G-PAC: Constructing Cohesive Pseudo-Features for Generalizable Physical Adversarial Camouflage

Tianrui Lou **Haoqing Zhang** **Jiawei Liang** **Puning Zhao** **Xiaochun Cao**
Sun Yat-sen University
loutianrui@gmail.com zhanghq59@mail2.sysu.edu.cn
liangjw57@mail2.sysu.edu.cn zhaopn@mail.sysu.edu.cn caoxiaochun@mail.sysu.edu.cn

Abstract

Physical-domain adversarial attacks present critical security threats to diverse AI systems, notably autonomous driving. Generalization is an inherent requirement and a primary obstacle in physical attacks, manifesting as the ability of adversarial patterns to persist across varying environmental configurations and to transfer to unseen victim models. While prior methods demonstrate generalization across specific models, their performance drops significantly on open-vocabulary foundation models, with some approaches becoming entirely ineffective. This lack of generalization stems from the failure of adversarial camouflage to construct a cohesive pseudo-feature across perspectives, lacking the stable semantic representation characteristic of natural objects. Such a deficiency arises because existing attack paradigms rely on independent view-level optimization and non-targeted objectives, which lead to semantically scattered features and inconsistent optimization directions. To address these limitations, we propose Generalizable Physical Adversarial Camouflage (G-PAC), a framework that transitions from view-specific suppression to coordinated object-level representation learning. G-PAC maintains a momentum-updated global pseudo-feature center as a stable adversarial anchor to guide optimization across diverse configurations. By leveraging the feature space of a self-supervised foundation model as a semantic prior, we introduce a contrastive regularization term to encourage feature aggregation while ensuring adversarial potency. Comprehensive digital and physical evaluations demonstrate the effectiveness of G-PAC. Notably, it outperforms the strongest baseline by an average AP@0.5 margin of 0.10 across seven diverse detector architectures in digital settings, and further extends this margin to 0.12 in real-world physical tests. Furthermore, G-PAC exhibits profound black-box transferability against highly resilient open-vocabulary foundation models, inducing an additional absolute AP@0.5 reduction of 0.30 on GLIP. Code is available at `https://github.com/TRLou/GPAC`.

1 Introduction

Deep learning models have made significant advancements across various fields, extending from computer vision [13] to natural language processing [42]. Nonetheless, the emergence of adversarial attacks exposes the fundamental weaknesses of these systems. Although digital-domain attacks have triggered extensive security concerns, physical-world attacks conducted in practical settings pose even greater threats. These hazards significantly hinder the practical use of neural networks in safety-critical scenarios, such as autonomous driving [46, 3, 5], security surveillance [31, 49, 51, 23, 18, 24, 17], and remote sensing [47, 20, 25]. This work investigates physical attacks within the context of autonomous driving, with a primary emphasis on leveraging adversarial camouflage to compromise the detection of vehicles.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The core challenge of physical-domain attacks lies in the inherent requirement for broad generalization. This requirement manifests across two critical dimensions: environmental robustness against varying configurations and architectural transferability to unknown victim models. Although prior methods achieve robust generalization on specific detectors by employing EoT-style [1] optimization within differentiable rendering pipelines across diverse simulated configurations, the performance diminishes substantially when targeting open-vocabulary foundation detectors. These foundation models, pre-trained on massive-scale datasets and possessing superior zero-shot generalization capabilities, exhibit significant resilience against such conventional adversarial patterns.

We reflect on a key question: *why does an adversarial camouflage fail to generalize like natural objects?* We observe that while natural objects maintain clustered features across various viewpoints, the features of camouflaged objects are semantically scattered in the latent space. We hypothesize that this latent dispersion prevents the construction of cohesive pseudo-features, ultimately leading to the poor generalization of existing camouflages. To empirically validate this causal link, we analyze the features of various methods via a vision foundation model (DINOv3 [36]). As shown in Fig. 1, a clear positive linear correlation exists between semantic dispersion (mean cosine distance) and detector performance (AP@0.5), confirming that semantically scattered adversarial features directly lead to degraded generalization. We attribute the emergence of such scattered features to two critical flaws in prior optimization paradigms: (1) **View-level Optimization.** Independent view-level gradients without object-level constraints cause optimization to degrade into isolated localized tasks; (2) **Non-targeted Attack Paradigm.** Unlike supervised training, the non-targeted paradigm only requires evading detection, leading to divergent optimization directions across environmental configurations. Consequently, these inconsistent directions yield the scattered features that hinder robust generalization.

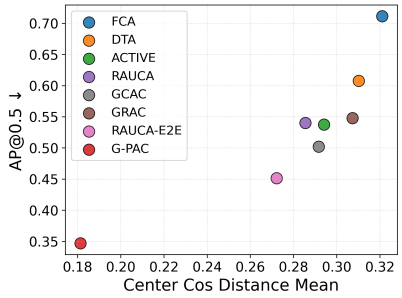


Figure 1: Correlation between latent feature dispersion and attack generalization. We utilize DINOv3 to extract representations from adversarial samples across various configurations and measure their dispersion via the mean cosine distance to the central point. A lower distance indicates more compact feature clusters. The results reveal a strong linear relationship: as adversarial features become more semantically scattered (higher distance), the transfer AP@0.5 increases, signifying a significant degradation in generalization.

To address these limitations, we propose **G**eneralizable **P**hysical **A**dversarial **C**amouflage (G-PAC), a framework that transitions from view-specific feature suppression to coordinated object-level representation learning. Central to G-PAC is the maintenance of a momentum-updated global pseudo-feature center that serves as a stable adversarial anchor to guide the camouflage optimization across diverse configurations. To prevent the optimization from overfitting to the local biases of specific surrogate detectors, we leverage the feature space of a self-supervised foundation model as a semantic prior. Within this robust embedding space, we introduce a contrastive regularization term based on the InfoNCE [32] objective. This mechanism encourages the adversarial features to aggregate toward the global center as positive pairs while simultaneously pushing them away from the clean features to ensure adversarial potency. We integrate this framework into an end-to-end training pipeline and employ a dedicated warm-up strategy to stabilize the global anchor before activating the contrastive loss. Through this paradigm shift, we achieve a substantial enhancement in the generalization of adversarial camouflages. Extensive experiments in both digital and physical domains demonstrate the consistent superiority of G-PAC. Notably, G-PAC establishes a new SOTA by outperforming the strongest baseline by an average AP@0.5 margin of 0.10 across seven architectures in digital evaluations, and 0.12 in physical deployments. It further induces an additional absolute AP@0.5 reduction of 0.30 on the highly resilient GLIP foundation detector through black-box transfer.

Our main contributions are threefold: **1)** We identify view-level optimization and non-targeted objectives as the root causes of scattered adversarial features and poor generalization in physical attacks. **2)** We introduce G-PAC, a framework leveraging a momentum-updated pseudo-feature center and contrastive regularization to achieve coordinated object-level learning. **3)** Extensive digital and physical experiments confirm the superiority of G-PAC, which outperforms the strongest baseline by an average AP@0.5 margin of 0.10 in digital settings and 0.12 in physical deployments.

2 Related Work

Physical Adversarial Attacks. Prior research has developed adversarial patterns for objects such as traffic signs [8, 11, 9, 37], pedestrians [38, 14, 15, 16, 54, 41], and vehicles [59, 52, 45, 60, 39, 40, 44, 43, 64, 65, 22, 21, 62, 48, 58, 29, 30], which are physically manifested through patches, clothes, or camouflage. Among various physical attack modalities, adversarial camouflage provides superior practical utility due to its sustained efficacy across diverse viewing configurations. Early methodologies addressed the complexities of mapping textures onto target surfaces through black-box strategies, such as neural network approximations [59] or genetic algorithms [52]. Recent advancements have transitioned toward white-box settings through differentiable rendering frameworks, which utilize either mesh-based neural renderers [44, 43, 39, 40, 45] or 3D Gaussian Splatting (3DGS) pipelines [29, 30]. Within these frameworks, several studies have focused on the suppression of attention maps to enhance adversarial transferability [44, 45]. Furthermore, the optimization of full-coverage camouflage has been proposed to mitigate the impacts of partial occlusion and long-distance detection [43]. To further strengthen robustness, research has integrated tri-planar mapping [40] and high-fidelity renderers [39], while gradient calibration and decorrelation strategies have been introduced to resolve optimization conflicts during multi-view updates [21]. Shifting toward representation-level manipulation, recent work such as PhyCamo [58] introduces contrastive learning to the physical domain, aiming to achieve feature-level stealth by explicitly pulling adversarial embeddings toward pure background representations. To address illumination and weather sensitivities, research adopts diverse relighting simulation strategies. One methodology, encompassing the RAUCA series [61, 62] and GRAC [22], utilizes environment feature extraction and UV map optimization. Another approach leverages GAN-based style learning to ensure visual plausibility [28]. Meanwhile, R-PGA [30] integrates relightable 3DGS with generative background inference via a hybrid rendering pipeline.

Transferability and Generalization in Attacks. The pursuit of generalization remains a core challenge in adversarial research. Early digital efforts enhanced transferability through gradient-based momentum [6] and input transformations [53], later shifting toward disrupting internal feature representationss [50, 56]. With the rise of foundation models, AnyAttack [57] introduces a self-supervised pre-training paradigm on massive datasets to enable label-free targeted attacks across diverse architectures. Simultaneously, GLEAM [27] leverages global-local image transformations combined with cross-modal feature alignment to enhance adversarial consistency. Despite the success of these digital methodologies, maintaining adversarial potency across varying environmental configurations and diverse architectures in the physical world remains an open challenge, particularly concerning the transfer to foundation models. To address this, we propose G-PAC, which achieves a substantial leap in transferability and generalization across various architectures including foundation models, thereby establishing a new performance benchmark for robust attacks against sophisticated detection systems.

