

1 Multi-Scene Physical Experiments

To further validate the generalizability and robustness of our approach across diverse real-world conditions, we conducted additional physical experiments in three distinct environments. These supplementary evaluations extend our main paper findings by demonstrating consistent performance improvements across varied physical settings.

The experimental protocol remains identical to the physical experiments described in the main paper, with each method evaluated using 10 sample patches per environment. As presented in Table 1, our results demonstrate remarkable consistency across all three testing environments (⑦-⑨).

MAGIC+AngleRoCL achieves the highest average AASR across all environments, with 61.87% in Environment ⑦, 55.33% in Environment ⑧, and 69.07% in Environment ⑨. Similarly, NDDA+AngleRoCL consistently outperforms its baseline counterpart, achieving 39.60%, 38.40%, and 46.67% respectively. The traditional AdvPatch method fails completely across all physical environments, reinforcing our earlier findings regarding the limitations of conventional adversarial patch approaches in real-world deployment.

Notably, the performance improvements remain substantial and consistent across diverse environmental conditions. MAGIC+AngleRoCL demonstrates relative improvements of 191.8%, 260.3%, and 174.1% over vanilla MAGIC in the three environments respectively, while NDDA+AngleRoCL achieves improvements of 60.6%, 105.7%, and 65.1% over baseline NDDA. These consistent gains across varied physical settings validate the environmental robustness of our angle-robust concept learning approach and confirm its practical applicability in diverse real-world scenarios.

Table 1: Angle-Aware Attack Success Rate (AASR) in physical environments. 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	42.00%	2.67%	9.33%	48.00%	21.33%	24.67%
	NDDA+AngleRoCL	51.33%	8.00%	44.00%	56.67%	38.00%	39.60%
	MAGIC	38.00%	6.00%	10.00%	44.67%	7.33%	21.20%
	MAGIC+AngleRoCL	82.00%	26.00%	64.67%	88.00%	48.67%	61.87%
Environment ⑧	AdvPatch	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	NDDA	37.33%	3.33%	8.67%	28.00%	16.00%	18.67%
	NDDA+AngleRoCL	49.33%	13.33%	44.67%	50.67%	34.00%	38.40%
	MAGIC	38.67%	8.67%	10.76%	14.00%	4.67%	15.35%
	MAGIC+AngleRoCL	71.33%	32.00%	58.00%	69.33%	46.00%	55.33%
Environment ⑨	AdvPatch	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	NDDA	49.33%	24.67%	14.00%	49.33%	4.00%	28.27%
	NDDA+AngleRoCL	56.00%	41.33%	60.00%	53.33%	22.67%	46.67%
	MAGIC	41.33%	24.67%	22.00%	33.33%	4.67%	25.20%
	MAGIC+AngleRoCL	82.00%	57.33%	76.00%	85.33%	44.67%	69.07%

2 Implementation Details

2.1 Computing Resources

All experiments are performed using two NVIDIA-GeForce-RTX-3090 GPUs. The overall duration of all the experiments in the paper was about six weeks.

2.2 NDDA Prompts

In our implementation, we utilize a comprehensive pool of 39 NDDA prompts derived from the open-source dataset provided in the original NDDA work [48]. Following the categorization framework established in MAGIC [58], we systematically classify these prompts based on which robust features have been modified. The complete prompt pool is presented below.

NDDA Prompt Pool

Unmodified:

- “a photo of a stop sign”

Single Feature Modification:

- *Shape modification:*
 - “a photo of a square stop sign”
 - “a photo of a triangle stop sign”
- *Color modification:*
 - “a photo of a blue stop sign”
 - “a photo of a yellow stop sign”
- *Text modification:*
 - “a photo of a stop sign with ‘abcd’ on it”
 - “a photo of a stop sign with ‘hello’ on it”
 - “a photo of a stop sign with ‘world’ on it”
- *Pattern modification:*
 - “a photo of a stop sign with checkerboard paint on it”
 - “a photo of a stop sign with polkadot paint on it”

Dual Feature Modification:

