles. (b) Digital and physical deployment results AngleRoCL patches (bottom rows) vs. NDDA baseline (top rows). (c) Statistical results across 15 angles, where orange intensity represents attack success rates.

in physical environments. As shown in Fig. 1, NDDA's patches can only mislead the stop sign detector (YOLOv5) at the front angles while failing at other angles in both digital and physical settings.

In practical scenarios, adversarial patches are typically observed from various viewpoints, making angle robustness essential for sustained attack effectiveness. Current approaches predominantly generate patches that demonstrate attack capabilities only under fixed or limited viewing angle ranges, significantly limiting their effectiveness in real-world applications. For example, patches designed to mislead stop sign detection may work when viewed head-on but fail when observed from oblique angles, reducing their practical impact.

In this paper, we first conduct a comprehensive investigation into the angle robustness of T2I adversarial patches. Our study reveals several important findings: ❶ the angle robustness of T2I adversarial patches varies significantly depending on the textual prompts used; ❷ simply augmenting prompts with task-specific linguistic instructions (*e.g.*, "detectable at multiple angles in all directions") fails to enhance angle robustness; and ❸ certain robust feature-related prompt elements have greater influence on angle robustness than others. Please refer to Sec. 3 for details.

Motivated by these insights, we introduce angle-robust concept learning (AngleRoCL), a novel approach that learns a concept representing the capability to generate angle-robust adversarial patches. Unlike previous methods that require specific optimization for each patch/environment, our approach encodes angle robustness as a learned concept in the embedding space of T2I models. The learned concept can then be incorporated into any subject-related textual prompt, guiding diffusion models to generate patches with attack effectiveness that inherently remains robust across multiple viewing angles. AngleRoCL offers several key advantages over existing approaches: ❶ environment-free learning that eliminates the need for user-provided environment images; ❷ detector-guided optimization that leverages feedback from target detectors to supervise the concept learning process; and ❸ consistent attack effectiveness maintained across diverse viewing angles. Through extensive experiments in both digital and physical environments, we demonstrate that AngleRoCL significantly enhances the angle robustness of T2I adversarial patches compared to baseline methods (See Fig. 1 (a)), achieving about relative improvement of 58.96% in digital environments and 82.41% in physical environments in attack effectiveness across multiple viewing angles. This research advances the understanding of physically robust adversarial patches and provides valuable insights into the relationship between textual concepts and physical properties in T2I-generated content.

2 Related Works

Physical adversarial attacks. Latest research indicates that deep neural network (DNN)-based object detection systems are vulnerable to adversarial patches that can induce erroneous outputs [13, 17, 36, 39, 41, 51, 19, 16, 60, 28, 30]. Physical adversarial patches pose severe threats to critical systems, including autonomous driving [55], classification models [5, 13, 63], and other safety-critical

domains [9, 10, 20, 50, 52]. These patches are highly reproducible and deployable [51], making it essential to investigate their attack effectiveness [5, 52]. Given that these physical adversarial patches are highly reproducible and deployable [48], their attack effectiveness [5, 52, 22, 18], stealthiness [11, 23, 26, 54, 25], and scenario plausibility [6, 58] warrant in-depth investigation. However, achieving efficient physical adversarial attacks remains challenging. Taking physical attacks targeting stop sign as a representative case, existing studies predominantly generate adversarial patches that only demonstrate effective attack capabilities under fixed or narrow viewing angle ranges, thereby exhibiting constrained attack effectiveness in practical applications. To address this limitation, our research focuses on generating angle-robust adversarial patches capable of maintaining attack efficacy across multiple viewing angles in physical deployment scenarios.

T2I generation for attacks. Traditional adversarial attacks predominantly operate within the digital domain, conducting targeted manipulations by introducing perturbations to image [7, 17, 39, 40, 41, 51] or by altering pixel values [51]. While these perturbations maintain visual imperceptibility to human observers, their efficacy degrades significantly when deployed in physical environments due to environmental factors [1, 35, 33, 59]. Recent advancements in diffusion models [21, 43, 45, 47] have enabled researchers to employ text-to-image (T2I) models for directly generating adversarial patches [38, 56, 34, 37, 31, 32]. Among them, the Natural Denoising Diffusion (NDD) attack [48, 58] achieves superior attack effectiveness compared to traditional methods in the physical world due to non-robust features that are predictive but incomprehensible to humans.

Angle robustness enhancements. The deployment of adversarial patches in practical physical environments necessitates systematic consideration of various complex environmental factors, with detector viewing angle variation constituting a critical influencing parameter. Previous research on angle robustness primarily focused on 3D object camouflage domains, where certain methods attempted to generate or modify surface textures of physical 3D objects to achieve multi-view disguise effects [2, 8, 12, 27, 42, 49, 64]. However, these specially engineered textures suffer from critical limitations including overfitting to specific viewing angles, poor transferability across objects, and visually unnatural appearances [8, 24]. In the 2D domain, existing studies primarily concentrate on evasion attacks - such as attaching adversarial stickers to existing objects to disrupt detector recognition processes [13, 57]. Some approaches have explored transformation-aware optimization frameworks that incorporate diverse geometric and photometric variations to enhance adversarial patch robustness in real-world deployments [5, 3, 49], yet demonstrate inadequate angular robustness [9, 62]. To our knowledge, no prior work has formally conceptualized angular robustness for adversarial patches, nor successfully integrated this property into practical patch generation frameworks for physical deployment scenarios.

3 Preliminaries and Discussions

3.1 T2I Adversarial Patches

The recent work [48] demonstrates that text-to-image (T2I) models can generate patches capable of misleading object detectors in the physical world by simply modifying input text prompts to remove robust features. This operational pipeline can be formalized as follows: given a textual prompt $\mathcal{T}$ describing both the target subject (*e.g.*, stop sign) and its associated robust features (e.g., shape, color, text, and pattern), a stable diffusion model $\mathcal{M}(\cdot)$ processes $\mathcal{T}$ to generate a patch $\mathbf{P}$, which can be formulated as $\mathbf{P} = \mathcal{M}(\mathcal{T})$. The work demonstrates that by modifying the textual prompt $\mathcal{T}$ to remove robust features, the T2I model generates patches that can effectively mislead various state-of-the-art object detectors in a black-box manner in physical-world scenarios. We denote the patch as *T2I adversarial patches* to distinguish them from conventional adversarial patches.