3 Methodology

3.1 Preliminary

Adversarial Camouflage Representation. Let $\mathbf{M}$ denote the 3D mesh of a target vehicle and $\mathbf{T} \in \mathbb{R}^{H \times W \times 3}$ be its corresponding texture stored as a UV map. The goal of physical adversarial camouflage is to find a perturbation $\boldsymbol{\delta}$ such that the adversarial texture $\mathbf{T}_{\text{adv}} = \mathbf{T} + \boldsymbol{\delta}$, when applied to the vehicle surface, can mislead object detectors. Following the end-to-end optimization paradigm in RAUCA-E2E [62], we optimize $\mathbf{T}_{\text{adv}}$ directly in the UV space to ensure precise mapping during deployment.

Differentiable Rendering in Simulated Environments. To bridge the gap between 3D textures and 2D images, we employ a differentiable renderer $\mathcal{R}$. We encapsulate the diverse physical setups into a unified configuration space $\mathcal{C}$. Specifically, a configuration $\mathbf{c} \in \mathcal{C}$ parameterizes both the geometric transformations (e.g., camera pitch, azimuth, viewing distance, and vehicle spatial location) and the environmental conditions (e.g., fog density and sun altitude). Given a vehicle mesh $\mathbf{M}$, an optimized adversarial texture $\mathbf{T}_{\text{adv}}$, and a background reference image $\mathbf{x}_{\text{ref}}$, the rendering process is

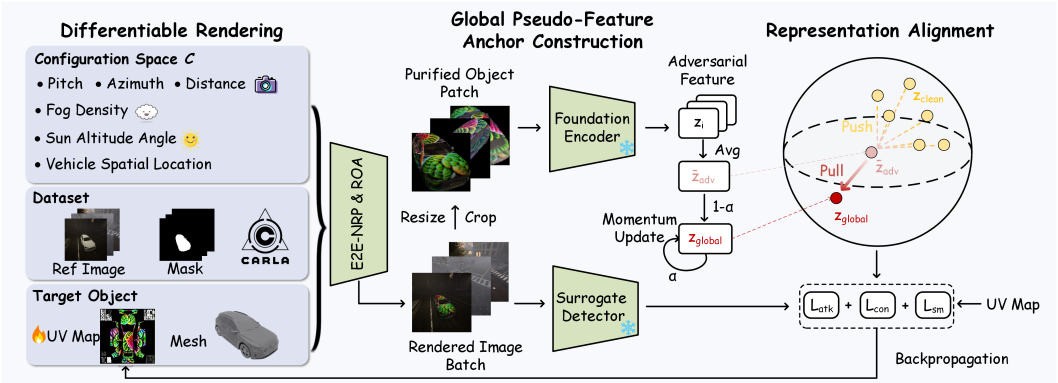


Figure 2: Architecture of the G-PAC framework. The differentiable rendering module produces batch images under diverse configurations. Global pseudo-feature anchor construction utilizes a foundation encoder and momentum updates to maintain a stable semantic target. Representation alignment enforces feature aggregation through contrastive learning. Finally, the combined loss facilitates backpropagation to optimize the UV map for coordinated object level representation learning.

formulated as:

$$\mathbf{x}_{\text{adv}} = \mathcal{R}(\mathbf{M}, \mathbf{T}_{\text{adv}}, \mathbf{x}_{\text{ref}}, \mathbf{c}). \tag{1}$$

To ensure the generated camouflage remains robust across realistic physical conditions, we utilize the CARLA autonomous driving simulator [7] to construct a high-fidelity multi-weather dataset. This dataset provides diverse pairs of $(\mathbf{x}_{\text{ref}}, \mathbf{c})$, effectively capturing the complex interplay of atmospheric scattering, dynamic lighting, and varying observation perspectives in real-world scenarios.

Traditional Optimization Objective. Physical attacks typically follow the Expectation over Transformation (EoT) framework [1] to ensure robustness. For a target detector Det and a benign image with its corresponding ground truth $(\mathbf{x}, y, \mathbf{b})$, where y is the class label and $\mathbf{b}$ is the bounding box, the non-targeted attack aims to minimize the detection performance by optimizing:

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{\mathbf{c}\sim\mathcal{C}}[\mathcal{L}_{\text{atk}}(Det(\mathcal{R}(\mathbf{M}, \mathbf{T}+\boldsymbol{\delta}, \mathbf{x}_{\text{ref}}, \mathbf{c})), y, \mathbf{b}) + \mathcal{L}_{\text{reg}}(\mathbf{M}, \mathbf{T}+\boldsymbol{\delta}, \mathbf{x}_{\text{ref}}, \mathbf{c})] \tag{2}$$

where $\mathcal{L}_{\text{atk}}$ is the attack loss, which typically involves suppressing the objectness score or maximizing the classification error, and $\mathcal{L}_{\text{reg}}$ represents the regularization terms, such as auxiliary constraints to enhance physical robustness and the smooth loss to maintain the visual naturalness of the texture.

3.2 Overview of G-PAC

The primary objective of G-PAC is to rectify the poor generalization of physical adversarial camouflage, a bottleneck fundamentally rooted in the scattered nature of adversarial features generated across diverse environmental configurations. To this end, our framework transitions the optimization paradigm from independent view-specific feature suppression toward a coordinated object-level representation learning process.

As illustrated in Fig 2, in each training iteration, we sample a batch of discrete configurations $\mathbf{c} \sim \mathcal{C}$ representing the distribution of realistic physical scenarios and render the corresponding adversarial images using Eq. 1. These images are subsequently projected into a robust feature space defined by representative visual foundation models, such as DINO [4, 33, 36], EVA [10], or CLIP [34], among others. Our framework is highly adaptable to different embedding spaces. Although these foundation models generally provide broad semantic priors that mitigate local architectural biases, their specific pre-training objectives influence the optimal degree of adversarial generalization. The core of G-PAC consists of two synergistic mechanisms, namely the maintenance of a momentum-updated global pseudo-feature center $\mathbf{z}_g$ that serves as a stable adversarial anchor (Sec 3.3) and the introduction of a contrastive alignment term (Sec 3.4). These mechanisms compute a contrastive regularization loss $\mathcal{L}_{\text{con}}$, which, together with the adversarial loss $\mathcal{L}_{\text{atk}}$ and smooth loss $\mathcal{L}_{\text{smooth}}$, forms the total objective for updating the texture $\mathbf{T}_{\text{adv}}$ via backpropagation following a specialized training strategy that includes a warm-up phase to stabilize the global anchor.

3.3 Object-level Optimization via Global Pseudo-Feature Anchor

To rectify the scattered nature of adversarial features inherent in view-level optimization, we introduce a Global Pseudo-Feature Anchor $\mathbf{z}_g$ as a unified optimization target. This anchor provides a stable semantic center that coordinates the features of visible texture components, ensuring that adversarial features aggregate into a cohesive cluster rather than remaining scattered.

Mask-guided Semantic Extraction. To isolate the target's semantic representation from environmental noise, we implement a mask-guided extraction process. For each configuration $\mathbf{c} \sim \mathcal{C}$, we first compute a purified object patch $\mathbf{x}_{\text{patch}}$ by applying the mask $\mathbf{m}$ (obtained from the semantic segmentation camera feature in CARLA [7, 62]) and cropping the rendered image $\mathbf{x}_{\text{adv}}$ to the minimal bounding box $\Omega_{\mathbf{m}}$ that encompasses the foreground pixels. This process is formally defined as:

$$\mathbf{x}_{\text{patch}} = \text{Resize}(\text{Crop}(\mathbf{x}_{\text{adv}} \odot \mathbf{m}, \Omega_{\mathbf{m}})), \tag{3}$$

where $\odot$ denotes the element-wise product. This patch is subsequently standardized and fed into the frozen visual foundation encoder $f(\cdot)$ to extract the embedding:

$$\mathbf{z} = \text{Norm}(f(\psi(\mathbf{x}_{\text{patch}}))), \tag{4}$$

where $\psi(\cdot)$ represents the normalization operation aligned with the pre-training specifications of the foundation encoder, and $\mathbf{z} \in \mathbb{R}^d$ is the L_2-normalized feature vector. By decoupling the target from its surroundings, this procedure ensures that the embedding $\mathbf{z}$ strictly represents the adversarial features of the camouflage, remaining independent of the surrounding environment.

Momentum-based Anchor Update. Relying solely on the instantaneous batch center $\bar{\mathbf{z}}_{\text{adv}}$ would impose a transient optimization target that fluctuates across iterations, forcing adversarial features into disjointed clusters rather than a coherent representation. To overcome this instability and achieve cross-batch consistency, we maintain $\mathbf{z}_g$ as a Global Pseudo-Feature Anchor that serves as a persistent memory of the object's cohesive pseudo-feature. This global anchor is updated via an Exponential Moving Average (EMA) mechanism:

$$\mathbf{z}_g \leftarrow \text{Norm}(\alpha \cdot \mathbf{z}_g + (1 - \alpha) \cdot \bar{\mathbf{z}}_{\text{adv}}), \tag{5}$$

where $\alpha \in [0, 1]$ is the momentum coefficient and $\text{Norm}(\cdot)$ denotes L_2 normalization. By integrating semantic information across the entire distribution $\mathcal{C}$, $\mathbf{z}_g$ serves as a stable anchor representing the global adversarial pseudo-feature. The introduction of this anchor preserves the adversarial potency inherent in gradient-based untargeted optimization while preventing the emergence of scattered representations across diverse environmental configurations.

