

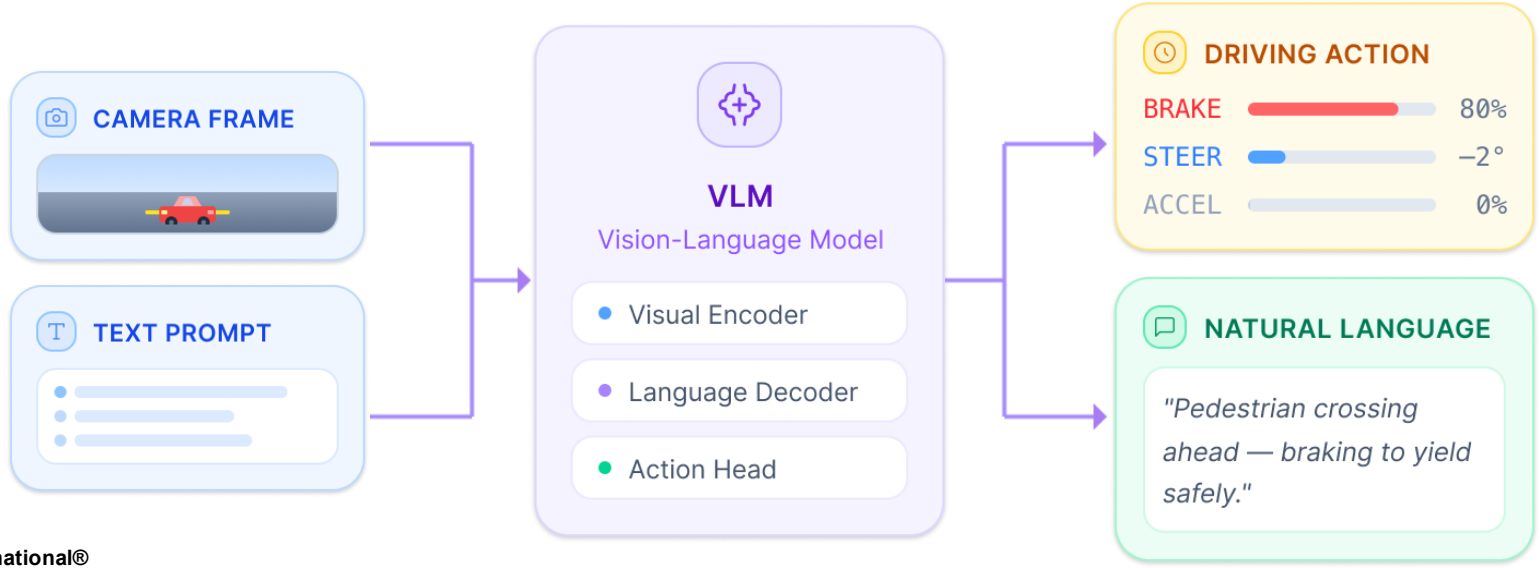
Understanding Adversarial Transferability in Vision-Language Models for Autonomous Driving: A Cross-Architecture Analysis

CLEMSON UNIVERSITY

TIGERSEC

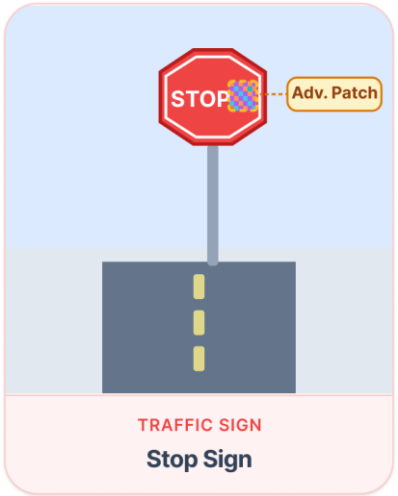
VLMs Are Reshaping Autonomous Driving

- VLMs combine visual perception + language reasoning
- Enable interpretable decisions: the car can "explain" why it brakes
- 20+ different VLM architectures now proposed for driving