3.2 Empirical Studies on Angle Robustness

Angle robustness testing. Although effective, previous works [48, 58] overlook the angle robustness of T2I adversarial patches—their effectiveness when viewed from different angles. To address this gap, we comprehensively evaluate the angle robustness of T2I adversarial patches. Specifically, we selected "stop sign" as our test subject, with the dual objective of generating patches that ❶ appear non-suspicious to human observers and ❷ cause detectors to misclassify them as a stop sign. Following the NDDA method [48], we constructed a prompt set $\{\mathcal{T}_i | i \in [1, ..., 15]\}$ including 15

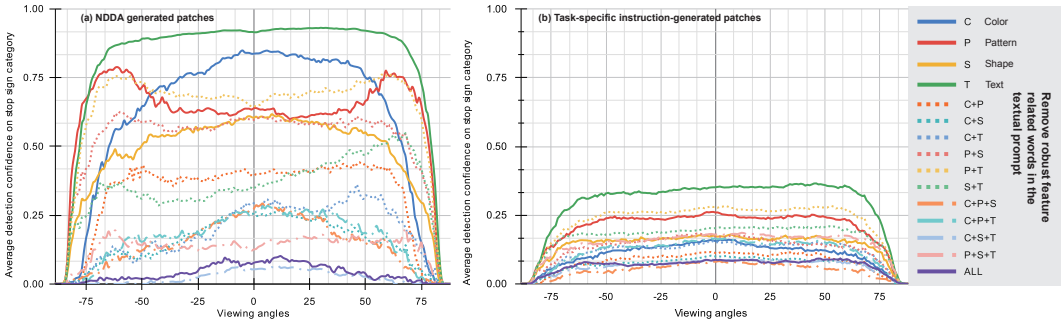

Figure 2: Average detection confidence (*i.e.*, $\mathcal{R}_\theta(\mathcal{T}_i)$) at different viewing angles (*i.e.*, θ) across image sets (*i.e.*, $\mathbf{I}_\theta^{(i,j)}$) generated by given attack strategy ($i \in [1, ..., 15], j \in [1, ..., K], \theta \in (-90°, 90°), \nabla\theta = 1°$). The y-axis labels indicate average detection confidence and the x-axis labels indicate different viewing angles. The rightmost panel demonstrates removed features: Color (C), Pattern (P), Shape (S), and Text (T). Patch generated by (a) NDDA prompt and (b) Task-specific instruction prompt.

types of prompt by removing different robust features from benign stop sign descriptions [2].See supp. for more details. For each prompt $\mathcal{T}_i$, we generated K patches $\{\mathbf{P}_i^j | j \in [1, ..., K]\}$. For each patch $\mathbf{P}_i^j$, we inserted it into an environment and captured images $\mathbf{I}_\theta^{(i,j)}$ from various angles θ, spanning the range $\theta \in (-90°, 90°)$ with a sampling interval of $\nabla\theta = 1°$. An angle of zero (*i.e.*, $\theta = 0°$) represents a frontal view (See Fig. 1 (b)). We then fed each image $\mathbf{I}_\theta^{(i,j)}$ into an object detector and recorded the false-positive confidence score $s_\theta^{(i,j)}$ (*i.e.*, probability of misclassification as "stop sign"). This procedure generated 180 test images per patch $\mathbf{P}_i^j$, resulting in $K \times 180$ evaluations per prompt $\mathcal{T}_i$. We defined the angle robustness metric for prompt $\mathcal{T}_i$ at viewing angle θ as the average of $s_\theta^{(i,j)}$ across K adversarial patches, *i.e.*, $\mathcal{R}_\theta(\mathcal{T}_i) = \frac{1}{K}\sum_{j=1}^{K} s_\theta^{(i,j)}$. We evaluated the 15 prompt types by setting $K = 50$ and digitally inserting the generated patches into environmental images.

We show the visualization of $s_\theta^{(i,j)}$ in Fig. 2 (a), and observe that: ❶ Most prompts $\mathcal{T}_i$ display centripetal confidence increase across 180° viewing angles ; ❷ Attack efficacy roughly presents an inverse correlation with the number of ablated features ; ❸ For $\mathcal{T}_i$ that removed the same counts of features, the performance differences are still significant ; ❹ Only a minority of prompts achieve statistically significant attack success. The experimental results demonstrate significant variance in attack efficacy across different prompts $\mathcal{T}_i$ under varying viewing angles.

Task-specific instruction for angle robustness enhancement. To investigate potential improvements, we implemented another attack strategy: augmenting the original prompts in $\{\mathcal{T}_i | i \in [1, ..., 15]\}$ with task-oriented narrative instructions, such as appending the phrase "detectable at multiple angles in all directions". Three modified prompt configurations were engineered: ❶ Prefix-enhanced: $\{\mathcal{T}_p + \mathcal{T}_i\}$. ❷ Infix-integrated: $\{\mathcal{T}_i + \mathcal{T}_m + \mathcal{T}_i\}$. ❸ Suffix-appended: $\{\mathcal{T}_i + \mathcal{T}_s\}$. All other experimental parameters remained unchanged except sample size K ($\theta \in (-90°, 90°), \nabla\theta = 1°$). As illustrated in Fig. 2 (b), the augmented prompts exhibited significant degradation in angular robustness metrics. This empirical evidence suggests that simply incorporating task-specific linguistic instructions fails to enhance the angle robustness. Consequently, we conclude that current T2I text encoders lack the capacity to interpret abstract, goal-oriented narrative commands for angle-robust purposes.

4 Methodology: Angle-Robust Concept Learning (AngleRoCL)

Our analyses show different textual prompts significantly affect angle robustness, while simple task-specific linguistic instructions fail to enhance this property. To address this limitation, we propose AngleRoCL to encode angle robustness as a latent concept, which can be plugged into textual prompts and guide T2I to generate patches that maintain attack efficacy across viewpoint variations while preserving physical deployability. Fig. 3 shows our AngleRoCL's pipeline.