3.4 Representation Alignment via Contrastive Learning

Based on the representations extracted in Sec 3.3, we define a contrastive alignment objective to guide the optimization of the adversarial texture. For each adversarial embedding $\mathbf{z}_{i,\text{adv}}$ within a batch of size B, we treat the global anchor $\mathbf{z}_g$ as the positive target to enforce representation aggregation toward the global adversarial pseudo-feature. Simultaneously, the corresponding set of clean embeddings $\{\mathbf{z}_{j,\text{clean}}\}_{j=1}^{B}$ serves as a negative sample pool, which encourages a distinct semantic shift away from the benign appearance. This alignment is formally achieved through the InfoNCE loss [32]:

$$\mathcal{L}_{\text{con}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\mathbf{z}_{i,\text{adv}}^{\top} \mathbf{z}_g / \tau)}{\exp(\mathbf{z}_{i,\text{adv}}^{\top} \mathbf{z}_g / \tau) + \sum_{j=1}^{B} \exp(\mathbf{z}_{i,\text{adv}}^{\top} \mathbf{z}_{j,\text{clean}} / \tau)} \tag{6}$$

where τ is a temperature hyper-parameter, which is assigned a default value of 0.1. Since all feature vectors are L_2-normalized on the unit hypersphere, the dot product directly represents the cosine similarity. By maximizing the similarity between $\mathbf{z}_{i,\text{adv}}$ and the stable anchor while minimizing its proximity to the clean semantic cluster, the resulting texture achieves robust generalization and configuration-invariance.

3.5 Overall Objective and Training Strategy

We optimize the adversarial texture by integrating our contrastive learning objective with the training framework from RAUCA-E2E [62].

Warm-up Strategy. To ensure optimization stability, we implement a warm-up phase of E_{warm} epochs. This warm-up phase prevents the optimization from prematurely converging toward a feature-stability shortcut at the expense of adversarial potency, while simultaneously enhancing stability by mitigating violent fluctuations during the initial training stage. During this period, the global anchor $\mathbf{z}_g$ is updated via the momentum mechanism (Sec 3.3) to establish a stable adversarial pseudo-feature, but the contrastive loss $\mathcal{L}_{\text{con}}$ does not back-propagate gradients to the texture.

Composite Loss. Following the engineering designs in RAUCA-E2E [62], we incorporate several auxiliary components to enhance physical robustness: the adversarial loss $\mathcal{L}_{\text{atk}}$ to suppress surrogate detector responses, $\mathcal{L}_{\text{sm}}$ for the smoothness of human vision, and the E2E-NRP renderer with ROA for realistic environmental simulation and data augmentation. The final optimization objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{atk}} + \mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{atk}} + \lambda_1 \mathcal{L}_{\text{sm}} + \lambda_2 \mathcal{L}_{\text{con}} \tag{7}$$

where λ_1 and λ_2 are hyper-parameters.

4 Experiments

This section first details the experimental setup and implementation specifics. We then validate the effectiveness of G-PAC through digital-domain experiments, including extensive qualitative and quantitative comparisons along with ablation studies (see Appendix.A for details). Finally, we present physical-domain experiments to demonstrate the robust performance of G-PAC on a 1:24 scale toy car.

4.1 Experimental Setup

Datasets. To thoroughly assess the efficacy of the proposed attack, we construct individual datasets for the digital and physical domains, respectively.

Regarding the **digital domain dataset**, we follow the experimental setup of previous studies by utilizing the CARLA simulator to ensure a fair and direct comparison. For the generation of adversarial camouflage, we adopt the same dataset configuration as RAUCA-E2E [62]. Specifically, this dataset utilizes the Audi E-Tron as the target vehicle and comprises 32,768 photo-realistic images for texture optimization. To enhance the challenge of the digital evaluation and verify true robustness, the experiments in the main text do not employ a test set from the same distribution as the training set, while the results for the same distribution test set are provided in the Appendix E. Instead, a more difficult evaluation set is constructed for every method with increased variation in lighting, weather, and camera perspectives. For each attack strategy, we generate a test set by rendering images from various viewpoints around a vehicle equipped with the specified adversarial camouflage. This data collection covers 6 weather conditions including dark, foggy, golden, hardnoon, normal, and overcast. Furthermore, the dataset incorporates 4 distinct distances of 5m, 10m, 15m, and 20m, along with 10 camera pitch angles ranging from 0° to 90° at 10° increments. At each configuration, images are captured every 20° to complete a 360° rotation. This process results in a comprehensive test set of 4,320 images for each method evaluated.

For the **physical domain dataset**, we deploy the adversarial camouflages generated by G-PAC and other SOTA methods on a 1:24 scale Audi Q5 model car. The camouflage patterns are realized via high-resolution printing and applied to the vehicle surface using stickers. To ensure consistency with our digital-domain evaluation, we construct the physical test set by capturing images under diverse real-world configurations. Specifically, our setup encompasses 4 camera pitch angles ($20°, 40°, 60°, 80°$), 2 shooting distances (50cm and 100cm), and 2 distinct lighting conditions with significant contrast. For each combination, we perform a $360°$ surround-view capture at $22.5°$ intervals. Consequently, for each comparative method, we assemble a physical test set comprising 256 images for comprehensive qualitative and quantitative analysis.

Target Models. We select seven widely adopted detection architectures for the experiments. These include four detectors pre-trained on the COCO dataset, categorized as follows: one-stage detectors, including Yolo-V3 with a DarkNet-53 backbone [35] and YoloX-L [12]; a two-stage detector,

Table 1: This table provides a comparative analysis of AP@0.5 across multiple detection architectures for different physical attack strategies. The reported values represent the average detection performance measured on a test dataset encompassing a variety of viewpoints, weather conditions, and shooting distances. Optimization of adversarial camouflage is performed with Yolo-V3, while black-box transferability, indicated by an asterisk (*), is assessed on models such as GLIP, Detic, Grounding-DINO, Yolo-X, C-RCNN, and DINO. In each column, the best results are highlighted in **bold**, and the second-best results are underlined.

Method	Paper Year	White-box	Black-box						Average
			Foundation Detector			One-Stage	Two-Stage	Transformer-based	
		Yolo-V3	Glip*	Detic*	Grounding-DINO*	Yolo-X*	C-RCNN*	DINO*	
Clean	-	0.5251	0.8897	0.7693	0.9090	0.7395	0.6892	0.9093	0.7759
FCA[43]	AAAI 2022	0.4981	0.8490	0.6833	0.9091	0.6549	0.4822	0.9017	0.7112
DTA[39]	CVPR 2022	0.3461	0.7702	0.6316	0.9089	0.4070	0.2944	0.8947	0.6076
ACTIVE[40]	ICCV 2023	0.1934	0.5868	0.6494	0.9081	0.3749	0.2376	0.8117	0.5374
RAUCA[61]	ICML 2024	0.1869	0.6553	0.5849	0.9090	0.3643	0.1987	0.8823	0.5402
GCAC[21]	IJCAI 2025	0.1207	0.5758	0.5760	0.9086	0.3084	0.1592	0.8644	0.5019
PGA[29]	ICCV 2025	0.1772	0.5934	0.5926	0.9080	0.3269	0.1951	0.8494	0.5204
GRAC[22]	ICCV 2025	0.1515	0.6875	0.6193	0.9091	0.3695	0.1992	0.8992	0.5479
RAUCA-E2E[62]	TDSC 2025	0.1121	0.5888	0.5089	0.9080	0.1303	0.1333	0.7792	0.4515
G-PAC	-	**0.1093**	**0.2849**	**0.3841**	**0.8174**	**0.1293**	**0.1229**	**0.5809**	**0.3470**

namely Cascade R-CNN with a ResNext-101-64x4d backbone [2]; and a transformer-based detector, specifically DINO utilizing a Swin-L backbone and 5-scale feature maps [55]. Furthermore, we evaluate three open-vocabulary foundation detectors, namely GLIP (Swin-L variant with Dynamic Head) [19], Detic (CenterNet2 with a Swin-B backbone, pre-trained on LVIS-COCO-IN21k) [63], and Grounding DINO with a Swin-L backbone [26]. These models demonstrate superior detection performance by leveraging extensive pre-training on massive multi-modal data to achieve robust zero-shot generalization.

Compared Methods. We select 8 state-of-the-art physical adversarial attack methods as our baseline for comparison, including FCA (AAAI-2022) [43], DTA (CVPR-2022) [39], ACTIVE (ICCV-2023) [40], RAUCA (ICML-2024) [61], GCAC (IJCAI-2025) [21], GRAC (ICCV-2025) [22], PGA (ICCV-2025) [29], RAUCA-E2E (TDSC-2025) [62].

Evaluation Metrics. We employ AP@0.5(%) to quantify the impact of various attack methods on detector performance. Following the convention in [39, 40, 61, 21, 22, 62, 29], this metric serves as a standard benchmark to account for both precision and recall at a detection IoU threshold of 0.5.