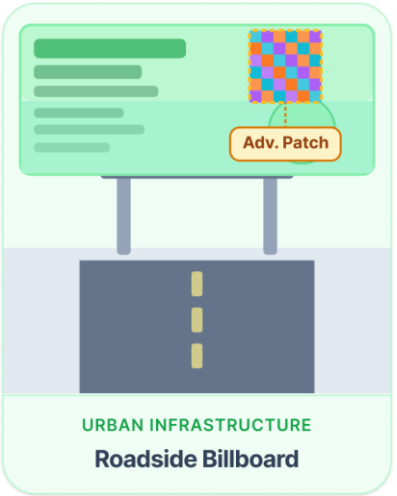


Physical Adversarial Patches: A Real-World Threat

- Adversarial patches: printed patterns that deceive AI systems
- Placed on everyday infrastructure
- Already demonstrated against classifiers and detectors

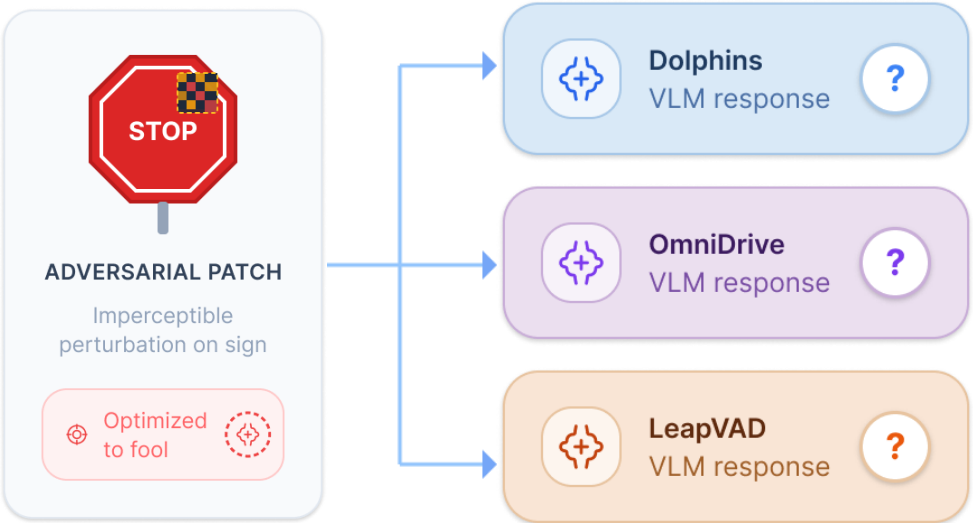






What About Transferability?

- Attackers don't know which VLM a vehicle uses
- **Can a single patch fool vehicles from different manufacturers?**



What we Investigate



RQ1: Cross-Architecture Transferability How well do adversarial patches optimized for one VLM transfer to other architectures?



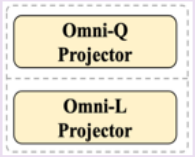
RQ2: Architecture-Specific Vulnerabilities Which VLM designs are more robust or more vulnerable to transfer attacks?



RQ3: Temporal Consistency Do transferred attacks maintain sustained effectiveness over time, frame-by-frame?