[2]In NDDA, we have an original prompt for "stop sign", *i.e.*, "a photo of a stop sign". We can remove robust features (*i.e.*, color (C), pattern (P), Shape (S), and Text (T)) by adding constraints into the standard description.

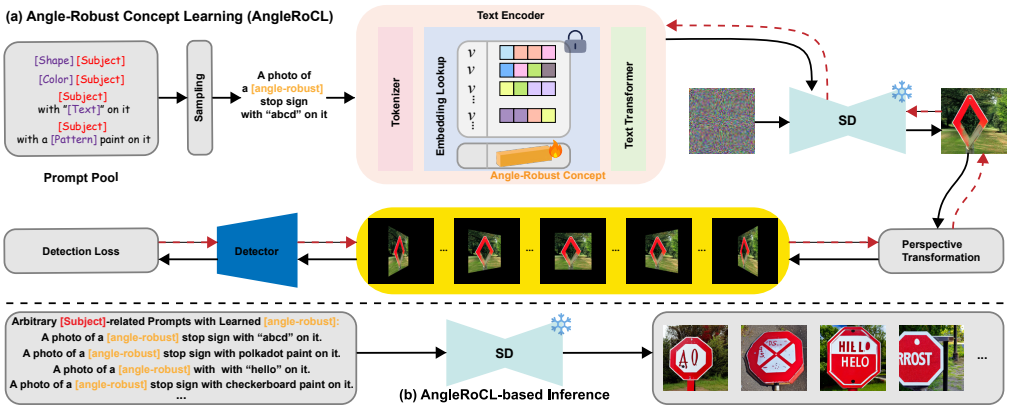

Figure 3: (a) Pipeline of our angle-robust concept learning (AngleRoCL). The latent code represents the learned concept <angle-robust>. (b) shows the inference results when we use the learned concept.

4.1 Problem Formulation

Given a pre-trained stable diffusion model $\mathcal{M}(\cdot)$, we generate an adversarial patch by feeding the model with a texture prompt $\mathcal{T}$ containing the subject and related constraints, such that $\mathbf{P} = \mathcal{M}(\mathcal{T})$. When this generated patch $\mathbf{P}$ is inserted into an environment denoted as E, it produces an image from observation angle θ, that is, $\mathbf{I}_\theta = \text{Obs}(\mathbf{P}, \text{E}, \theta)$. Our objective is to ensure the patch consistently misleads object detectors across various viewing angles. Intuitively, following the existing solution, we can add optimized adversarial noise to the generated patch to achieve the goal [62]. However, such a solution needs to perform the optimization for each patch, and the optimized noise can hardly be transferred into the physical world and keep attack effectiveness against different detectors. In this work, we propose to learn a concept to represent the capability of generating angle-robust adversarial patches, and the learned concept could be defined through new "words" in the embedding space and be inserted into any existing subject-related prompts to generate angle-robust patches. We can formulate the angle robust concept learning (AngleRoCL) as

$$\mathcal{C} = \arg\min_{\mathcal{C}_*} \mathbb{E}_{\theta \sim (-90°, 90°)} \mathcal{L}_{\text{det}}(\text{Obs}(\mathbf{P}, \text{E}, \theta), y) \text{d}\theta, \text{subject to}, \mathbf{P} = \mathcal{M}(\Lambda(\mathcal{T}, \mathcal{C}_*)), \tag{1}$$

where $\mathcal{C}$ denotes the words representing the learned angle-robust concept and $\Lambda(\mathcal{T}, \mathcal{C}_*)$ is the function to insert the concept words into the textual prompt. The loss function $\mathcal{L}_{\text{det}}(\cdot)$ measures whether the attack objective indicated by y has been achieved, for example, the generated patch is misdetected as "Stop Sign". Once we obtain $\mathcal{C}$, we can incorporate it into any subject-related textual prompts to generate patches that maintain robust performance across multiple viewing angles (See Fig. 3 (b)).

4.2 AngleRoCL as the Angle-aware Textual Inversion

Textual inversion was proposed for personalizing diffusion-based generation, enabling the model to learn a specific concept of an object or style from user-provided images [15]. This technique allows for consistent regeneration of the same object by simply incorporating the concept's learned token embedding into the generation model's input prompt. Inspired by this approach, we formulate angle-robust concept learning as a textual inversion problem. Unlike previous work, AngleRoCL should have the following key properties: ❶ **Environment-free learning.** Rather than requiring user-provided images of a concept, our learned angle-robust concept remains effective across any environment without relying on specific environmental images. ❷ **Detector-guided optimization.** AngleRoCL should leverage feedback from the target detector to directly supervise the optimization of the angle-robust concept. ❸ **Consistent attack effectiveness across viewing angles.** During the learning process, we optimize not only for image quality but specifically for adversarial performance maintained across multiple angular perspectives. *Our method should uniquely combine angle-awareness with environment-free concept learning through detector feedback across diverse observation angles.*

Specifically, we start with the input textual prompt with the $\Lambda(\mathcal{T}, \mathcal{C})$ where we tend to insert the targeted $\mathcal{C}$ (*i.e.*, "angle-robust") into an existing textual prompt $\mathcal{T}$. Then, we have a text encoder like CLIP to extract the embeddings of $\Lambda(\mathcal{T}, \mathcal{C})$ and get $\mathbf{F}_\mathcal{T}$ and $\mathbb{F}_\mathcal{C}$. Here, we aim to learn a new

embedding of $\mathcal{C}$ to replace the original one (*i.e.*, $\mathbb{F}_{\mathcal{C}}$), which denotes the angle-robust concept and should make the subsequent patch generation angle-robust. Then, we can reformulate the Eq. (1)

$$\mathbf{F}_{\mathcal{C}} = \arg\min_{\mathbf{F}_{\mathcal{C}*}} \mathbb{E}_{\theta\sim(-90^\circ,90^\circ)} \mathcal{L}_{\text{det}}(\text{Obs}(\mathbf{P}, \text{E}, \theta), y), \text{subject to}, \mathbf{P} = \mathcal{M}([\mathbf{F}_{\mathcal{T}}, \mathbf{F}_{\mathcal{C}*}]). \tag{2}$$