Training Details. In the camouflage generation phase, we employ the Adam optimizer with a learning rate of 0.01. Following the methodology of RAUCA-E2E [62], we utilize the NRP-E2E framework, integrated with a pre-trained EFE (Environment Feature Extractor), to facilitate differentiable rendering, relighting, and weather simulation. The hyper-parameters λ_1 and λ_2 are empirically set to 0.01 (following RAUCA-E2E) and 1, respectively. The vehicle masks are extracted directly from the semantic segmentation camera features within CARLA. For the calculation of z_g, we utilize DINOv3 ViT-S+/16 distilled [36] as the visual foundation model, which leverages a Vision Transformer (ViT) architecture to provide robust semantic features. The momentum α for updating the Global Adversarial Feature Anchor is set to 0.9 by default. The camouflage texture is initialized randomly and optimized over four epochs. All experiments are conducted on a NVIDIA A100 GPU.

4.2 Digital Experiments

In this section, we conduct a series of comprehensive digital-domain experiments to compare G-PAC with current SOTA baselines. These evaluations span multiple dimensions, providing a multifaceted demonstration that substantiates the distinct advantages of G-PAC from various perspectives.

Digital World Attack. We compare the digital attack performance of G-PAC against SOTA methods across various detectors and physical configurations. In these experiments, all methods utilize Yolo-V3 as the white-box surrogate model to guide the generation of adversarial camouflages. To rigorously evaluate generalization, we construct a test set encompassing a wide array of viewpoints, weather conditions, and shooting distances that include previously unseen physical configurations. We conduct white-box testing on YOLO-v3 and subsequently perform black-box transferability assessments on

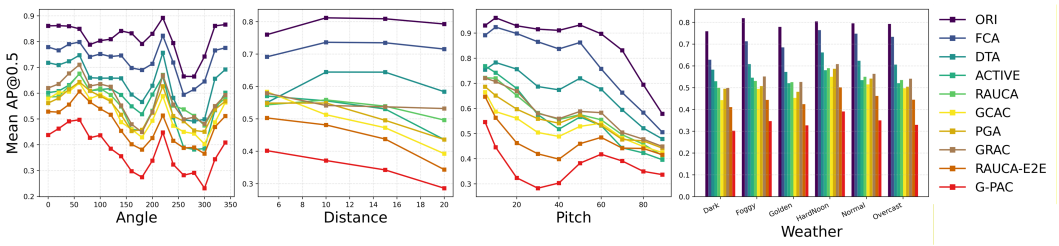


Figure 3: Robustness evaluation of various attack methods across diverse geometric and environmental configurations. The first three subplots illustrate the mean AP@0.5 as line curves across varying azimuth angles, distances, and pitch angles, respectively. The final subplot presents a bar chart evaluation across six distinct weather conditions.

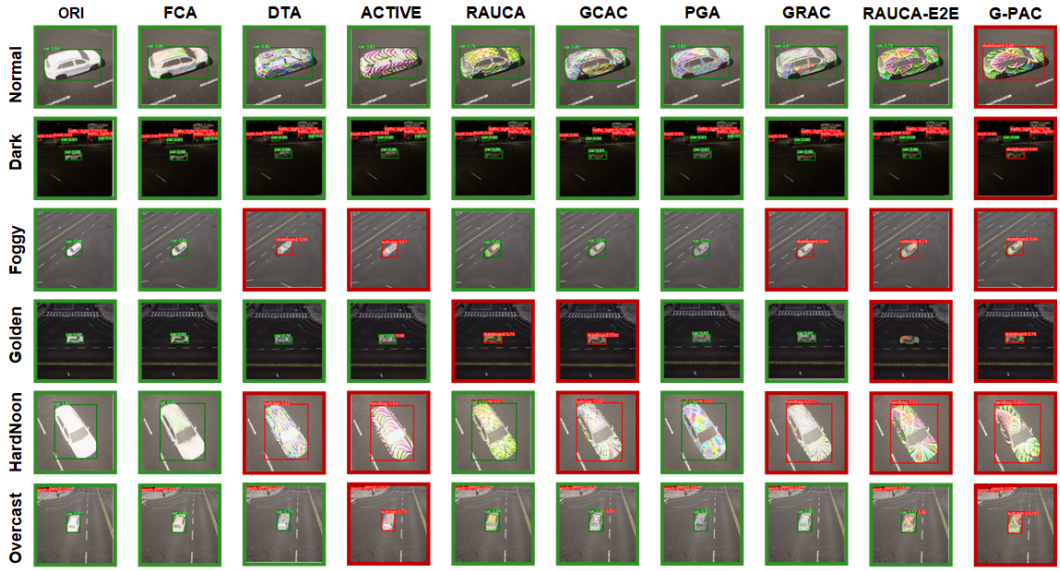


Figure 4: Visual comparison of detection results across different methods in the digital domain. We display camouflaged vehicles captured under diverse configurations, including varying pitch angles, azimuths, distances, and weather conditions. Green bounding boxes indicate correct detections (attack failure), while red bounding boxes denote incorrect detections (successful evasion).

six highly representative detectors.The results, summarized in Tab 1, demonstrate the superior performance of G-PAC. It is evident that foundation detectors such as GLIP*, Detic*, and DINO* possess high intrinsic robustness; even the strongest baseline, RAUCA-E2E, only manages to reduce their AP@0.5 to 0.5888, 0.5089, and 0.7792, respectively. In contrast, G-PAC achieves an additional AP@0.5 reduction of 0.3039, 0.1248, and 0.1983, successfully reaching much lower levels of 0.2849, 0.3841, and 0.5809. Furthermore, Grounding DINO* emerges as an extremely resilient black-box victim; none of the eight comparative methods can produce an effective attack, with their AP@0.5 values remaining near 0.90. However, G-PAC effectively breaks through this defense, lowering the AP@0.5 to 0.8174. Crucially, G-PAC achieves the lowest average AP@0.5 of 0.3470 across all seven evaluated models, representing a substantial improvement over the state-of-the-art RAUCA-E2E (0.4515) by a margin of 10.45%. These results provide compelling evidence of G-PAC's exceptional cross-architecture transferability and its robust effectiveness against advanced foundation models.

We further evaluate G-PAC's robustness across diverse configurations in Fig 3. The line plots indicate that while detection accuracy fluctuates with varying azimuths, distances, and pitches, G-PAC consistently maintains a substantial lead with minimal sensitivity to geometric changes. Notably, our method remains the most potent attack throughout the entire range as distance and pitch increase. Additionally, the bar chart confirms G-PAC's stable superiority across all evaluated weather conditions. Collectively, these multidimensional evaluations underscore the broad-spectrum robustness of G-PAC in various scenarios.

Visualization. Visualizations in Fig. 4 compare G-PAC with baseline methods, utilizing the open-vocabulary foundation detector GLIP (Swin-L) to assess the detection outcomes. To comprehensively

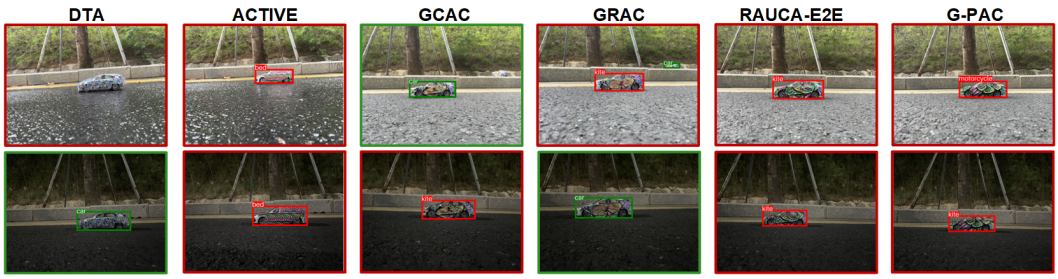


Figure 5: Visual comparison of detection results across different methods in the physical domain.

Table 2: Comparison of detection performance on the physical domain test set across diverse architectures. Each method utilizes adversarial textures generated via Yolo-V3 to evaluate black-box transferability on other detectors in physical environments. In each column, the best results are highlighted in **bold** and the second best results are underlined.

Method	Paper Year	White-box	Black-box						Average
			Foundation Detector			One-Stage	Two-Stage	Transformer-based	
		Yolo-V3	Glip*	Detic*	Grounding-DINO*	Yolo-X*	C-RCNN*	DINO*	
Clean	-	0.5915	0.9537	0.9044	0.9926	0.8742	0.7924	0.9850	0.8705
DTA[39]	CVPR 2022	0.3479	0.6020	0.7048	0.9610	0.6301	0.5656	0.9970	0.6869
ACTIVE[40]	ICCV 2023	0.2109	0.4825	0.6514	0.9352	0.2455	0.3052	0.9943	0.5464
GCAC[21]	IJCAI 2025	0.1998	0.4635	0.5974	0.9914	0.5343	0.3676	0.7008	0.5507
GRAC[22]	ICCV 2025	0.1866	0.4881	0.6466	0.9225	0.3506	0.3681	0.8153	0.5397
RAUCA-E2E[62]	TDSC 2025	0.1598	0.4350	0.7019	0.9743	0.1830	0.2514	0.6202	0.4751
G-PAC	-	**0.1133**	**0.2033**	**0.5331**	**0.8812**	**0.1399**	**0.1224**	**0.4899**	**0.3547**

validate the adversarial potency and cross-configuration robustness of G-PAC, we evaluate a diverse range of scenarios that encompass the various settings of pitch, azimuth, distance, and weather used in our quantitative analysis. The results indicate that G-PAC achieves superior adversarial effectiveness across these configurations compared to SOTA methods. Notably, G-PAC maintains consistent attack efficacy even in demanding configurations where other adversarial camouflages prove ineffective.