- *Color + Shape:*
 - “a photo of a blue square stop sign”
 - “a photo of a yellow triangle stop sign”
- *Color + Text:*
 - “a photo of a blue stop sign with ‘abcd’ on it”
 - “a photo of a blue stop sign with ‘hello’ on it”
 - “a photo of a yellow stop sign with ‘world’ on it”
- *Color + Pattern:*
 - “a photo of a blue stop sign with checkerboard paint on it”
 - “a photo of a yellow stop sign with polkadot paint on it”
- *Shape + Text:*
 - “a photo of a square stop sign with ‘abcd’ on it”
 - “a photo of a square stop sign with ‘hello’ on it”
 - “a photo of a triangle stop sign with ‘world’ on it”
- *Shape + Pattern:*
 - “a photo of a square stop sign with checkerboard paint on it”
 - “a photo of a triangle stop sign with polkadot paint on it”
- *Text + Pattern:*
 - “a photo of a stop sign with ‘abcd’ on it and checkerboard paint on it”
 - “a photo of a stop sign with ‘hello’ on it and checkerboard paint on it”
 - “a photo of a stop sign with ‘world’ on it and polkadot paint on it”

Triple Feature Modification:

- *Color + Shape + Text:*
 - “a photo of a blue square stop sign with ‘abcd’ on it”
 - “a photo of a blue square stop sign with ‘hello’ on it”
 - “a photo of a yellow triangle stop sign with ‘world’ on it”
- *Color + Shape + Pattern:*
 - “a photo of a blue square stop sign with checkerboard paint on it”
 - “a photo of a yellow triangle stop sign with polkadot paint on it”
- *Color + Text + Pattern:*
 - “a photo of a blue stop sign with ‘abcd’ on it and checkerboard paint on it”

- “a photo of a blue stop sign with ‘hello’ on it and checkerboard paint on it”
- “a photo of a yellow stop sign with ‘world’ on it and polkadot paint on it”
- *Shape + Text + Pattern:*
 - “a photo of a square stop sign with ‘abcd’ on it and checkerboard paint on it”
 - “a photo of a square stop sign with ‘hello’ on it and checkerboard paint on it”
 - “a photo of a triangle stop sign with ‘world’ on it and polkadot paint on it”

Complete Feature Modification:

- “a photo of a blue square stop sign with ‘abcd’ on it and checkerboard paint on it”
- “a photo of a blue square stop sign with ‘hello’ on it and checkerboard paint on it”
- “a photo of a yellow triangle stop sign with ‘world’ on it and polkadot paint on it”

2.3 Task-oriented Narrative Instructions

To investigate whether T2I models could be directly instructed to generate angle-robust patches, we augmented the original NDDA prompts with task-oriented narrative instructions specifically addressing angle robustness. As described in the main paper, we implemented three distinct prompt modification strategies: prefix-enhanced, infix-integrated, and suffix-appended. Each strategy incorporates explicit linguistic instructions directing the model to generate images that maintain effectiveness across multiple viewing angles.

Task-oriented Narrative Instructions

Prefix Modifiers:

- “To enable the detection from multiple horizontal angles, I need a [base prompt]”
- “To enable the recognition from numerous horizontal angles, I need a [base prompt]”
- “To enable the identification from several horizontal angles, I need a [base prompt]”

Infix Modifiers:

- “[base prompt part 1] that is detectable at multiple angles in all horizontal directions [base prompt part 2]”
- “[base prompt part 1] that is recognizable at numerous angles in all horizontal orientations [base prompt part 2]”
- “[base prompt part 1] that is identifiable at several angles in all horizontal perspectives [base prompt part 2]”

Suffix Modifiers:

- “[base prompt] which can be detected by the detector from different horizontal angles of observation”
- “[base prompt] which can be recognized by the sensor from various horizontal angles of viewing”
- “[base prompt] which can be identified by the system from diverse horizontal angles of scanning”

Implementation Details:

- For prefix modifications, the instruction was placed at the beginning of the prompt, followed by the original NDDA prompt
- For infix modifications, we identified the phrase “stop sign” in the original prompt and inserted the instruction after it using a natural connecting phrase “that is”
- For suffix modifications, the instruction was appended to the end of the original prompt
- For each base NDDA prompt, one modifier was randomly selected from the corresponding category (prefix, infix, or suffix)
- All other generation parameters remained identical to the standard NDDA prompt generation process