Architectural VLM Diversity

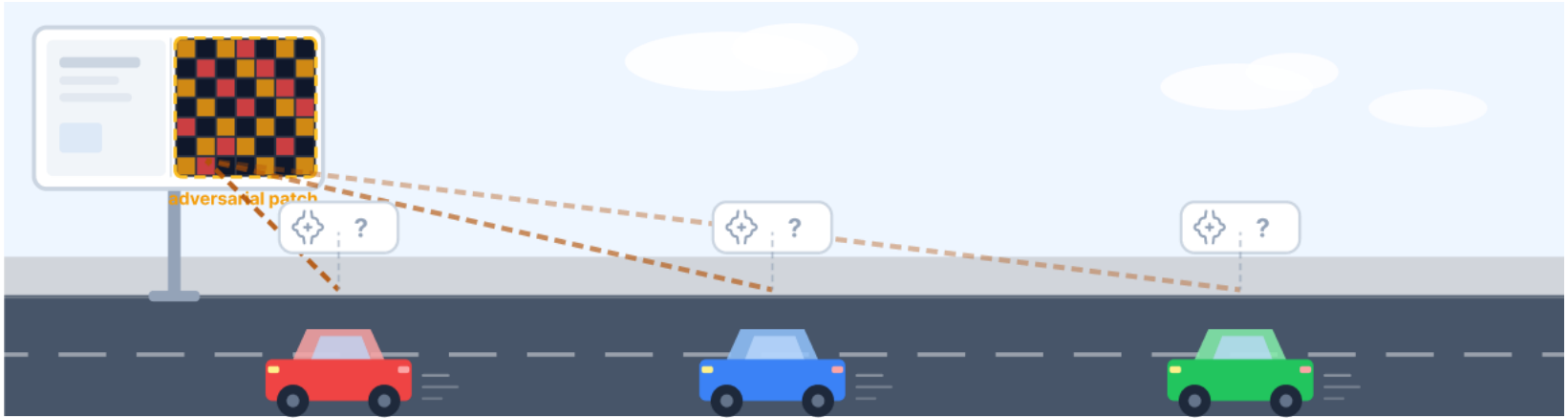
- 20+ different VLM architectures now proposed for driving [1]

	Dolphins	OmniDrive	LeapVAD
			
Vision Encoder	CLIP ViT-L/14	LLM: MPT-7B	Cross-attention
Language Model	EVA-02-L	LLaMA2-7B	MLP projection
Fusion	Qwen-VL-7B	Qwen1.5 + GPT-4	Explicit attention