However, it requires high costs to sample different angles during the optimization. Hence, we perform the projective transformation on the image captured at the angle $\theta = 0$ to produce the images captured at other angles. Then, the Eq. (2) is reformulated as

$$\mathbf{F}_{\mathcal{C}} = \arg\min_{\mathbf{F}_{\mathcal{C}*}} \mathbb{E}_{\theta\sim(-90^\circ,90^\circ)} \mathcal{L}_{\text{det}}(\text{Proj}(\mathbf{I}_{0^\circ}, \theta), y), \tag{3}$$

$$\text{subject to}, \mathbf{I}_{0^\circ} = \text{Obs}(\mathbf{P}, \text{E}, 0^\circ), \mathbf{P} = \mathcal{M}([\mathbf{F}_{\mathcal{T}}, \mathbf{F}_{\mathcal{C}*}]), \tag{4}$$

where $\text{Proj}(\cdot)$ denotes the projective transformation corresponding to the observation angle.

Angle robustness loss. We adopt a loss function $\mathcal{L}_{\text{det}}(\cdot)$ to maximize the adversarial effect across multiple viewing angles, which is defined as follows

$$\mathcal{L}_{\text{det}}(\mathbf{I}_\theta, y) = \max(y - \text{Det}(\mathbf{I}_\theta), 0) \cdot \lambda, \tag{5}$$

where $\text{Det}(\mathbf{I}_\theta)$ is the confidence score of a detector on the interested category (*e.g.*, "stop sign"). Here, we use YOLOv5. y is the detection threshold and λ is a scaling factor. This loss penalizes viewing perspectives where detection confidence falls below the threshold, encouraging wide-angle effectiveness. By minimizing this loss across different angles, we optimize the $\mathbf{F}_{\mathcal{C}}$ to guide the patch generation that maintains high detection confidence across all viewing angles.

Angle-robust patch generation. Once the angle robustness concept is learned and encapsulated in the `<angle-robust>` token, generating angle-robust adversarial patches becomes remarkably straightforward. Our method's key advantage lies in its plug-and-play nature, requiring minimal modifications to existing text-to-image workflows. To generate angle-robust patches, users simply need to load our trained embedding into Stable Diffusion and incorporate the `<angle-robust>` placeholder into their existing attack prompts. For example, a standard NDDA prompt like "a blue square stop sign" can be enhanced to "a `<angle-robust>` blue square stop sign" to produce a patch with inherent angle robustness. This approach is compatible with various Natural Denoising Diffusion (NDD) attack frameworks, including NDDA [48] and MAGIC [58], without requiring any architectural modifications or additional optimization steps.

4.3 Implementation Details and Physical Deployment

Implementation details. We implement our approach using the Stable Diffusion v1.5 as the base diffusion model with DPMSolver++ for denoising. We set the classifier-free guidance scale to 7.5 and use 25 denoising steps. For the angle-robust concept, we use CLIP embedding of `<angle-robust>` as the initialization. Focusing on the stop sign category, we utilize a total of 39 NDDA prompt templates that incorporate various robustness features including shape (*e.g.*, square, triangle), color (*e.g.*, blue, yellow), text (*e.g.*, "hello", "abcd", "world"), and pattern (checkerboard, polkadot). Examples include "a blue square stop sign", "a stop sign with 'hello' on it", and "a yellow triangle stop sign with polkadot paint on it" (see supp. for more details). During the training process, we sample 9 angles, *i.e.*, $\{-72^\circ, -54^\circ, -36^\circ, -18^\circ, 0^\circ, 18^\circ, 36^\circ, 54^\circ, 72^\circ\}$, which are symmetrically and equally spaced within the range from -90° to 90° (excluding the endpoints). The detection loss parameter y and the scaling factor λ are respectively set to 0.8 and 10. We train for 50,000 steps using AdamW optimizer with learning rate 10^{-4}, updating only the `<angle-robust>` embedding while keeping all other parameters frozen.

Physical deployment. The physical deployment of our angle-robust patches follows a straightforward process validating their real-world effectiveness. We directly print the generated adversarial patches using a standard color printer on regular office paper, without requiring specialized printing techniques or materials. These patches are then affixed to target objects (stop signs) in various environments.

5 Experimental Results

5.1 Setups

Digital & Physical environments. Following the evaluation protocol in [58], we adopt the nuImage dataset [14], selecting one representative image from each of the six car-mounted camera views (front,

front left, front right, back, back left, back right) for digital evaluation. We also validate AngleRoCL in real-world scenarios, which are collected by ourselves. The main paper presents results from one physical environment, while cross-scene validation across three physical environments is provided in the supplementary material. See supp. for more details.

Baseline methods. To demonstrate the effectiveness of AngleRoCL in improving angle robustness for physical adversarial patches, we establish comparisons with four baseline methods: one traditional physical adversarial patch approach (AdvPatch [5]) and two NDD attacking methods (NDDA [48] and MAGIC [58]). ❶ For the traditional methods, we generate AdvPatch with Adversarial Robustness Toolbox , selecting only the highest-performing patch from each method for evaluation. ❷ For NDD attacking methods, we follow the methodology outlined in previous work [58], generating 50 patches for each "remove text or pattern" text prompt, and randomly selecting 100 patches as our NDDA baseline set. ❸ For MAGIC, we utilize its generation component to produce 100 patches for each digital environment using identical prompt configurations to ensure fair comparison. To isolate the effect of our proposed approach, NDDA+AngleRoCL and MAGIC+AngleRoCL patches were generated concurrently with the same seed and generator. The only difference in the generation process was the inclusion of our angle-robust embedding in the text prompt. We extend our evaluation from digital to physical environments, acknowledging the higher cost and complexity of physical experimentation. For physical validation, we selected 25 matched pairs of patches from each method (NDDA, NDDA+AngleRoCL, MAGIC, and MAGIC+AngleRoCL).