As demonstrated in our empirical studies (Fig. 2 in the main paper), these linguistically-enhanced prompts not only failed to improve angle robustness but actually exhibited significant degradation in angular robustness metrics compared to the original NDDA prompts. The quantitative analysis presented in Table 2 provides compelling evidence of this phenomenon: across all feature modification scenarios, the task-oriented narrative instructions consistently underperform the baseline NDDA method. Specifically, the prefix-enhanced prompts show the most severe degradation, with average AASR dropping from 60.35% (NDDA baseline) to 34.31% (NDDA+Prefix), representing a 43.1% relative performance decrease. The infix-integrated and suffix-appended modifications also

demonstrate substantial performance losses, achieving only 49.05% and 49.90% average AASR respectively. This finding provides compelling evidence that current T2I text encoders struggle to interpret abstract, goal-oriented instructions for physical properties like angle robustness, highlighting the need for our concept learning approach.

Table 2: Angle-Aware Attack Success Rate (%) comparison between task-oriented narrative instructions and baseline NDDA method when robust features are removed. Prefix, Infix, and Suffix refer to different positions where angle-robustness instructions are inserted into the original NDDA prompts. 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					Avg.
Shape	Color	Text	Pattern		Faster R-CNN	YOLOv3	YOLOv5	RT-DETR	YOLOv10	
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%
NDDA+Prefix	✓				36.43%	35.44%	24.78%	44.37%	30.54%	34.31%
		✓			10.48%	11.99%	5.21%	21.29%	8.14%	11.42%
			✓		27.65%	22.76%	8.97%	44.37%	12.10%	23.17%
				✓	32.47%	31.56%	19.45%	46.28%	22.81%	30.51%
				✓	18.00%	15.91%	7.64%	27.17%	10.97%	15.94%
	✓	✓	✓	✓	18.17%	15.86%	2.98%	37.26%	6.49%	16.15%
NDDA+Infix	✓				51.03%	48.63%	40.85%	61.10%	43.65%	49.05%
		✓			26.07%	23.14%	18.31%	43.35%	21.94%	26.56%
			✓		46.06%	41.99%	12.97%	61.78%	18.91%	36.34%
				✓	49.56%	47.12%	39.61%	60.68%	41.42%	47.68%
				✓	38.62%	44.12%	33.27%	56.34%	33.33%	41.14%
	✓	✓	✓	✓	34.60%	26.71%	6.49%	47.70%	10.54%	25.21%
NDDA+Suffix	✓				47.21%	50.43%	43.98%	64.85%	43.02%	49.90%
		✓			23.51%	27.39%	20.07%	40.47%	25.06%	27.30%
			✓		53.58%	44.25%	9.14%	65.69%	15.58%	37.65%
				✓	41.39%	46.61%	35.28%	59.39%	34.19%	43.37%
				✓	32.01%	36.76%	23.66%	41.81%	27.61%	32.37%
	✓	✓	✓	✓	39.03%	29.69%	9.81%	51.80%	13.76%	28.82%

2.4 Details of Digital&Physical Experiments

Experimental environments. We conducted comprehensive evaluations across diverse environments to validate the effectiveness and generalizability of our approach. As illustrated in Fig. 1, our experimental setup encompasses both digital and physical environments. The digital evaluation utilizes six representative environments (①-⑥) selected from the nuImage dataset, which provide diverse urban driving scenarios including highway scenes, intersection contexts, and various lighting conditions as detailed in the main paper. For physical world validation, we carefully selected three real-world environments (⑦-⑨) that simulate authentic road scenarios and viewing perspectives. Environment ⑦ represents the primary physical testing site discussed in the main paper, while environments ⑧ and ⑨ serve as additional validation scenes to demonstrate cross-environment effectiveness. The selection criteria for physical environments prioritized realistic road scenarios that closely mirror actual driving conditions, ensuring ecological validity while maintaining controlled experimental conditions. Multiple physical environments were deliberately chosen to validate the robustness of our method across different real-world settings. It is important to note that all physical experiments were conducted in controlled environments without causing any disruption or safety concerns to other road users or pedestrians.