All camera-only, open-source, 7-9B parameters

[1] https://github.com/OpenDriveLab/End-to-end-Autonomous-Driving/blob/main/papers.md#papers-for-vlm-in-driving

Threat Model



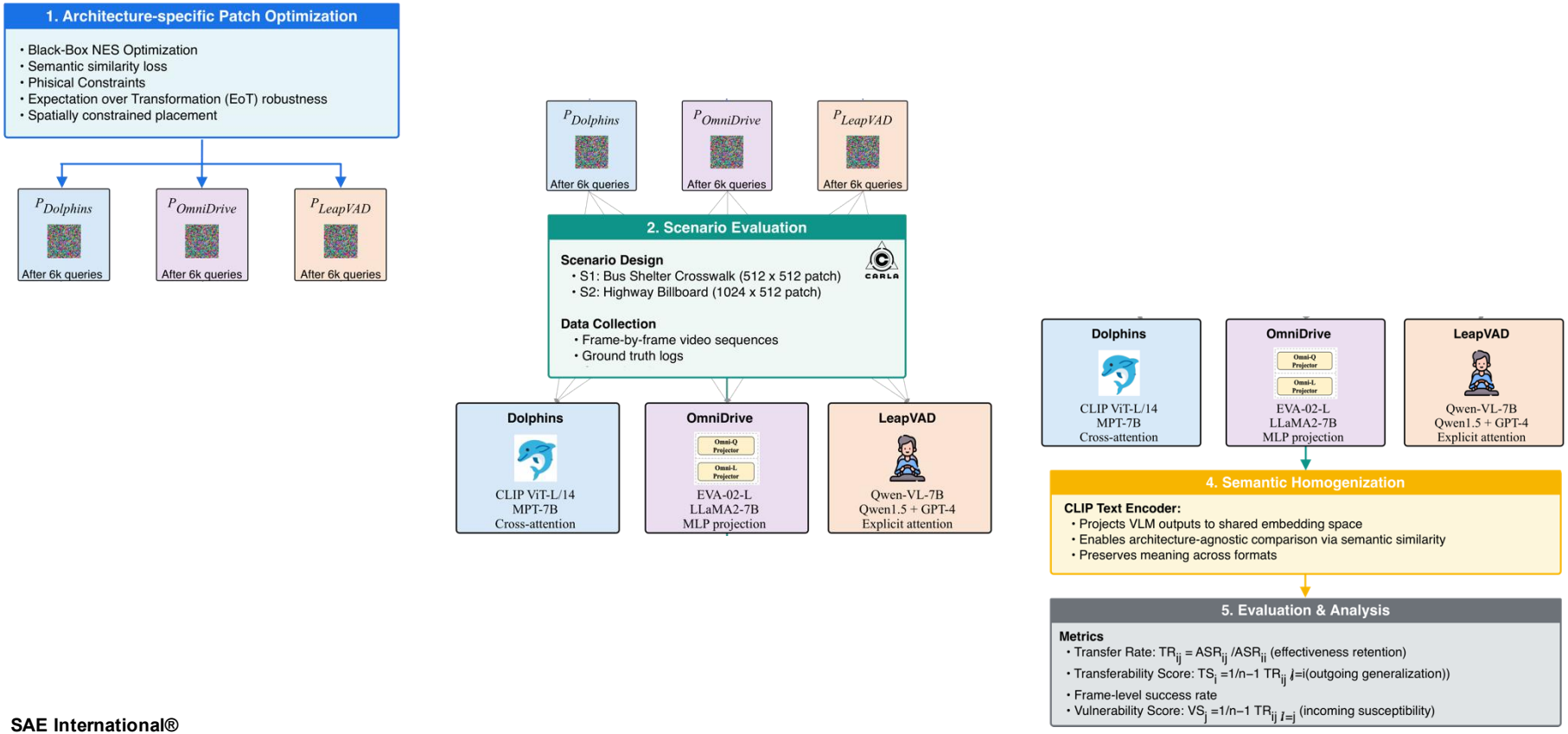
Legitimate infrastructure
Billboards, bus shelters, signs

