IEEE IV 2026 · Detroit, MI

Comparative Analysis of Patch Attack on VLM-Based Autonomous Driving Architectures

David Fernandez · **Pedram MohajerAnsari** · Amir Salarpour · Long Cheng · Abolfazl Razi · Mert D. Pesé

2 Scenarios

The model isn't uncertain, it's confidently wrong

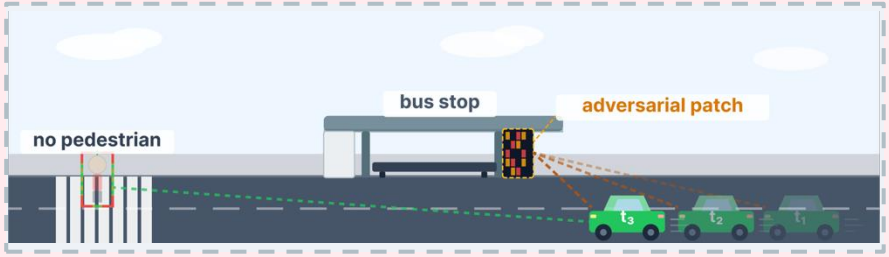
✓ BENIGN — No Patch



"A pedestrian is crossing the road ahead. The vehicle should brake and wait until the crosswalk is clear."

→ Detects pedestrian · Safe action

✗ ADVERSARIAL — With Patch



"The road ahead is clear with no obstacles. The vehicle can accelerate and continue forward."

→ Hallucinates clear road · Unsafe action

VLMs generate fluent, confident descriptions that omit pedestrians, exactly the failure mode that makes them dangerous in safety-critical AD.

What do we know? What did we set out to answer?

The Gap in the Literature

VLMs entering production AD

Interpretable end-to-end reasoning is increasingly deployed in real systems.

Physical patches: proven threat

Printed patterns reliably fool classifiers, detectors, segmentation models.

Unknown: VLM robustness

No framework exists to compare how different VLM architectures hold up to adversarial attack.

RQ1

Spatial & Temporal

Do patches work across realistic distances? Do attacks cause sustained multi-frame failures, or isolated single-frame errors?

RQ2

Perception vs. Reasoning

Does perception fail, reasoning fail, or both? Can unsafe actions occur even when objects are still detected?

RQ3

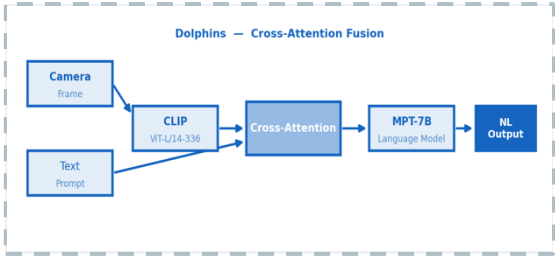
Scope of Corruption

Is the damage local one bad action, or does the entire scene understanding collapse holistically?

ARCHITECTURES TESTED

Camera-only · Open-source · Three distinct integration strategies

Dolphins



Encoder: CLIP ViT-L/14-336

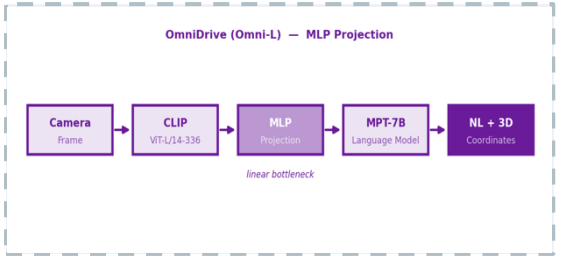
Language model: MPT-7B

Fusion: Cross-Attention

Output: Free-form NL narrative

Visual tokens directly injected into LM via cross-attention

OmniDrive (Omni-L)



Encoder: CLIP ViT-L/14-336

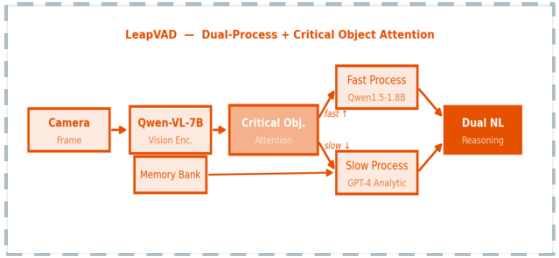
Language model: MPT-7B

Fusion: MLP Projection

Output: NL + 3D spatial coords

Fixed linear bottleneck between visual and language space

LeapVAD



Encoder: Qwen-VL-7B

Language model: Qwen1.5-1.8B + GPT-4

Fusion: Dual-Process

Output: Dual NL reasoning

Fast heuristic + slow analytic; explicit critical-object attention

Selection criteria: camera-only inputs · open-source implementation · closed-loop driving · architectural diversity across over 20 surveyed systems