Ablation Study. We conduct extensive ablation experiments to verify each component of G-PAC. Beyond this, we investigate the influence of various foundation encoders (category and scale), the sensitivity of the contrastive loss weight, and performance on in-distribution data. Due to space constraints, detailed results and analysis are provided in the **Appendix**.

4.3 Physical Experiments

Quantitative and qualitative results for the physical domain are reported in Tab. 2 and Fig. 5, respectively. The experimental results indicate that G-PAC significantly compromises the performance of the victim detectors, reducing the average AP@0.5 from 0.8705 to 0.3547, which constitutes a substantial overall reduction of 0.5158. Furthermore, compared to the most competitive baseline, RAUCA-E2E (0.4751), our method achieves an additional average performance degradation of 0.1204. This demonstrates that the maintenance of a momentum-updated global pseudo-feature center as a stable adversarial anchor, combined with contrastive regularization, effectively facilitates coordinated object-level representation learning. Consequently, these mechanisms successfully guarantee the adversarial potency and cross-configuration generalization of G-PAC in real-world environments.

5 Conclusion

In this paper, we propose G-PAC to address the limited generalization of physical attacks across varying environmental configurations and unseen detection architectures. By leveraging a momentum-updated global pseudo-feature center and contrastive regularization, G-PAC transitions the optimization paradigm from independent view-specific suppression to coordinated object-level representation learning. Extensive evaluations in both digital and physical domains demonstrate that G-PAC achieves superior robustness and cross-architecture transferability across all tested detectors.

Acknowledgments and Disclosure of Funding

This work was supported by National Natural Science Foundation of China (No.62625602, U2541229), Shenzhen Science and Technology Program (No.KQTD20221101093559018, SYSRD20250529113401002, KJZD20240903095730039), Guangdong Provincial Key Laboratory of Information Security Technology.

References

[1] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[2] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[3] Y. Cao, S. H. Bhupathiraju, P. Naghavi, T. Sugawara, Z. M. Mao, and S. Rampazzi. You can't see me: Physical removal attacks on {lidar-based} autonomous vehicles driving frameworks. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2993–3010, 2023.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE, 2020.

[6] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[7] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[8] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. 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.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. 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.

[10] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.

[11] W. Feng, B. Wu, T. Zhang, Y. Zhang, and Y. Zhang. Meta-attack: Class-agnostic and model-agnostic physical adversarial attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7787–7796, 2021.

[12] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[15] Z. Hu, W. Chu, X. Zhu, H. Zhang, B. Zhang, and X. 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.

[16] L. Huang, C. Gao, Y. Zhou, C. Xie, A. L. Yuille, C. Zou, and N. Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 720–729, 2020.

[17] D. Kong, S. Liang, and W. Ren. Environmental matching attack against unmanned aerial vehicles object detection. *arXiv preprint arXiv:2405.07595*, 2024.

[18] D. Kong, S. Liang, X. Zhu, Y. Zhong, and W. Ren. Patch is enough: naturalistic adversarial patch against vision-language pre-training models. *Visual Intelligence*, 2(1):1–10, 2024.

[19] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[20] J. Lian, S. Mei, S. Zhang, and M. Ma. Benchmarking adversarial patch against aerial detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[21] J. Liang, S. Liang, J. Huang, C. Si, M. Zhang, and X. Cao. Physical adversarial camouflage through gradient calibration and regularization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1521–1529, 2025.

[22] J. Liang, S. Liang, T. Lou, M. Zhang, W. Li, D. Fan, and X. Cao. Gradient-reweighted adversarial camouflage for physical object detection evasion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13880–13889, 2025.

[23] S. Liang, L. Li, Y. Fan, X. Jia, J. Li, B. Wu, and X. Cao. A large-scale multiple-objective method for black-box attack against object detection. In *European Conference on Computer Vision*, 2022.

[24] S. Liang, W. Wang, R. Chen, A. Liu, B. Wu, E.-C. Chang, X. Cao, and D. Tao. Object detectors in the open environment: Challenges, solutions, and outlook. *arXiv preprint arXiv:2403.16271*, 2024.

[25] A. Liu, J. Guo, J. Wang, S. Liang, R. Tao, W. Zhou, C. Liu, X. Liu, and D. Tao. {X-Adv}: Physical adversarial object attacks against x-ray prohibited item detection. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

[26] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[27] Y. Liu, X. Ouyang, and X. Cui. Gleam: Enhanced transferable adversarial attacks for vision-language pre-training models via global-local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1665–1674, 2025.

[28] Z. Liu, Z. Yan, Q. Ning, Y. Lu, Z. Wang, and H. Wang. Naturalistic physical adversarial camouflage for object detection via differentiable rendering and style learning. *Pattern Recognition*, page 112621, 2025.

[29] T. Lou, X. Jia, S. Liang, J. Liang, M. Zhang, Y. Xiao, and X. Cao. 3d gaussian splatting driven multi-view robust physical adversarial camouflage generation. *arXiv preprint arXiv:2507.01367*, 2025.

[30] T. Lou, S. Liang, J. Liang, Y. Gao, and X. Cao. R-pga: Robust physical adversarial camouflage generation via relightable 3d gaussian splatting. *arXiv preprint arXiv:2603.26067*, 2026.

[31] K. Nguyen, T. Fernando, C. Fookes, and S. Sridharan. Physical adversarial attacks for surveillance: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[32] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[35] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[36] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[37] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018.

[38] J. Sun, W. Yao, T. Jiang, D. Wang, and X. Chen. Differential evolution based dual adversarial camouflage: Fooling human eyes and object detectors. *Neural Networks*, 163:256–271, 2023.

[39] N. Suryanto, Y. Kim, H. Kang, H. T. Larasati, Y. Yun, T.-T.-H. Le, H. Yang, S.-Y. Oh, and H. 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.

[40] N. Suryanto, Y. Kim, H. T. Larasati, H. Kang, T.-T.-H. Le, Y. Hong, H. Yang, S.-Y. Oh, and H. Kim. Active: Towards highly transferable 3d physical camouflage for universal and robust vehicle evasion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4305–4314, 2023.

[41] S. Thys, W. Van Ranst, and T. 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.

[42] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[43] D. Wang, T. Jiang, J. Sun, W. Zhou, Z. Gong, X. Zhang, W. Yao, and X. Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2414–2422, 2022.

[44] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. 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.

[45] J. Wang, X. Liu, Z. Yin, Y. Wang, J. Guo, H. Qin, Q. Wu, and A. Liu. Generate transferable adversarial physical camouflages via triplet attention suppression. *International Journal of Computer Vision*, pages 1–17, 2024.

[46] N. Wang, Y. Luo, T. Sato, K. Xu, and Q. A. 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.

[47] X. Wang, S. Mei, J. Lian, and Y. Lu. Fooling aerial detectors by background attack via dual-adversarial-induced error identification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[48] Y. Wang, L. Wu, Y. Cao, J. Jin, Z. Zhang, E. Wang, C. Ma, and Y. Zhao. A highly transferable camouflage attack against object detectors in the physical world. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[49] Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi. advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8341–8350, 2019.

[50] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021.

[51] X. Wei, S. Liang, N. Chen, and X. Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.

[52] T. Wu, X. Ning, W. Li, R. Huang, H. Yang, and Y. Wang. Physical adversarial attack on vehicle detector in the carla simulator. *arXiv preprint arXiv:2007.16118*, 2020.

[53] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[54] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 665–681. Springer, 2020.

[55] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*.

[56] J. Zhang, W. Wu, J.-t. Huang, Y. Huang, W. Wang, Y. Su, and M. R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14993–15002, 2022.

[57] J. Zhang, J. Ye, X. Ma, Y. Li, Y. Yang, Y. Chen, J. Sang, and D.-Y. Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909, 2025.

[58] X. Zhang, J. Chen, H. Zheng, and Z. Liu. Phycamo: A robust physical camouflage via contrastive learning for multi-view physical adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10230–10238, 2025.

[59] Y. Zhang, H. Foroosh, P. David, and B. Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2018.

[60] Y. Zhang, Z. Gong, Y. Zhang, K. Bin, Y. Li, J. Qi, H. Wen, and P. Zhong. Boosting transferability of physical attack against detectors by redistributing separable attention. *Pattern Recognition*, 138:109435, 2023.

[61] J. Zhou, L. Lyu, D. He, and Y. Li. Rauca: A novel physical adversarial attack on vehicle detectors via robust and accurate camouflage generation. In *International Conference on Machine Learning*, pages 62076–62087. PMLR, 2024.

[62] J. Zhou, L. Lyu, D. He, and Y. Li. Toward robust and accurate adversarial camouflage generation against vehicle detectors. *IEEE Transactions on Dependable and Secure Computing*, 2025.

[63] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022.

[64] H. Zhu and D. Rong. Multiview consistent physical adversarial camouflage generation through semantic guidance. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[65] Z. Zhu, X. Yang, H. Su, and S. Zheng. Camoenv: Transferable and environment-consistent adversarial camouflage in autonomous driving. *Pattern Recognition Letters*, 188:95–102, 2025.