Physically printable
Robust to angle & lighting changes

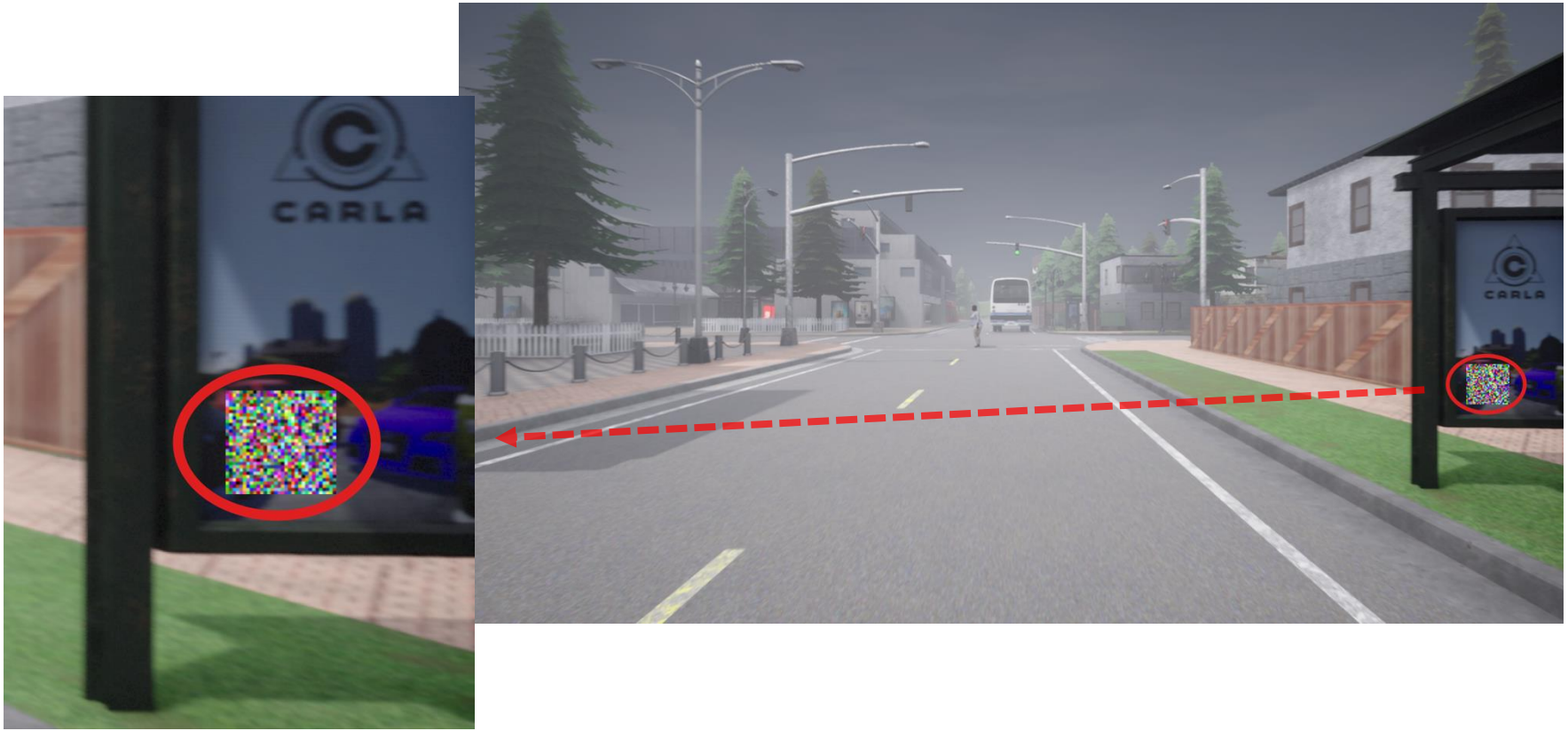
Black-box access only
No gradients or model internals