Digital experiment protocol. In digital environments, patches undergo projective transformation to simulate viewing angle variations, then are digitally inserted into the selected scenes. Each patch (150 × 150 pixels) is positioned at the image center and scaled to occupy approximately 1.5% of the scene area (1600 × 900 pixels), mirroring the physical experiment configuration to ensure realistic simulation. A single image containing the transformed patch at a specific viewing angle constitutes one perspective sample, which is subsequently fed to the detector for evaluation. All detectors operate with default hyperparameter settings throughout the evaluation process.

Physical experiment protocol. For physical validation, patches are printed on standard A4 paper and fixed at a predetermined location. Images are captured by moving the camera around the stationary patch while maintaining constant distance across all viewing angles. The patch size is configured to occupy 1.5–2% of the scene, positioned at the image center to maintain consistency with digital experiments. Captured images are directly processed by detectors using default hyperparameter configurations for evaluation.



Figure 1: Overview of experimental environments used in our evaluation. Top two rows show the six digital environments (①-⑥) selected from the nuImage dataset, representing diverse urban driving scenarios. Bottom row displays the three physical environments (⑦-⑨) used for real-world validation, where ⑦ corresponds to the primary physical testing site mentioned in the main paper, and ⑧-⑨ are additional environments for cross-scene validation.

3 Effect of Observation Distance

Our AngleRoCL training process employs projective transformations at a fixed observation distance to generate multi-angle images for concept learning. This raises an important question: how does varying observation distance affect the performance of our learned angle-robust concept and the relative importance of different robustness features?

To investigate this phenomenon, we conducted a comprehensive analysis using the NDDA patch dataset categorized by robust feature removal, as described in Sec. 3 of the main paper. We evaluated patches at two distinct observation distances—close (where patches occupy approximately 30% of the scene area) and far (where patches occupy approximately 5% of the scene area)—across multiple viewing angles and computed the corresponding AASR values. The results, presented in Fig. 2, reveal significant insights into the distance-dependent behavior of different robustness features.

Overall distance sensitivity. As illustrated in Fig. 2, increased observation distance leads to performance degradation across most feature categories, with the original patches (ORIGIN) maintaining the highest AASR at both distances (91.73% at close distance vs. 87.34% at far distance). This 4.39% absolute decrease demonstrates that even patches with complete robust features experience some distance-related performance loss, indicating the inherent challenge of maintaining attack effectiveness across varying observation distances.

Shape feature exhibits the most pronounced distance sensitivity. Most notably, patches with removed shape features (S) show the most dramatic performance degradation when observation distance increases, plummeting from 51.71% AASR at close distance to 20.42% at far distance—a 31.29% absolute decrease representing a 60.5% relative decline. This extreme sensitivity far exceeds all other feature categories and demonstrates that geometric features become critically important for maintaining detection confidence as observation distance increases, likely due to the severely reduced visual salience of shape information at greater distances.

Color features demonstrate significant distance sensitivity. Patches with removed color features (C) also experience substantial performance degradation, dropping from 67.75% AASR at close

distance to 52.73% at far distance—a 15.02% absolute decrease (22.2% relative decrease). This significant drop suggests that color information becomes increasingly important for patch recognition at greater distances, as chromatic cues help compensate for the loss of fine-grained visual details.

Complex feature interactions emerge at different distances. Interestingly, certain feature combinations exhibit counterintuitive behavior: patches with removed pattern features (P) actually show improved performance at far distances (68.61% to 73.01%), suggesting that the absence of pattern details may be beneficial when overall visual resolution decreases. However, patches with text and pattern removal (T+P) show performance decline (71.51% to 65.60%), while those with multiple feature removals involving shape (e.g., S+P: 58.57% to 28.36%) demonstrate severe performance degradation, indicating that the compound effect of shape feature absence becomes extremely pronounced at greater observation distances.