Appendix

Structure of Appendix. The supplementary material is organized as follows. Appendix A provides comprehensive ablation studies on framework components and foundation encoders. Appendices B, C, and D detail the algorithmic pseudo-code, computational overhead, and adversarial loss formulation, respectively. Appendix E reports in-distribution evaluations, while Appendix F extends our evaluation to Vision-Language Models (VLMs). Appendix G provides theoretical analyses via variance reduction. Finally, Appendix H visualizes the optimized UV maps, Appendix I confirms the statistical stability of our method, and Appendix J discusses limitations and future work against commercial VLMs

A Ablation Study

Component-wise Analysis. To thoroughly investigate the individual contributions of the proposed mechanisms within G-PAC, we conduct a component-wise ablation study, as detailed in Tab 3. We systematically ablate the momentum-updated mechanism, the warm-up strategy, and the contrastive regularization ($\mathcal{L}_{con}$). As observed in the results, removing specific components can occasionally yield marginally better attack performance (i.e., lower AP@0.5) on certain individual detectors. For instance, some ablated variants achieve slightly lower AP@0.5 scores on the white-box surrogate YOLO-v3, as well as on Detic and Cascade R-CNN. This localized performance boost indicates that without the comprehensive constraints of our framework, the optimization tends to overfit to the inherent feature biases of specific architectures. However, this narrow overfitting comes at the severe cost of broader transferability. When all components are synergistically integrated, the complete G-PAC framework mitigates this degradation and achieves the most significant overall performance drop, yielding the lowest average AP@0.5 of 0.3470 across all seven evaluated models. Notably, the full framework demonstrates a profound capability to compromise highly resilient open-vocabulary foundation models, securing substantial performance reductions on GLIP, Grounding-DINO, and DINO. This confirms that while individual components might restrict aggressive overfitting to a single surrogate, their combination is indispensable for establishing stable semantic representations, ultimately leading to superior cross-architecture generalization in physical attacks.

Momentum	Warm-up	L_{con}	White-box	Black-box						Average
				Foundation Detector			One-Stage	Two-Stage	Transformer-based	
			Yolo-V3	Glip*	Detic*	Grounding-DINO*	YoloX*	C-RCNN*	DINO*	
			0.5251	0.8897	0.7693	0.9090	0.7395	0.6892	0.9093	0.7759
	✓	✓	**0.0724**	0.4549	**0.3513**	0.9055	0.1935	**0.1139**	0.6685	0.3943
✓		✓	0.0787	0.3925	0.3627	0.8270	0.2010	0.1221	0.6172	0.3716
✓	✓		0.2045	0.6664	0.5000	0.9077	0.1517	0.2045	0.8752	0.5014
✓	✓	✓	0.1093	**0.2849**	0.3841	**0.8174**	**0.1293**	0.1229	**0.5809**	**0.3470**

Table 3: Component-wise ablation study of the proposed G-PAC framework. We evaluate the individual contributions of the momentum-updated mechanism, the warm-up strategy, and the contrastive regularization ($\mathcal{L}_{con}$). Performance is measured in AP@0.5 across the white-box surrogate (Yolo-v3), three foundation detectors, and three standard black-box models. Lower AP@0.5 values indicate stronger adversarial generalization.

Impact of Foundation Encoder Capacity. To understand how the scale of the semantic prior affects attack performance, we evaluate G-PAC utilizing different variants of the DINOv3 encoder, ranging from the lightweight ViT-S/16 (21M) to the massive ViT-H+/16 (840M). As shown in the experimental data (Fig.6), the relationship between the encoder's parameter capacity and the overall adversarial generalization is not monotonically positive. The average AP@0.5 initially improves from 0.3872 (ViT-S/16) to an optimal 0.3470 using the distilled ViT-S+/16 (29M). However, as the parameter size scales up to 86M (ViT-B/16), 300M (ViT-L/16), and ultimately 840M (ViT-H+/16), the average AP@0.5 notably degrades to 0.4027, 0.4996, and 0.5063, respectively.

We attribute this phenomenon to two primary factors: semantic over-abstraction and representational misalignment. First, ultra-large foundation models are trained to capture highly abstract, invariant global semantics, rendering their latent spaces inherently insensitive to the localized, texture-level perturbations characteristic of physical camouflage. Consequently, utilizing them as a feature anchor

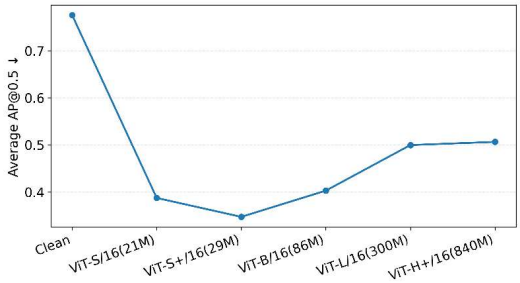


Figure 6: Impact of visual foundation encoder scale on adversarial generalization. We investigate the relationship between the parameter capacity of the foundation encoder (using various DINOv3 variants) and the average attack performance.

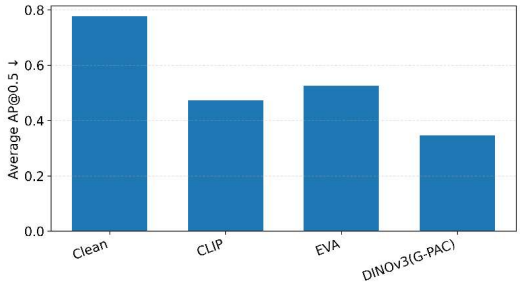


Figure 7: Impact of visual foundation encoder category on adversarial generalization. We compare the average AP@0.5 achieved by G-PAC when utilizing different foundation models (CLIP, EVA, and DINOv3) as the semantic prior.

may over-regularize the optimization process, suppressing the generation of aggressive adversarial patterns. Second, a massive foundation encoder introduces a significant capacity disparity compared to standard victim detectors. An overly complex anchor space may misalign with the feature spaces utilized by these standard architectures, thereby hindering cross-model transferability. The ViT-S+/16 (29M) model strikes the optimal balance, providing sufficient semantic stability to prevent feature dispersion while maintaining representational compatibility with the target detectors to facilitate potent gradient propagation.

Impact of Foundation Encoder Category. To validate the selection of the visual foundation encoder, we compare DINOv3 [36] against two other widely adopted foundation models: CLIP [34] and EVA [10]. Fig 7 illustrates the average detection AP@0.5 across all seven evaluated models when these different semantic priors are integrated into our framework. While both CLIP and EVA successfully reduce the average AP@0.5 compared to the clean baseline (0.7759), achieving 0.4740 and 0.5253 respectively, they fall significantly short of the 0.3470 average achieved by DINOv3. We attribute this performance gap to the inherent differences in their pre-training paradigms. CLIP is explicitly optimized to align images with textual descriptions, resulting in a latent space heavily biased toward high-level linguistic semantics, which often neglects the fine-grained visual textures essential for robust physical camouflage. EVA, which primarily relies on masked image modeling, excels at structural reconstruction but does not inherently enforce the tight feature clustering needed for an effective adversarial anchor. In contrast, DINOv3 utilizes self-distillation to learn instance-level discrimination, naturally forming highly compact and semantically coherent feature clusters for distinct objects. This specific property aligns perfectly with G-PAC's core objective of pulling scattered view-specific adversarial features into a stable, unified global representation, rendering DINOv3 the most optimal choice for our framework.

Impact of Contrastive Loss Weight. To determine the optimal balance between representation alignment and adversarial potency, we evaluate the impact of the contrastive loss weight (λ_2) on the overall attack performance. Fig 8 illustrates the average AP@0.5 across all seven detectors

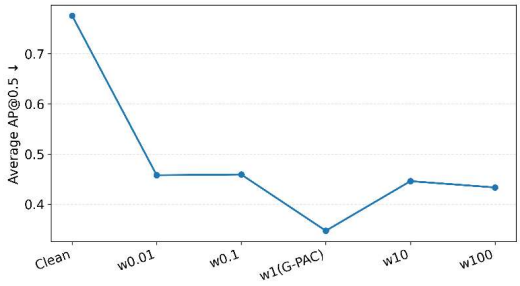


Figure 8: Impact of the contrastive loss weight on adversarial generalization. The line chart demonstrates the average AP@0.5 across seven victim models when varying the weight parameter λ_2 from 0.01 to 100. A weight of 1 achieves the optimal trade-off between representation alignment and adversarial potency.

when varying λ_2 from 0.01 to 100. The results exhibit a clear U-shaped trend. When the weight is set too low (e.g., 0.01 or 0.1), the contrastive regularization is insufficient to pull the scattered view-specific features toward the global anchor, resulting in suboptimal generalization with average AP@0.5 values of 0.4578 and 0.4592, respectively. Conversely, assigning an excessively high weight (e.g., 10 or 100) causes the representation alignment objective to dominate the optimization process. This over-regularization suppresses the primary adversarial loss ($\mathcal{L}_{atk}$), degrading the fundamental attack potency and causing the average AP@0.5 to rebound to 0.4459 and 0.4334. Setting $\lambda_2 = 1$ achieves the optimal equilibrium, effectively enforcing feature aggregation while maintaining robust adversarial strength, thereby yielding the lowest average AP@0.5 of 0.3470.

B Pseudo Code

Algorithm 1 formalizes the implementation of the G-PAC framework. The process involves rendering adversarial images across diverse configurations and extracting semantic features through a frozen foundation encoder. A momentum mechanism updates the global anchor to provide a stable optimization target. To maintain potency while ensuring feature aggregation, the contrastive loss is introduced after a designated warm-up period.