Generator & Detectors. For consistency with prior work [48, 58], we employ Stable Diffusion v1.5 [46] as our image generator. Our evaluation spans multiple object detection architectures, including YOLOv5 [29], YOLOv3 [44], Faster R-CNN [44], and RT-DETR [61]. We additionally evaluate against YOLOv10 [53], a more robust contemporary detector, as our primary benchmark. For traditional adversarial patches, we use the same detectors for both training and evaluation as in their original works to ensure fair comparison. For implementation frameworks, YOLOv5 and YOLOv10 are evaluated using the API from ultralytics, while Faster R-CNN, YOLOv3, and RT-DETR are implemented through the MMDetection framework to ensure consistent evaluation procedures.

Evaluation metrics. Previous studies have primarily relied on **Attack Success Rate (ASR)** to evaluate adversarial patch effectiveness. However, this conventional metric typically assesses performance only under fixed or limited viewpoints, failing to capture robustness across diverse viewing angles—a critical factor for real-world deployment. To address this limitation, we introduce a novel evaluation metric: **Angle-Aware Attack Success Rate (AASR)**. This metric comprehensively quantifies a patch's effectiveness across varied viewing perspectives. Let Ω represent the complete angular space of interest. The AASR is defined as a weighted integral of ASR across this angular domain: $\text{AASR} = \int_{\Omega} w(\theta) \cdot \text{ASR}(\theta) \cdot d\theta \times 100\%$ where $\text{ASR}(\theta)$ represents the traditional attack success rate when patches are viewed at angle $\theta \in \Omega$, and $w(\theta)$ is a normalized weighting function such that $\int_{\Omega} w(\theta) \cdot d\theta = 1$. In our evaluation, we adopt uniform weighting, *i.e.*, $w(\theta) = \frac{1}{|\Omega|}$ for all angles, ensuring equal contribution from each viewing angle to the AASR calculation. This generalized formulation allows for evaluation across arbitrary angular domains, accommodating various real-world deployment scenarios.

5.2 Digital Comparative Results

We evaluate AngleRoCL in digital environments across multiple detectors, comparing 4 T2I-based attacks (NDDA, NDDA+AngleRoCL, MAGIC, MAGIC+AngleRoCL) with a traditional physical adversarial approach (AdvPatch). Our evaluation simulates angles from $-90°$ to $90°$ using projective transformations, with patches placed at environment centers to eliminate positional bias. Table 1 summarizes the AASR performance across 5 detectors and 6 environments. We have the following observations: ❶ AngleRoCL significantly outperforms traditional physical adversarial approach. While AdvPatch only achieves 5.54% average AASR with substantial fluctuations across environments (0.00% to 14.21%), our NDDA+AngleRoCL reaches 36.02% and MAGIC+AngleRoCL 32.51%. ❷ AngleRoCL consistently enhances angle robustness of NDD-based methods, increasing NDDA's average AASR from 23.79% to 36.02% (51.4% improvement) and MAGIC's from 26.26% to 32.51% (23.8% improvement). ❸ Improvements persist across diverse environments, from 47.61% AASR in favorable Environment ② to 24.24% in challenging Environment ④. ❹ AngleRoCL enhances performance across all detection architectures, including YOLOv10 (72.8% improvement) and YOLOv5 (81.2% improvement), validating its cross-detector transferability.

Table 1: Angle-Aware Attack Success Rate (AASR) in digital environments. Results across five detectors in six environments, measured from -90° to 90° with 1° intervals. Best average highlighted in red, second best in blue. Best detector results **underlined+bold**, second best **bold**.

Environment	Method	Faster R-CNN	YOLOv3	YOLOv5	RT-DETR	YOLOv10	Avg.
Environment ①	AdvPatch	11.51%	2.52%	0.00%	0.00%	0.00%	0.78%
	NDDA	25.96%	1.88%	14.62%	14.57%	10.02%	13.41%
	NDDA+AngleRoCL	**39.79%**	**6.38%**	**36.58%**	19.30%	**23.82%**	25.17%
	MAGIC	29.35%	1.97%	18.55%	**25.52%**	11.67%	17.41%
	MAGIC+AngleRoCL	**39.72%**	**5.94%**	**34.08%**	**30.07%**	**20.24%**	26.01%
Environment ②	AdvPatch	23.31%	43.53%	2.25%	1.97%	0.00%	14.21%
	NDDA	41.64%	32.86%	29.99%	45.43%	28.99%	35.78%
	NDDA+AngleRoCL	**50.67%**	**50.76%**	**46.04%**	**47.64%**	**42.95%**	47.61%
	MAGIC	44.22%	34.94%	30.76%	42.80%	31.12%	36.77%
	MAGIC+AngleRoCL	**45.45%**	**43.34%**	**42.32%**	**44.29%**	**39.98%**	43.08%
Environment ③	AdvPatch	8.99%	15.73%	8.71%	0.00%	0.00%	6.69%
	NDDA	30.27%	10.60%	25.67%	23.81%	22.18%	22.51%
	NDDA+AngleRoCL	**43.94%**	**25.74%**	**43.38%**	**33.40%**	**36.11%**	36.51%
	MAGIC	31.13%	14.01%	32.46%	25.29%	24.70%	25.52%
	MAGIC+AngleRoCL	**36.04%**	**25.99%**	**42.71%**	**31.11%**	**34.16%**	34.00%
Environment ④	AdvPatch	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	NDDA	20.80%	3.61%	10.31%	18.07%	13.11%	13.18%
	NDDA+AngleRoCL	**31.25%**	**6.74%**	**27.34%**	**29.99%**	**25.87%**	24.24%
	MAGIC	21.24%	3.71%	14.46%	22.90%	15.98%	15.66%
	MAGIC+AngleRoCL	**28.42%**	**6.75%**	**22.68%**	**31.03%**	**22.24%**	22.22%
Environment ⑤	AdvPatch	9.55%	10.11%	0.00%	6.46%	0.00%	5.22%
	NDDA	38.70%	12.41%	29.07%	52.45%	21.93%	30.91%
	NDDA+AngleRoCL	**47.52%**	**26.26%**	**42.80%**	**54.23%**	**36.50%**	41.46%
	MAGIC	38.04%	14.67%	34.19%	**53.77%**	26.88%	33.51%
	MAGIC+AngleRoCL	**41.44%**	**24.38%**	**38.74%**	53.28%	**31.76%**	37.92%
Environment ⑥	AdvPatch	0.56%	31.15%	0.00%	0.00%	0.00%	6.34%
	NDDA	28.55%	25.95%	20.17%	**39.56%**	20.46%	26.94%
	NDDA+AngleRoCL	**42.02%**	**43.05%**	**39.14%**	**45.13%**	**36.41%**	41.15%
	MAGIC	**30.53%**	28.56%	24.40%	35.49%	24.58%	28.71%
	MAGIC+AngleRoCL	29.72%	**33.24%**	**31.81%**	37.16%	**27.33%**	31.85%

Table 2: Angle-Aware Attack Success Rate (AASR) in physical environment. Results across five detectors, measured from -70° to 70° with 10° intervals. Best average highlighted in red, second best in blue. Best detector results **underlined+bold**, second best **bold**.