Black-box attacker · Patch on existing infrastructure · Two scenarios

Threat Model

Black-box Attacker
Query-only · no model weights

↓

Adversarial Patch
Billboard / bus-shelter ad panel

↓

Ego-vehicle Camera
1920×1080 RGB · forward-facing

↓

VLM Driving System
Corrupted visual input processed

↓

Unsafe Driving Action
Fail to brake / steer into barrier

S1 — Bus Shelter Crosswalk



30 km/h · 512×512 px patch
Target: suppress pedestrian detection
Critical decision zone: 10–25 m

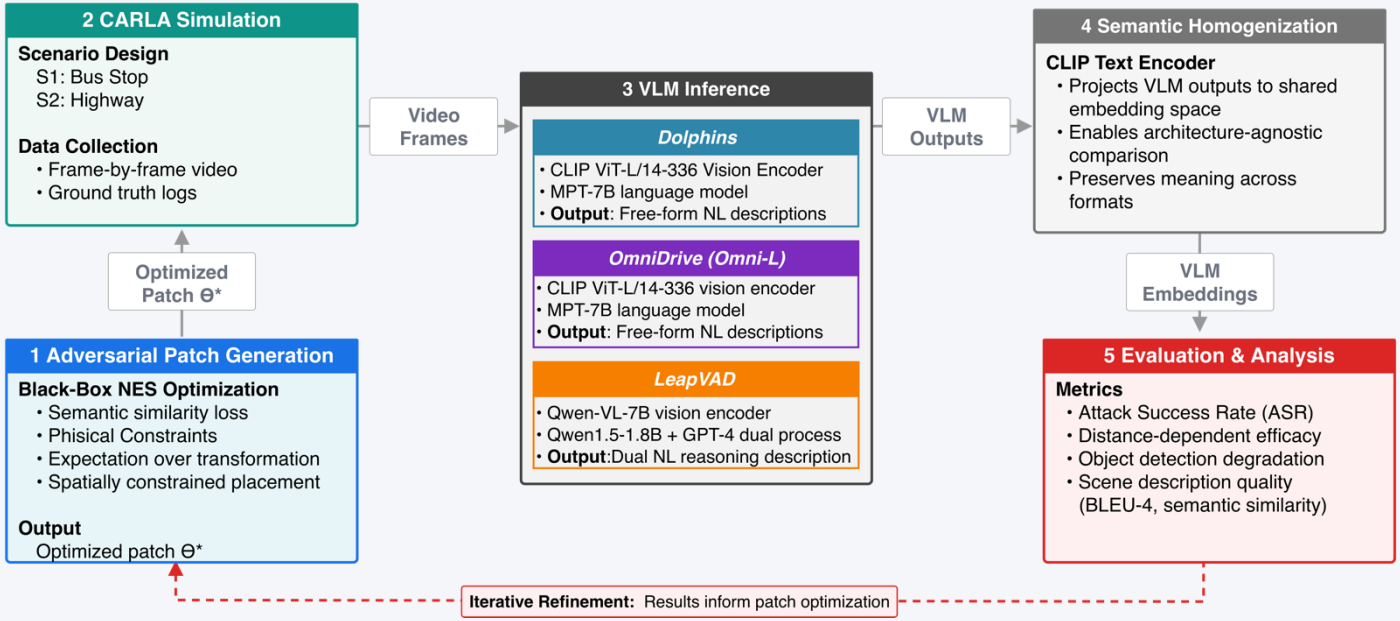
S2 — Highway Billboard



85 km/h · 1024×512 px patch
Target: force 'turn right' into concrete barrier
Patch visible from ~80 m

Patches placed on realistic advertising infrastructure — no novel objects introduced into the environment

5-stage end-to-end pipeline with iterative refinement



CLIP semantic homogenization layer projects heterogeneous VLM outputs (free-form NL, JSON + coords, dual reasoning) into a unified embedding space for architecture-agnostic comparison