Implications for concept learning. These findings have critical implications for our angle-robust concept learning approach. Since our training process uses a fixed observation distance, the learned concept may be heavily biased toward the distance-specific importance hierarchy of robust features. The extreme sensitivity of shape-related patches to distance variations (60.5% relative decrease) suggests that if training were conducted at greater distances, the learned concept would need to develop much stronger associations with shape-related tokens in the embedding space. This could potentially improve performance for far-distance scenarios where geometric features are paramount, while requiring careful balancing to maintain effectiveness in close-distance scenarios.

Future directions. This analysis reveals a critical research direction: developing distance-robust concept learning that can adapt to varying observation distances. The dramatic performance variations observed, particularly the 60.5% relative decrease for shape-related features, highlight the urgent necessity for adaptive approaches. Future work could involve multi-distance training protocols or distance-adaptive concept embeddings that dynamically adjust the relative importance of different robust features based on the observation context. Such approaches could help mitigate the extreme distance sensitivity we observed and represent a significant step toward comprehensive robustness that encompasses both angular variations and distance-related challenges in real-world deployment scenarios.

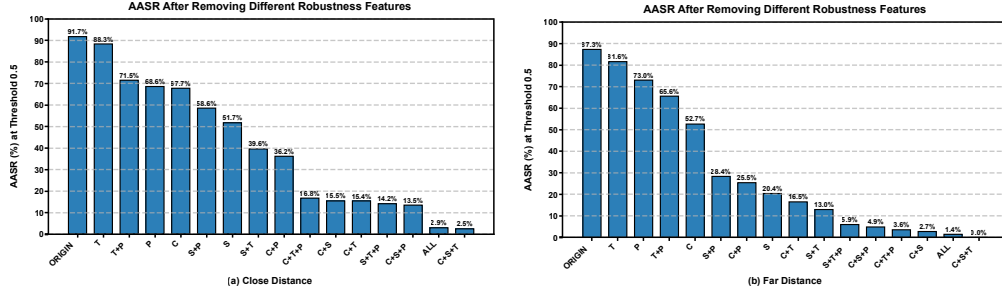


Figure 2: Effect of observation distance on AASR across different feature removal categories. (a) Close distance and (b) far distance results. X-axis labels indicate removed robust features: ORIGIN (no removal), S=Shape, C=Color, T=Text, P=Pattern, and their combinations.

4 Additional Visualization Results

Comparison of patches with and without angle-robust concept. To illustrate the visual impact of our angle-robust concept, we present comparative visualizations of patches generated with and without the learned `<angle-robust>` concept. Fig. 3 shows NDDA patch pairs comparing baseline patches with AngleRoCL-enhanced patches generated from identical prompts. Fig. 4 presents MAGIC patch pairs with direct comparisons between baseline and AngleRoCL versions. These visualizations demonstrate the systematic changes introduced by our learned concept in the patch generation process.

Multi-view detection results. We will make our complete experimental results publicly available online, including detection results across multiple environments and detectors. Additionally, we have created demonstration videos showcasing multi-angle detection results in real-world scenarios. These videos will be included in the supplementary materials to provide visual evidence of our method’s angle-robust performance.