Environment	Method	Faster R-CNN	YOLOv3	YOLOv5	RT-DETR	YOLOv10	Avg.
Environment ⑦	AdvPatch	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	NDDA	40.27%	7.73%	15.20%	56.27%	22.40%	28.37%
	NDDA+AngleRoCL	**69.87%**	**16.00%**	**60.80%**	**69.87%**	**42.20%**	51.75%
	MAGIC	39.73%	4.27%	17.33%	42.67%	9.60%	22.72%
	MAGIC+AngleRoCL	**83.73%**	**27.73%**	**72.53%**	**89.60%**	**55.73%**	65.86%

5.3 Physical Comparative Results

To validate our findings in real-world scenarios, we conducted physical experiments comparing the same six methods as in the digital environment. We printed patches on standard paper and deployed them in a regular road next to a college. Due to the higher costs of physical experiments, we employed a reduced angular sampling strategy compared to the digital evaluation. Observations were made at a fixed distance across 15 viewing angles from $-70°$ to $70°$ at $10°$ intervals. Table 2 summarizes the AASR performance across five detectors. The evaluation reveals two findings: ❶ AngleRoCL significantly outperforms traditional approach in physical environments. While AdvPatch completely failed in physical settings (0.00% AASR), our methods demonstrated robust performance with NDDA+AngleRoCL achieving 51.75% and MAGIC+AngleRoCL reaching 65.86% AASR. This stark contrast highlights the superior physical-world transferability of our approach compared to traditional methods that are highly sensitive to environmental changes. ❷ AngleRoCL consistently enhances NDD methods in physical settings, with NDDA+AngleRoCL achieving 51.75% AASR compared to 28.37% for vanilla NDDA (82.4% improvement), and MAGIC+AngleRoCL reaching 65.86% versus 22.72% for original MAGIC (189.9% improvement). These substantial enhancements were consistent across all viewing angles, confirming that our learned angular robustness concept transfers effectively to real-world scenarios without environment-specific optimization.

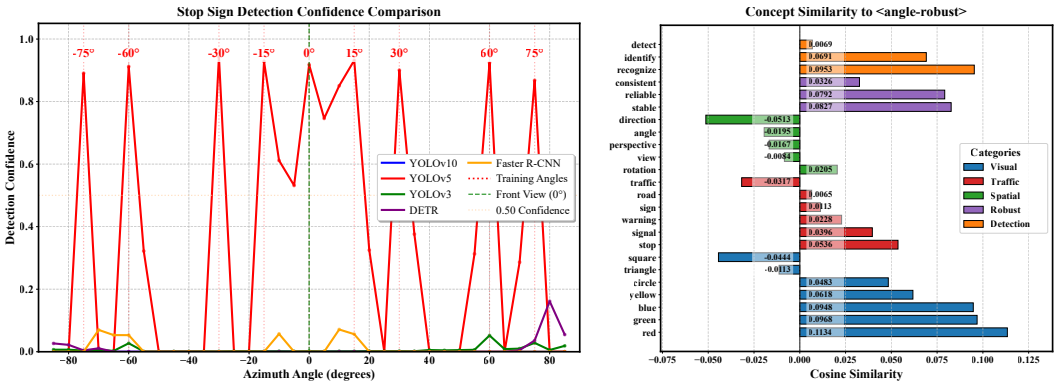

Figure 4: (a) Direct optimization shows severe overfitting to training angles (orange) and trained detector (YOLOv5). (b) Cosine similarity between learned `<angle-robust>` embedding and top-correlated tokens.

6 Ablation Study and Discussion

Ablation study. To validate our approach, we compared AngleRoCL with a direct optimization method that applies gradients directly to patch pixels instead of performing AngleRoCL. In Fig. 4, while directly optimized patches achieved high attack rates at trained angles and on trained detectors, they performed poorly on unseen angles and detectors. In contrast, AngleRoCL maintained consistent performance across all conditions, confirming that our embedding-based concept learning enables critical cross-angle and cross-detector generalization for real-world applications.

Embedding analysis. In Fig. 4, the cosine similarity analysis shows our `<angle-robust>` concept has developed meaningful associations with robustness-related features. Color tokens show the highest correlations (red: 0.1134, green: 0.0968, blue: 0.0948), while shapes show interesting patterns with circle being positive (0.0483) and square/triangle negative. To further validate these embedding insights, we regenerated the NDDA dataset using our learned angle-robust concept. Following the original NDDA methodology, we generated patches for all 39 prompts (50 patches per prompt) with and without the `<angle-robust>` token, and evaluated their angle robustness by digitally placing them in the center of blank environments across multiple viewing angles. The results are presented in Table 3. The embedding analysis results align with the experimental results. For NDDA on YOLOv5, removing color features caused the most significant performance drop (from 56.64% to 21.17%), matching our embedding analysis which identified color tokens as having the highest correlation with our concept. Moreover, AngleRoCL maintains a 62.33% detection rate even without color features—a 41.16 percentage point improvement over NDDA, confirming our approach effectively compensates for the removal of critical robustness features.

The correlations clearly show our learned concept captures essential visual attributes proportional to their importance for achieving angle robustness. This explains our method's effectiveness and further suggests angle robustness is fundamentally related to visual elements that maintain distinctive properties across different viewpoints.

Table 3: Angle-Aware Attack Success Rate (%) comparison between NDDA and AngleRoCL when robust features are removed. Each row indicates which features were removed from the prompt (checkmark means removed). For each detector and configuration, the best performance is highlighted in **bold**.