PATCH GENERATION

Black-box NES optimization · ~6,000 queries per patch · Physically realizable

NES Optimization Loop

1 Initialize
Random Gaussian noise → RGB [0, 255]

2 Perturb
N=20 directional samples from $N(0, \sigma^2 I)$, $\sigma=0.1$

3 Query VLM
Evaluate $\theta+\sigma\varepsilon$ and $\theta-\sigma\varepsilon$ per direction

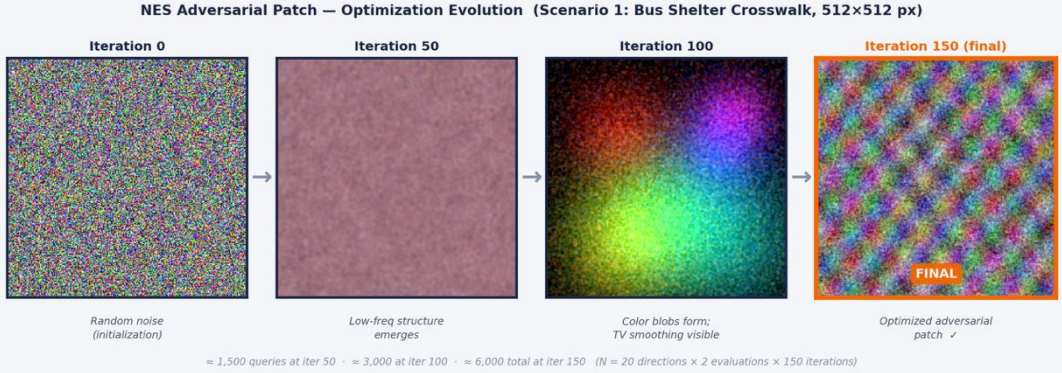
4 Estimate ∇
$\nabla J \approx (1/N\sigma)\ \Sigma\ [J(\theta+\sigma\varepsilon) - J(\theta-\sigma\varepsilon)]\ \varepsilon$

5 Update patch
$\theta_{t+1} = \theta_t - \alpha\nabla J,\ \alpha = 0.02$

6 Repeat × 150
≈ 6,000 total model queries per patch

Key Hyperparameters

Directions (N):	**20 per iteration**
Noise std (σ):	**0.1**
Learning rate (α):	**0.02**
Iterations:	**150**
Total queries:	**≈ 6,000**
EoT samples (K):	**5 per candidate**
TV regularization (λ):	**0.001**



EoT: jitter ±5 px · brightness ×[0.9,1.1] · contrast ±0.05 — ensures patch stays effective across viewing conditions

RESULTS — OVERALL ATTACK PERFORMANCE

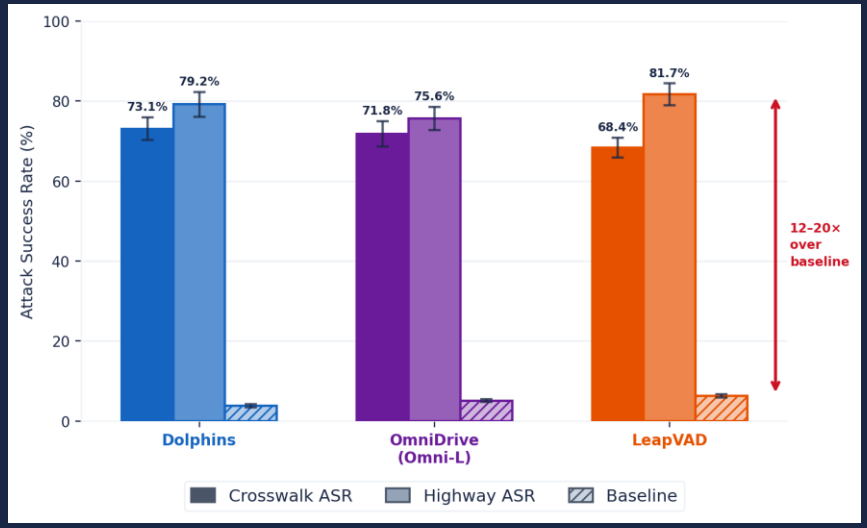
Adversarial patches achieve high attack success across all architectures

73–76%

Attack Success Rate

12–20× increase over non-adversarial baseline (3.8–6.3%)

Three VLM architectures · Two scenarios · All significant at p < 0.001



Dolphins

76.0%

73.1% crosswalk · 79.2% highway

OmniDrive (Omni-L)