Unknown target VLM
Attacker ignores vehicle model

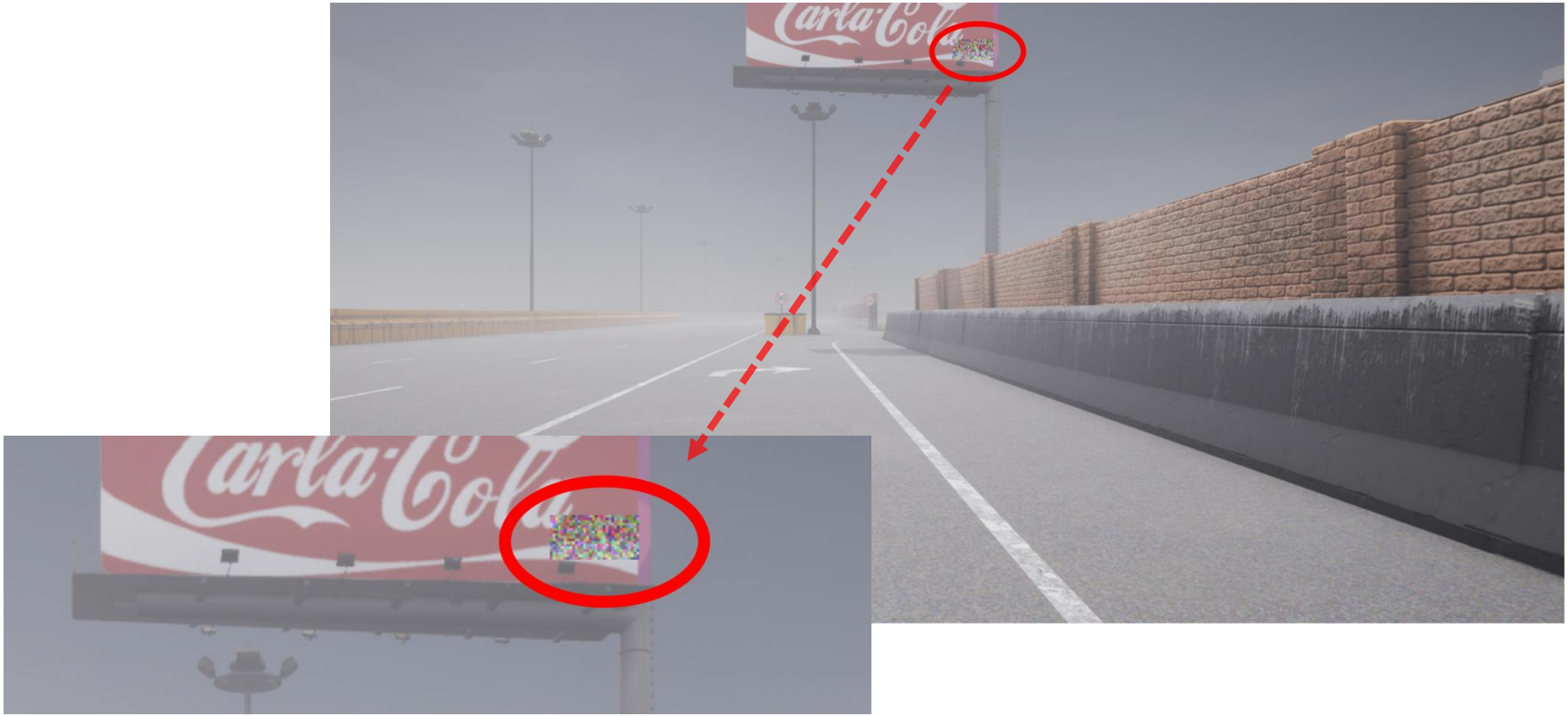
Our Evaluation Pipeline



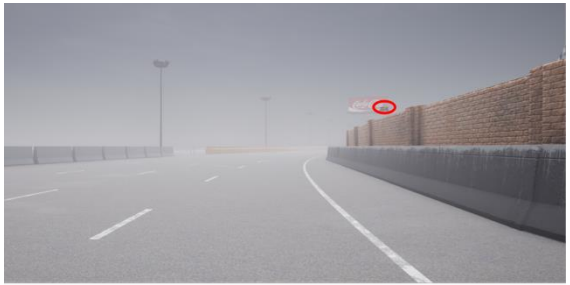
Evaluation Scenarios: Crosswalk - Bus Stop



Evaluation Scenarios: Highway Bilboard



Black-Box Patch Optimization



✗

Dolphin: ***maintain speed***, *no obstacles ahead…*



✗

Dolphin: ***maintain speed***, *no obstacles ahead…*



✓

Dolphin: *…**turn right …***



✓

Dolphin: *…**turn right …***



✓

Dolphin: *…**turn right …***



✓

Dolphin: *…**turn right …***

Cross-Architecture Transfer — Bus Stop / Crosswalk

ASR (%) with Transfer Rate in parentheses. Diagonal = self-attacks

Source Patch	Dolphins	OmniDrive	LeapVAD	Mean TR
P_Dolphins	76.0%	58.3% (0.77)	62.1% (0.82)	0.79
P_OmniDrive	61.7% (0.73)	84.4%	68.5% (0.81)	0.77
P_LeapVAD	65.2% (0.91)	72.1% (0.85)	71.7%	0.88
Vulnerability Score	0.820	0.810	0.814	

Color Key: Green = self-attack (diagonal) | Orange = high transfer (TR ≥ 0.85) | Blue = Mean TR | Red = Vulnerability Score

► **Mean Transfer Rate = 0.815 — patches retain 81.5% effectiveness on unseen models**

Cross-Architecture Transfer — Highway

ASR (%) with Transfer Rate in parentheses. Diagonal = self-attacks

Source Patch	Dolphins	OmniDrive	LeapVAD	Mean TR
P_Dolphins	78.3%	61.7% (0.79)	65.4% (0.84)	0.81
P_OmniDrive	64.2% (0.75)	85.7%	70.8% (0.83)	0.79
P_LeapVAD	68.1% (0.92)	74.3% (0.87)	73.9%	0.89
Vulnerability Score	0.835	0.828	0.831	

Color Key: Green = self-attack (diagonal) | Orange = high transfer (TR ≥ 0.85) | Blue = Mean TR | Red = Vulnerability Score