C Computational Overhead Analysis

G-PAC builds upon the RAUCA-E2E framework by integrating a frozen foundation encoder, specifically the DINOv3 ViT-S+/16 distilled variant, to facilitate mask-guided semantic extraction and contrastive alignment. Despite the addition of this auxiliary semantic prior, the computational overhead remains manageable. Efficiency evaluations conducted on a single NVIDIA A100 GPU demonstrate that G-PAC requires 3.6359 seconds per batch with a batch size of 8, while RAUCA-E2E consumes 3.2656 seconds under identical conditions. This indicates that the inclusion of the foundation encoder results in a marginal increase of 0.3703 seconds, which corresponds to an 11.34% increment in training time per iteration.

D Detailed Implementation of $\mathcal{L}_{atk}$

To suppress the responses of the surrogate detector, the adversarial objective $\mathcal{L}_{atk}$ computes a specialized detection loss based on the output of the target architecture. Following the formulation established in previous studies [61, 62], the attack loss is derived from a composite detection score designed to mislead the bounding box predictions. Given an augmented adversarial image x'_{adv} and its corresponding ground truth label y, the detector Det generates a set of bounding box predictions along with their confidence scores.

For each predicted bounding box, we first calculate an initial detection score by multiplying the objectness confidence with the class confidence. To ensure that the optimization focuses strictly on predictions relevant to the target vehicle, we incorporate the Intersection over Union (IoU) between

Algorithm 1 Generalizable Physical Adversarial Camouflage (G-PAC)

Require: 3D Mesh M, Initial UV Texture T, Original Texture T_{orig}, Target Detector Det, Frozen Foundation Encoder $f(\cdot)$, Differentiable Renderer $\mathcal{R}$, Environment Distribution $\mathcal{C}$.
Require: Hyper-parameters: α (momentum), τ (temperature), $\lambda_{1,2}$ (weights), E (total epochs), E_{warm} (warmup epochs).
Initialize global adversarial anchor $z_g \leftarrow$ None
for epoch $e = 1$ to E **do**
  for each batch in dataset **do**
    Sample a batch of background images and configurations $(x_{ref}, c) \sim \mathcal{C}$
    Render adversarial images: $x_{adv} = \mathcal{R}(M, T, x_{ref}, c)$
    Render clean images: $x_{clean} = \mathcal{R}(M, T_{orig}, x_{ref}, c)$
    Apply ROA augmentation to obtain $\{x'_{adv}, x'_{clean}\}$ and corresponding masks m'
    Extract object patches: $x_{patch} = \text{Resize}(\text{Crop}(x'_{adv} \odot m', \Omega_{m'}))$
    Compute embeddings: $z_{adv} = \text{Norm}(f(\psi(x_{patch})))$
    Extract clean embeddings z_{clean} from x'_{clean} using the same process
    Compute batch mean: $\bar{z}_{adv} = \text{Norm}(\frac{1}{B}\sum z_{adv})$
    if z_g is None **then**
      $z_g \leftarrow \bar{z}_{adv}$
    else
      $z_g \leftarrow \text{Norm}(\alpha \cdot z_g + (1-\alpha) \cdot \bar{z}_{adv})$
    end if
    Compute attack loss: $\mathcal{L}_{atk} = \text{ComputeLoss}(Det(x'_{adv}), y)$
    Compute smooth loss: $\mathcal{L}_{sm} = \text{TotalVariation}(T \odot m_{UV})$
    if $e > E_{warm}$ **then**
      Compute contrastive loss using Eq. 6:
      $\mathcal{L}_{con} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(z_{i,adv}^\top z_g/\tau)}{\exp(z_{i,adv}^\top z_g/\tau) + \sum_{j=1}^{B} \exp(z_{i,adv}^\top z_{j,clean}/\tau)}$
      $\mathcal{L}_{total} = \mathcal{L}_{atk} + \lambda_1 \mathcal{L}_{sm} + \lambda_2 \mathcal{L}_{con}$
    else
      $\mathcal{L}_{total} = \mathcal{L}_{atk} + \lambda_1 \mathcal{L}_{sm}$
    end if
    Update T via backpropagation: $T \leftarrow T - \eta \nabla_T \mathcal{L}_{total}$
  end for
end for
return Optimized texture T_{adv}

the predicted bounding box and the ground truth box as a spatial weight. Consequently, the composite detection score H_d for a prediction is defined as the product of the objectness score, the class confidence, and the IoU value. This weighting mechanism assigns a non-zero score only to bounding boxes that spatially intersect with the target object. Finally, the algorithm identifies the maximum composite score among all predictions and computes the adversarial loss using a logarithmic formulation: $\mathcal{L}_{atk} = -\log(1 - \max(H_d))$. By minimizing this objective via backpropagation, the framework effectively forces the surrogate detector to either misclassify the target vehicle or fail to detect it entirely.

E Evaluation on the In-Distribution Test Set

As discussed in Section 4.1, the primary digital experiments in the main text evaluate the adversarial camouflages on a highly challenging evaluation set with unseen configurations to rigorously assess out-of-distribution physical generalization. For completeness and as a supplementary sanity check, this section reports the attack performance on the standard in-distribution (i.i.d.) test set. This set is constructed by performing an 8:1:1 random split on the original 32,768 simulated images, ensuring that the test images follow the exact same environmental and geometric distribution as the training phase.

The detection results across the seven evaluated architectures are summarized in Tab.4. As expected, the overall performance trends perfectly corroborate the out-of-distribution results presented in the

main text. G-PAC consistently maintains its superiority, achieving the lowest average AP@0.5 of 0.2708. More importantly, when targeting the highly resilient open-vocabulary foundation models in this in-distribution setting, G-PAC induces an even more pronounced performance degradation compared to the baseline methods. Specifically, our framework reduces the AP@0.5 of GLIP and Grounding-DINO to 0.3085 and 0.6814, respectively, establishing a substantial margin over the SOTA RAUCA-E2E. These supplementary results confirm that while G-PAC is explicitly designed for robust cross-configuration generalization, it simultaneously maximizes its adversarial potency within the observed training distribution.

Method	Paper Year	White-box	Black-box						Average
			Foundation Detector			One-Stage	Two-Stage	Transformer-based	
		Yolo-V3	Glip	Detic	Grounding-DINO	Yolo-X	C-RCNN	DINO	
Clean	-	0.5684	0.9641	0.8483	0.9936	0.8274	0.7649	0.9492	0.8451
FCA[43]	AAAI 2022	0.4252	0.9406	0.7119	0.9404	0.6007	0.5040	0.9186	0.7202
DTA[39]	CVPR 2022	0.4655	0.9195	0.7215	0.9404	0.6125	0.5607	0.9454	0.7379
ACTIVE[40]	ICCV 2023	0.3863	0.9098	0.8011	0.934	0.4798	0.4296	0.8566	0.6853
RAUCA[61]	ICML 2024	0.2244	0.8491	0.5597	0.9107	0.3018	0.1925	0.7864	0.5464
GCAC[21]	IJCAI 2025	0.2958	0.8831	0.6448	0.9349	0.4748	0.3429	0.8753	0.6359
GRAC[22]	ICCV 2025	0.3304	0.9110	0.6467	0.9437	0.4748	0.3608	0.9086	0.6537
RAUCA-E2E[62]	TDSC 2025	0.1003	0.6655	0.5005	0.7655	**0.0654**	0.0939	0.4884	0.3828
G-PAC	-	**0.0882**	**0.3085**	**0.4115**	**0.6814**	0.0674	**0.0313**	**0.3076**	**0.2708**

Table 4: Comparison of detection performance (AP@0.5) on the in-distribution test set. This table evaluates the adversarial potency of various physical attack strategies across seven distinct architectures under the exact same environmental and geometric distribution as the training phase.

F Evaluation on Vision-Language Models (VLMs)

	Clean	FCA	DTA	ACTIVE	RAUCA	GCAC	GRAC	RAUCA-E2E	G-PAC
llava15-7b*	0.7196	0.6406	0.6018	0.6362	0.6721	0.6485	0.6553	0.5487	**0.4116**

Table 5: Comparison of detection performance (AP@0.5) against the Vision-Language Model (LLaVA-1.5-7B). The results demonstrate the superior transferability of G-PAC against robust multimodal architectures.

To comprehensively assess the threat of our physical camouflage against next-generation multimodal architectures, we extend our evaluation to VLMs. Specifically, we test the proposed G-PAC and baseline methods against LLaVA-1.5 (7B), a highly representative open-source VLM.

Experimental Setup. Because VLMs generate natural language responses rather than standardized bounding box tensors, we formulate the object detection task as a two-stage Visual Question Answering (VQA) and grounding pipeline. To rigorously evaluate cross-modal transferability, the adversarial camouflages for all compared methods are optimized using the white-box YOLO-v3 surrogate detector (consistent with our main digital experiments) and subsequently transferred to the LLaVA-1.5-7B model for black-box testing. The pipeline consists of the following stages:

- **Stage 1: Existence Judgment.** We first query the VLM to determine the presence of the target object using the prompt: *"Is there a car present in this image? Answer strictly with 'Yes' or 'No'."* If the model outputs "No" or fails to provide a valid affirmative response, it is deemed to have missed the target entirely, and the attack is recorded as successful. If the response contains "Yes", the pipeline proceeds to the second stage.
- **Stage 2: Box Grounding.** To localize the identified vehicle, we utilize the prompt: *"Locate the car in this image and output its bounding box."* The generated text is then systematically parsed to extract the coordinate values, which are normalized to the $[0, 1]$ spatial range.