	Removed Robust Features				Object Detectors					
	Shape	Color	Text	Pattern	Faster R-CNN	YOLOv3	YOLOv5	RT-DETR	YOLOv10	Avg.
NDDA					59.58%	57.98%	56.64%	74.96%	52.57%	60.35%
	✔				30.38%	28.0%	27.78%	44.50%	23.52%	30.84%
		✔			61.53%	49.24%	21.17%	**71.75%**	20.84%	44.91%
			✔		51.39%	53.69%	43.84%	60.30%	39.04%	49.65%
				✔	41.95%	41.26%	29.69%	47.94%	27.98%	37.76%
	✔	✔	✔	✔	38.83%	31.09%	6.86%	52.43%	9.44%	27.73%
AngleRoCL					**74.16%**	**77.16%**	**77.07%**	**79.61%**	**72.96%**	**76.19%**
	✔				**37.09%**	**44.11%**	**40.75%**	**54.40%**	**38.69%**	**43.01%**
		✔			**64.08%**	**66.89%**	**62.33%**	71.54%	**60.46%**	**65.06%**
			✔		**53.50%**	**62.31%**	**66.33%**	**69.68%**	**60.33%**	**62.43%**
				✔	**51.14%**	**62.47%**	**58.38%**	**56.81%**	**54.74%**	**56.71%**
	✔	✔	✔	✔	**41.60%**	**46.12%**	**44.50%**	**57.01%**	**40.70%**	**45.99%**

Cross-Detector generalization While AngleRoCL improves angle robustness across multiple detectors, the degree varies between architectures. As shown in Table 3, different detectors show varying sensitivities to specific robustness features—when color features are removed, YOLOv5 shows dramatic improvement with our method (21.17% to 62.33%), while RT-DETR shows minimal change (71.75% to 71.54%). Since our framework uses only YOLOv5 for feedback during training, the resulting concept inherently captures YOLOv5's definition of robustness, explaining the most pronounced improvements on YOLOv5 (81.2% relative improvement). This detector-specific bias, while still enabling cross-detector generalization, limits achieving optimal universal robustness.
Limitations. We trained with only 9 sampled angles in the horizontal plane, which simplifies but doesn't fully capture continuous 3D angle variations, including vertical perspectives in real-world scenarios. While experiments show AngleRoCL significantly enhances patch robustness even with this limited sampling, the optimal angle sampling density and distribution remain unexplored.

7 Conclusion

In this paper, we introduced Angle-Robust Concept Learning (AngleRoCL), addressing the critical challenge of maintaining attack effectiveness across multiple viewing angles for T2I adversarial patches. Our comprehensive experiments in both digital and physical environments demonstrate that AngleRoCL significantly outperforms baseline methods without requiring environmental optimization. By encoding angle robustness as a learned concept, our method enables the generation of physically robust adversarial patches with consistent performance across viewpoints. The plug-and-play nature of our approach allows seamless integration with existing T2I attack frameworks. Beyond technical contributions, this work advances understanding of textual concepts and physical properties in diffusion-generated content, providing valuable insights for developing more robust defense mechanisms against angle-invariant adversarial attacks in real-world environments.

Acknowledgments

This research was supported by the NSFC (No. 62476143), Shenzhen Science and Technology Program (No. JCYJ20240813114237048), "Science and Technology Yongjiang 2035" key technology breakthrough plan project (No. 2025Z053), and Chinese government-guided local science and technology development fund projects (scientific and technological achievement transfer and transformation projects) (No. 254Z0102G). This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG4-GC-2023-008-1B), and National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore, and Infocomm Media Development Authority.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 3

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 3

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018. 3

[4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 1

[5] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *NeurIPS 2017 Workshop on Machine Learning and Computer Security*, 2017. 2, 3, 7

[6] Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25050–25059, 2025. 3

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017. 3

[8] Meng Chen, Jiawei Tu, Chao Qi, Yonghao Dang, Feng Zhou, Wei Wei, and Jianqin Yin. Towards physically-realizable adversarial attacks in embodied vision navigation. *arXiv preprint arXiv:2409.10071*, 2024. 3

[9] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 3

[10] Nhat Chung, Sensen Gao, Tuan-Anh Vu, Jie Zhang, Aishan Liu, Yun Lin, Jin Song Dong, and Qing Guo. Towards transferable attacks against vision-llms in autonomous driving with typography. *arXiv preprint arXiv:2405.14169*, 2024. 3

[11] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1000–1008, 2020. 3

[12] Mehmet Ergezer, Phat Duong, Christian Green, Tommy Nguyen, and Abdurrahman Zeybey. One noise to rule them all: Multi-view adversarial attacks with universal perturbation. In *International Conference on Artificial Intelligence and its Application*, pages 515–527. Springer, 2024. 3

[13] 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, 3

[14] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *RA-L*, 2022. 6, 17, 19

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 5

[16] Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, pages 442–460. Springer, 2024. 2

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2, 3

[18] Qi Guo, Shanmin Pang, Zhikai Chen, and Qing Guo. Towards robust deepfake distortion attack via adversarial autoaugment. *Neurocomputing*, 617:129011, 2025. 3

[19] Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 20:1333–1348, 2024. 2

[20] Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu. Spark: Spatial-aware online incremental attack against visual tracking. In *European conference on computer vision*, pages 202–219. Springer, 2020. 3

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[22] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading deepfake detectors via adversarial statistical consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12271–12280, 2023. 3

[23] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7848–7857, 2021. 1, 3

[24] Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16975–16984, 2023. 3

[25] Yihao Huang, Felix Juefei-Xu, Qing Guo, Geguang Pu, and Yang Liu. Natural & adversarial bokeh rendering via circle-of-confusion predictive network. *IEEE Transactions on Multimedia*, 26:5729–5740, 2023. 3

[26] Yihao Huang, Liangru Sun, Qing Guo, Felix Juefei-Xu, Jiayi Zhu, Jincao Feng, Yang Liu, and Geguang Pu. Ala: Naturalness-aware adversarial lightness attack. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2418–2426, 2023. 3