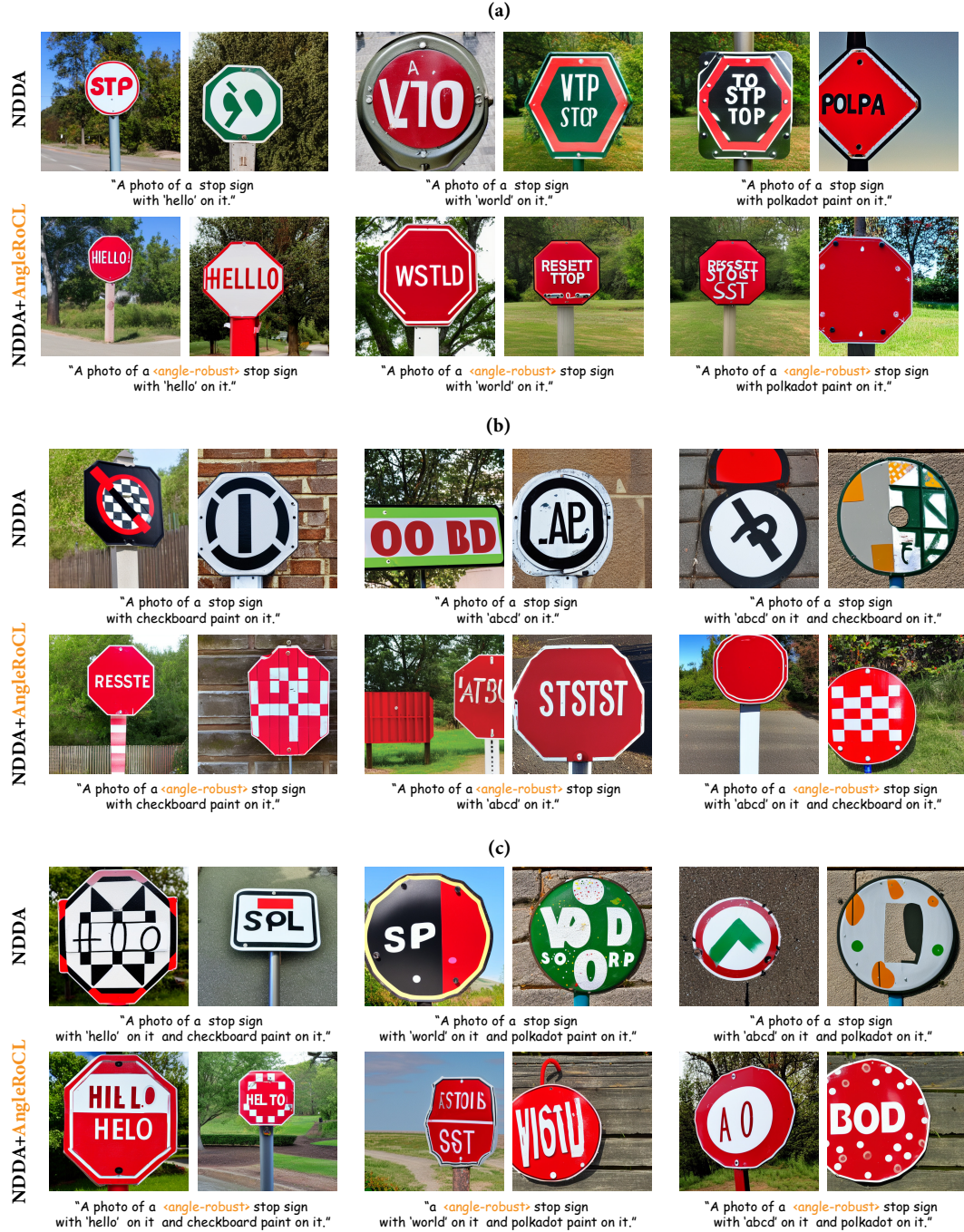


Figure 3: Visual comparison of NDDA patches generated with and without angle-robust concept. (a)-(c) Patch pairs organized for layout purposes, each showing baseline NDDA patches (top row of each section) vs NDDA+AngleRoCL patches (bottom row of each section) generated from identical prompts.



Figure 4: Visual comparison of MAGIC patches generated with and without angle-robust concept. (a)-(b) Patch pairs organized for layout purposes, each showing baseline MAGIC patches (top row of each section) vs MAGIC+AngleRoCL patches (bottom row of each section) generated from identical prompts.

5 Broader Impact

This work explores angle-robust adversarial patches that maintain effectiveness across multiple viewing angles, which has implications for both security and safety of AI systems. While our research advances understanding of physically robust attacks against object detectors, potentially revealing vulnerabilities in critical systems like autonomous vehicles and surveillance, it simultaneously provides valuable insights for developing more robust defense mechanisms. By demonstrating that text-to-image models can generate angle-invariant adversarial examples, we highlight a previously underexplored vulnerability that security researchers and system designers should address. Our work follows responsible disclosure practices by using common benchmark datasets and focusing on well-studied target objects (stop signs). We believe that understanding these vulnerabilities is essential for developing more resilient detection systems that maintain reliability across varying real-world viewing conditions, ultimately leading to safer AI deployment in critical applications.