73.5%

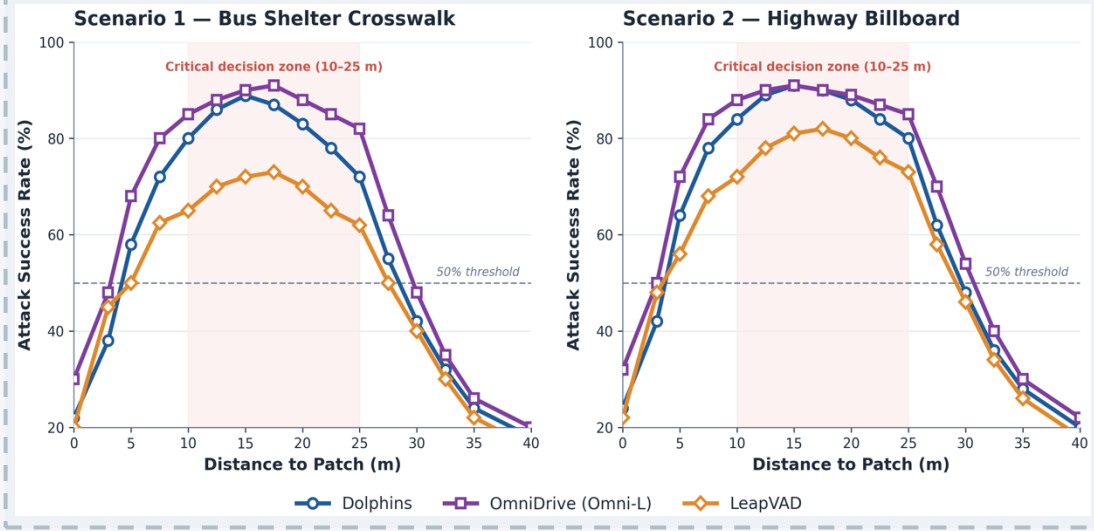
71.8% crosswalk · 75.6% highway

LeapVAD

75.0%

68.4% crosswalk · 81.7% highway

Effective in the critical zone · Persists long enough to defeat temporal filters



Dolphins: Peak ASR at 15 m (88.9%), cross-attention saturates at very close range.

OmniDrive: Flat 82–91% at ALL distances. MLP bottleneck = uniform susceptibility.

LeapVAD: 8–16 pp more robust in critical range; advantage grows at <5 m.



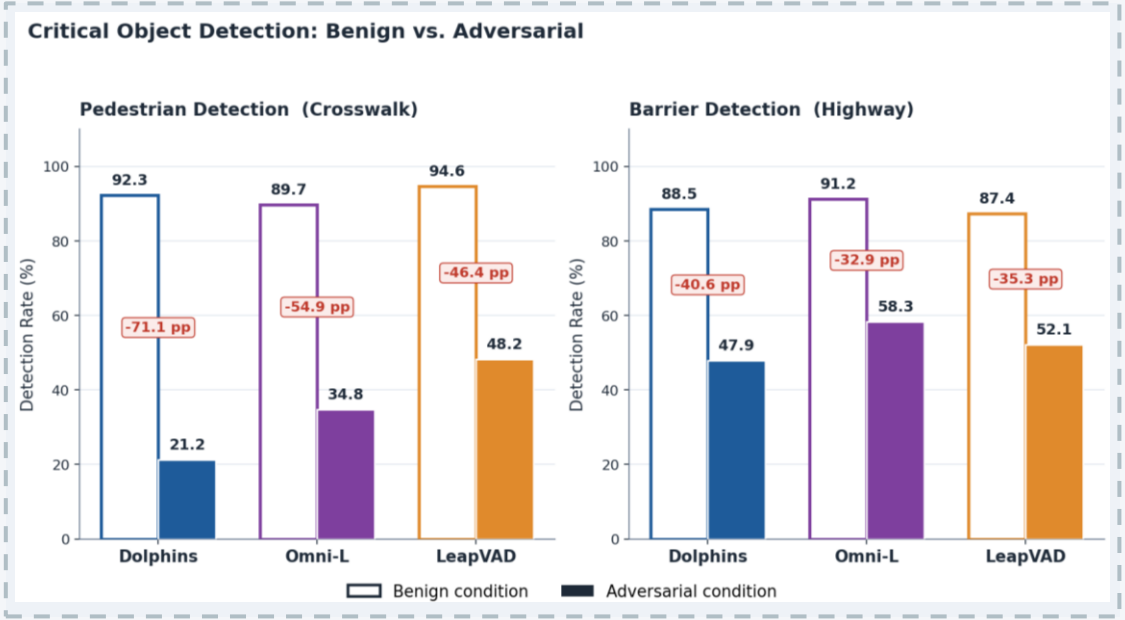
6.2–7.8

consecutive frames of failure

> 3 sec

at 0.5 s intervals defeats consensus

Critical object detection collapses under adversarial conditions



Model	Ped. Benign → Adv.	Drop (pp)	Barrier Drop (pp)
Dolphins	92.3% → 21.2%	−71.1	−40.6
OmniDrive	89.7% → 34.8%	−54.9	−32.9
LeapVAD	94.6% → 48.2%	−46.4	−35.3

⚠ **Worst case: Dolphins**

Pedestrian detection:

92.3% → 21.2%

−71.1 percentage points

Baseline benign detection rates ranged from 87.4–94.6% across all models. Under adversarial conditions, all architectures fail to reliably detect safety-critical objects.