[27] Matthew Hull, Zijie J Wang, and Duen Horng Chau. Revamp: Automated simulations of adversarial attacks on arbitrary objects in realistic scenes. *arXiv preprint arXiv:2310.12243*, 2023. 3

[28] Xiaojun Jia, Jindong Gu, Yihao Huang, Simeng Qin, Qing Guo, Yang Liu, and Xiaochun Cao. Transegpgd: Improving transferability of adversarial examples on semantic segmentation. *arXiv preprint arXiv:2312.02207*, 2023. 2

[29] Glenn Jocher. Ultralytics yolov5, 2020. 7, 19

[30] Dehong Kong, Siyuan Liang, Xiaopeng Zhu, Yuansheng Zhong, and Wenqi Ren. Patch is enough: Naturalistic adversarial patch against vision-language pre-training models. *arXiv preprint arXiv:2410.04884*, 2024. 2

[31] Hui Kuurila-Zhang, Haoyu Chen, and Guoying Zhao. Venom: Text-driven unrestricted adversarial example generation with diffusion models, 2025. 3

[32] Jin Li, Ziqiang He, Anwei Luo, Jian-Fang Hu, Z. Jane Wang, and Xiangui Kang. Advad: Exploring non-parametric diffusion for imperceptible adversarial attacks, 2025. 3

[33] Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsheng Xu, and Hongkai Yu. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15205–15215, 2024. 3

[34] Shuo-Yen Lin, Ernie Chu, Che-Hsien Lin, Jun-Cheng Chen, and Jia-Ching Wang. Diffusion to confusion: Naturalistic adversarial patch generation based on diffusion model for object detector, 2023. 3

[35] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017. 3

[36] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 2

[37] Linye Lyu, Jiawei Zhou, Daojing He, and Yu Li. Cnca: Toward customizable and natural generation of adversarial camouflage for vehicle detectors, 2024. 3

[38] Yasamin Medghalchi, Moein Heidari, Clayton Allard, Leonid Sigal, and Ilker Hacihaliloglu. Prompt2perturb (p2p): Text-guided diffusion-based adversarial attacks on breast ultrasound images. *arXiv preprint arXiv:2412.09910*, 2024. 3

[39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2, 3

[40] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 3

[41] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 2, 3

[42] Scott Oslund, Clayton Washington, Andrew So, Tingting Chen, and Hao Ji. Multiview robust adversarial stickers for arbitrary objects in the physical world. *Journal of Computational and Cognitive Engineering*, 1(4):152–158, 2022. 3

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 7, 19

[45] 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. 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 7

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[48] Takami Sato, Justin Yue, Nanze Chen, Ningfei Wang, and Qi Alfred Chen. Intriguing properties of diffusion models: An empirical study of the natural attack capability in text-to-image generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24635–24644, 2024. 1, 3, 6, 7, 19

[49] Samridha Shrestha, Saurabh Pathak, and Eduardo K Viegas. Towards a robust adversarial patch attack against unmanned aerial vehicles object detection. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3256–3263. IEEE, 2023. 3

[50] Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Dta: Physical camouflage attacks using differentiable transformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2022. 3

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2, 3

[52] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[53] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 7, 19

[54] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8565–8574, 2021. 3

[55] Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4412–4423, 2023. 2

[56] Zhixiang Wang, Xingjun Ma, and Yu-Gang Jiang. Badpatch: Diffusion-based generation of physical adversarial patches, 2025. 3

[57] Xingxing Wei, Ying Guo, and Jie Yu. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2711–2725, 2022. 3

[58] Yun Xing, Nhat Chung, Jie Zhang, Yue Cao, Ivor Tsang, Yang Liu, Lei Ma, and Qing Guo. Magic: Mastering physical adversarial generation in context through collaborative llm agents, 2025. 1, 3, 6, 7, 17, 19

[59] Liming Zhai, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Lei Ma, Wei Feng, Shengchao Qin, and Yang Liu. It's raining cats or dogs? adversarial rain attack on dnn perception. *arXiv preprint arXiv:2009.09205*, 2020. 3

[60] Qian Zhang, Qing Guo, Ruijun Gao, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Adversarial relighting against face recognition. *IEEE Transactions on Information Forensics and Security*, 2024. 2

[61] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 7

[62] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 1989–2004, 2019. 1, 3, 5

[63] 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. 2

[64] Zijian Zhu, Xiao Yang, Hang Su, and Shibao Zheng. Camoenv: Transferable and environment-consistent adversarial camouflage in autonomous driving. *Pattern Recognition Letters*, 188:95–102, 2025. 3

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and the scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitation in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have described the proposed method in Sec. 4 and the corresponding implementation details in Sec. 4.3. Additionally, the code and models will be made publicly available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available assets, including the evaluation protocol [58] and the nuImage dataset [14], both of which are released under open-access licenses. All relevant licenses are respected and cited in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All training details including data splits, hyperparameters, and optimization methods are specified in Section 4.3 and Appendix **??**. Complete implementation code is provided in our repository.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments ran on 2 NVIDIA 3090 GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research fully conforms to the NeurIPS Code of Ethics in all aspects of methodology and reporting.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper studies physical attack on object detector. It reveals the potential flaws of the existing object detectors and provides our perspective for subsequent research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The data we release complies with the open-source agreement.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available assets, including the evaluation protocol from NDDA [48] and MAGIC [58], the nuImage dataset [14], and various detector implementations (YOLOv5 [29], YOLOv10 [53], Faster R-CNN [44], etc.). All assets are properly cited with their original publications. For detector implementations, we used the official releases from Ultralytics (for YOLOv5/v10) and MMDetection (for others) under their respective open-source licenses. The Stable Diffusion v1.5 model was used under the CreativeML Open RAIL-M license. All licenses and terms of use were respected throughout our research.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new concept embedding for angle robustness (`<angle-robust>`), which will be released as part of our code repository. The repository includes comprehensive documentation detailing the model architecture, training procedures, hyperparameters, and usage instructions. We also provide the trained embedding weights, sample prompts, and evaluation scripts to reproduce our results. The code repository is structured to facilitate both immediate application and further research extensions. All assets are released under the MIT license, and we have anonymized the repository link for the submission process.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.