Based on the bounding boxes successfully extracted from this two-stage process, we assign a true positive (TP) flag if the Intersection over Union (IoU) with the ground truth exceeds 0.5. The final performance is measured using the standard AP@0.5 metric.

Results and Analysis. The evaluation results are presented in Tab. 5. On clean, unmodified images, LLaVA-1.5-7B achieves an AP@0.5 of 0.7196. This indicates that while VLMs may not match the absolute localization precision of dedicated dense object detectors, they exhibit a profound inherent robustness against texture-level perturbations. This resilience is largely attributed to their massive multimodal pre-training, which encourages the model to rely on broader contextual semantics rather than localized visual features.Consequently, conventional physical attacks struggle to bypass the semantic priors of VLMs. Baseline methods such as RAUCA, GRAC, and GCAC cause only marginal performance degradation, maintaining AP@0.5 scores around 0.64 to 0.67. Even the most competitive baseline, RAUCA-E2E, only reduces the AP@0.5 to 0.5487.Despite this immense architectural robustness, G-PAC significantly outperforms all baselines, successfully driving the AP@0.5 down to **0.4116**. This superior performance is a direct result of G-PAC's coordinated object-level representation learning. Rather than relying on fragile, view-specific noise that VLMs can easily ignore, G-PAC leverages a momentum-updated global anchor and contrastive alignment to forge cohesive, large-scale adversarial pseudo-features. These stable semantic patterns are robust enough to distort the fundamental visual representations extracted by the VLM, proving that G-PAC's optimization paradigm remains highly potent even against advanced, context-aware multimodal systems.

G Theoretical Analysis

The effectiveness of G-PAC in achieving cross-configuration generalization can be mathematically attributed to explicit variance reduction within the semantic representation space. In standard view-level optimization paradigms, optimizing the adversarial texture T_{adv} purely against detection loss across discrete configurations $c \in \mathcal{C}$ lacks constraints on the resulting intermediate representations. This unconstrained optimization inevitably leads to high variance among the adversarial features extracted across different views, manifesting as semantic dispersion. G-PAC formally mitigates this issue through the contrastive objective $\mathcal{L}_{con}$ defined in Eq. 6. Since the individual adversarial embeddings $z_{i,adv}$ and the global anchor z_g are mathematically constrained to be L_2-normalized vectors on a unit hypersphere, maximizing their dot product $z_{i,adv}^{\top} z_g$ in the InfoNCE numerator is mathematically equivalent to minimizing their squared Euclidean distance $||z_{i,adv} - z_g||_2^2$. Therefore, optimizing $\mathcal{L}_{con}$ explicitly bounds the variance of the adversarial features across the configuration distribution $\mathcal{C}$. By mathematically enforcing this variance reduction, G-PAC guarantees that the scattered view-specific features aggregate robustly toward the stable anchor z_g, resulting in a configuration-invariant representation that naturally exhibits superior transferability.

H UV Map Visualization

To intuitively illustrate the qualitative differences between the adversarial textures generated by various methods, we visualize the corresponding flat UV maps applied to the target vehicle mesh, as shown in Fig 9.

Visual inspection reveals a striking distinction in the structural characteristics of the generated camouflages. Baseline methods such as FCA, GCAC, GRAC, and RAUCA-E2E predominantly produce high-frequency, fragmented noise patterns. These textures lack coherent semantic structures and appear as disjointed pixel aggregations tailored to suppress local, view-specific responses. Such fragmented patterns are inherently fragile and struggle to maintain consistency across diverse configurations or transfer effectively to robust foundation models.

In contrast, the UV map optimized by the proposed G-PAC framework exhibits remarkably distinct, large-scale semantic structures. Rather than relying on scattered noise, G-PAC constructs continuous, organic-like textures with high contrast and well-defined macroscopic patterns (e.g., distinct leaf-like or scale-like structures) that span across different vehicle components coherently. This visual coherence directly reflects the theoretical objective of G-PAC: transitioning from isolated view-level suppression to coordinated object-level representation learning. By forcing the adversarial features to aggregate toward a stable global anchor, the optimization naturally gravitates toward generating these robust, unified, and semantically dense pseudo-features, which ultimately ensures the superior physical generalization demonstrated in our experiments.

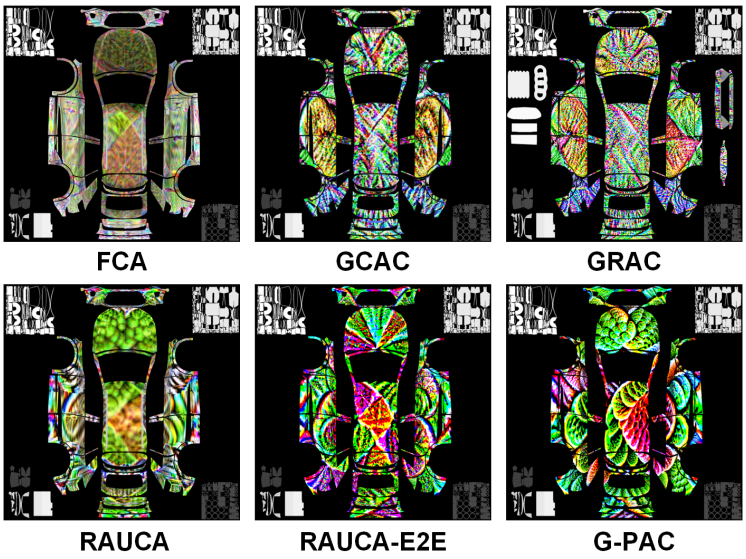


Figure 9: Visualization of optimized UV maps across different attack methods.

I Statistical Stability.

To verify that the performance gains of G-PAC are not artifacts of a specific random initialization, we repeatedly optimized the adversarial texture five times using different random seeds. Under the standard evaluation setting (YOLO-v3 as surrogate), the standard deviation of the resulting Average AP@0.5 through all detectors was within ± 0.008. This low variance demonstrates that G-PAC is highly stable across different initializations and that the reported performance improvements over baseline methods are statistically significant.

J Limitation and Future Work

Despite the significant improvements in cross-architecture generalization achieved by G-PAC, our method, along with existing physical camouflage paradigms, exhibits inherent limitations when targeting massive, closed-source commercial vision-language models (VLMs). These advanced foundation models demonstrate profound resilience against texture-based physical attacks due to several intersecting factors. First, they possess powerful contextual reasoning capabilities, allowing them to confidently infer the presence and category of a vehicle strictly from environmental cues, such as roads, driving lanes, and surrounding traffic dynamics. Second, practical physical attacks are strictly constrained by vehicle functionality, inevitably leaving critical components like wheels, glass windows, and roof racks exposed. Commercial VLMs seamlessly utilize these uncamouflaged sub-structures to recover the vehicle's identity, effectively bypassing the adversarial texture. Lastly, the unprecedented scale of their pre-training data endows them with strong semantic priors, making them naturally robust to adversarial noise patterns applied exclusively to the object's surface.

These limitations highlight a critical operational boundary for object-centric physical camouflage. To bridge this gap, our future work will pivot from localized texture optimization to exploring macroscopic contextual disruption. We aim to investigate whether manipulating the broader physical context can effectively deceive commercial VLMs in the physical domain. This direction will further probe the fundamental vulnerabilities of next-generation multi-modal autonomous systems and evaluate their true robustness in complex, real-world deployment scenarios.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly define the research scope and main contributions in the abstract and introduction, both of which are supported by experimental results.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 discuss in detail in Appendix J the limitations of the method.

Guidelines:

- The answer [N/A] 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: [Yes]

   Justification: We provide rigorous theoretical analysis based on variance reduction and normalization in Appendix G, with all assumptions derived from the formulas in the main text.

   Guidelines:

   - The answer [N/A] 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 disclose all hyper-parameters, model architectures, and procedures in Section 4.1 (Training Details) and Appendix B (Algorithm 1, Pseudo code).

Guidelines:

- The answer [N/A] 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: [No]

   Justification: The code will be released after acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Relevant details are elaborated in Lines 274-283 of the main text (training details) and Appendix D (implementation details).

   Guidelines:

   - The answer [N/A] 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: We conducted multiple runs for our main framework with different random initializations to assess optimization stability in Appendix I.

   Guidelines:

   - The answer [N/A] 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: We specify the use of a single NVIDIA A100 GPU on Line 283 and report the detailed training time in Appendix C (Lines 580-588).

   Guidelines:

- The answer [N/A] 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 aims to enhance the security of AI systems by evaluating attacks, comply with ethical guidelines, and avoid bias or unauthorized exploitation.

   Guidelines:

   - The answer [N/A] 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: In the introduction (Lines 25-33), we point out the potential safety hazards in autonomous driving scenarios, and this research helps enhance the defense capability of future systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This project does not involve the release of high-risk pre-trained models or sensitive datasets that may be misused.

Guidelines:

- The answer [N/A] 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 adopted the open-source CARLA simulator and pre-trained models such as DINOv3, with proper citations of the original literature.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The main contribution of this project lies in the algorithm framework, with no new large-scale datasets or model weights publicly released.

Guidelines:

- The answer [N/A] 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: [N/A]

    Justification: This study is entirely based on computer simulations and scaled model tests and does not involve human experiments.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This study is entirely based on computer simulations and scaled model tests and does not involve human experiments.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is only used to assist in paper writing and polishing, without altering the core research methodology or experimental rigor.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.