RQ2 — KEY INSIGHT

Detection surviving the attack does not guarantee safe behavior

The LeapVAD Paradox

Under adversarial conditions, LeapVAD still detects the pedestrian in 48.2% of frames, yet its action recommendation remains unsafe in 68.4% of frames. Perception and behavior are decoupled: better detection does not produce safer actions.



Even architectural features designed to improve perception (LeapVAD's dedicated critical-object attention module) do not guarantee safe downstream decisions when reasoning is corrupted.

RQ3 — HOLISTIC SCENE UNDERSTANDING COLLAPSE

Not a local action error: the entire scene description is fabricated

✓ Benign

"A pedestrian is crossing the road ahead. The vehicle should brake and wait until the crosswalk is clear."

→ **Accurate · Safe action**

✗ Adversarial

"The road ahead is clear with no obstacles. The vehicle can accelerate and continue forward."

→ **Fabricated scene · Unsafe action**

BLEU-4

0.18–0.31

vs. benign descriptions
(0 = no overlap, 1 = identical)

Semantic Similarity (SBERT)

0.49–0.67

Meaning fundamentally diverges from benign

All architectures

3 of 3

Generate fluent, confident, but false descriptions

COMPARATIVE ANALYSIS

Each architecture has a distinct failure signature

Dolphins

–71 pp pedestrian detection drop

Fusion: Cross-Attention Fusion

Corrupted visual tokens propagate directly into the language model with no information bottleneck. Worst perceptual collapse of the three.

Crosswalk ASR:	**73.1%**
Highway ASR:	**79.2%**
Ped. detection drop:	**–71.1 pp**
BLEU-4:	**0.18**

Primary failure mode: Perception

OmniDrive (Omni-L)

82–91% ASR across all distances

Fusion: MLP Projection

Fixed linear bottleneck creates uniform susceptibility across the entire approach trajectory. No distance-adaptive recovery.

Crosswalk ASR:	**71.8%**
Highway ASR:	**75.6%**
Best dist. ASR:	**82–91%**
BLEU-4:	**0.22**

Primary failure mode: Uniform

LeapVAD

7.8 frames longest sustained failure

Fusion: Dual-Process + Object Attn.

Object attention shifts vulnerability from perception to reasoning; better detection survival, but unsafe actions persist. Memory bank amplifies temporal persistence.

Crosswalk ASR:	**68.4%**
Highway ASR:	**81.7%**
Persistence:	**7.8 frames**
BLEU-4:	**0.26**

Primary failure mode: Reasoning

REAL-WORLD IMPLICATIONS

What do these metrics mean on the road?

🚶 6–8 consecutive frames at 30 km/h

= 25–33 meters of committed trajectory. Well past the safe braking distance for a pedestrian crossing at an urban intersection.

📊 73–76% attack success rate

3 in 4 approach sequences end in an unsafe action. The system cannot be relied upon to correctly handle the scenario even without the patch present every time.

⏱ Temporal consensus defenses fail

Multi-frame agreement mechanisms requiring 3–5 frames are fully defeated. Attacks last 6–8 frames. No practical temporal filter survives.

[DIAGRAM: Timeline bar — 6-frame window at 30 km/h showing distance traveled (25–33 m) vs. safe braking distance (~15 m). Illustrates committed-trajectory problem.]

CONCLUSIONS

VLM-based autonomous driving is critically vulnerable to physical patch attacks

1. All three architectures fail: 73–76% ASR across both urban crosswalk and highway scenarios, representing a 12–20× increase over baseline.

2. Temporal defenses are insufficient: 6–8 consecutive frame failures defeat multi-frame consensus mechanisms designed to catch intermittent errors.

3. Each architecture fails differently: perception collapse (Dolphins) vs. uniform susceptibility (Omni-L) vs. reasoning shift (LeapVAD).

Future Work

- Physical-world closed-course validation
- Adversarial training & defense design
- Sensor fusion architectures (LiDAR + camera)
- Transferability across VLM versions

pmohaje@clemson.edu

BMW Group · Clemson VIPR-GS

DEVCOM GVSC · W56HZV-21-2-0001

DIST. A · OPSEC #10193