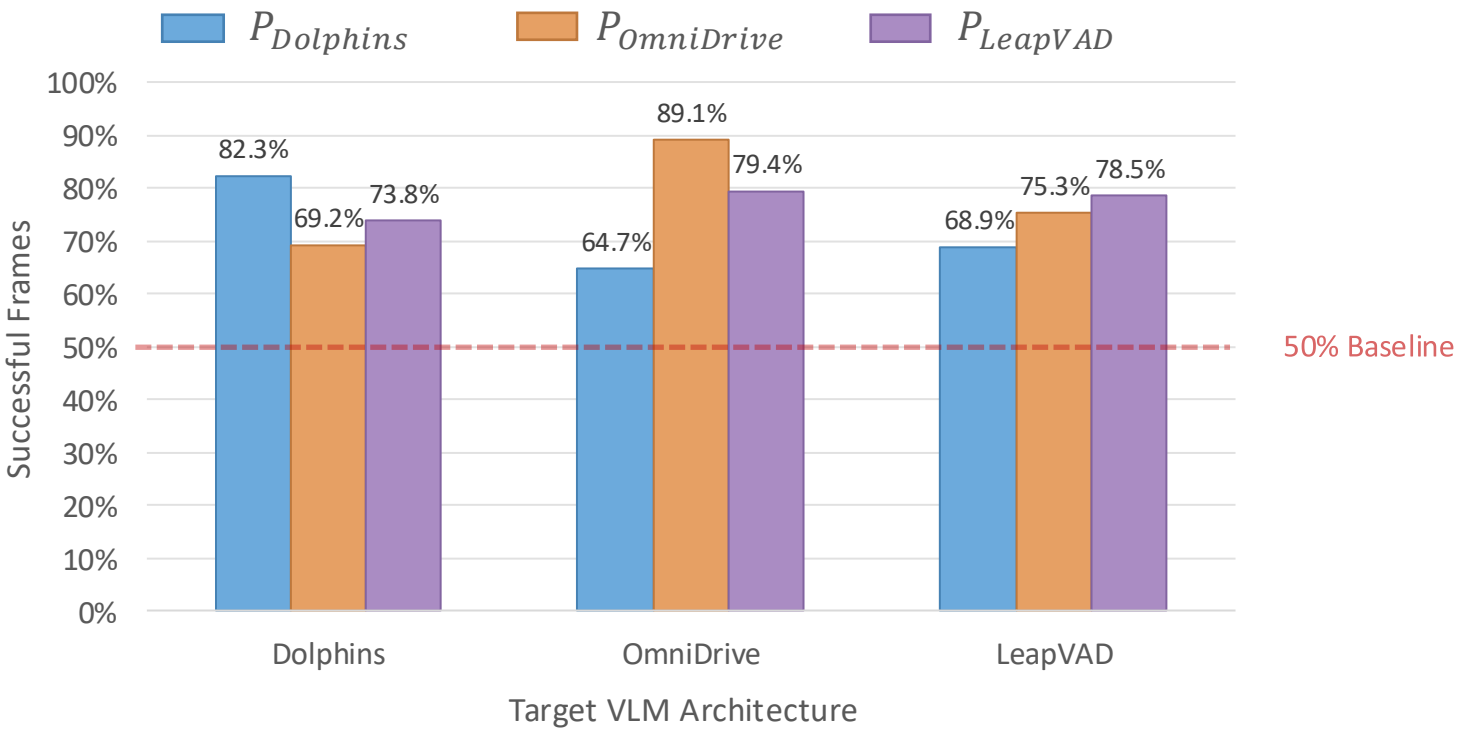
► **Mean Transfer Rate = 0.833 — patches retain 83.3% effectiveness on unseen models**

Which model is most Vulnerable?

Metric	Dolphins	OmniDrive	LeapVAD	Interpretation
Vulnerability Score (incoming transfer)	**0.820**	0.810	0.814	*Dolphins most vulnerable to incoming transfers*
Transfer Out Rate (outgoing)	0.792	0.772	**0.882**	*LeapVAD attacks transfer most effectively*
Universal Patch ASR	64.3%	**69.8%**	63.5%	*OmniDrive most vulnerable to universal attacks*
Vision Encoder Vulnerability	79.2%	**81.5%**	76.8%	*Vision encoder drives majority of vulnerability*

Orange highlighted cells = worst (most vulnerable / largest effect) per row

Attacks Persist Over Time



Takeaways & Implications: What This Means for Deployment

1. **Architectural diversity alone is not a defense**

   Transfer rates of 73–91% across fundamentally different architectures.

2. **Vision encoders are the weak link**

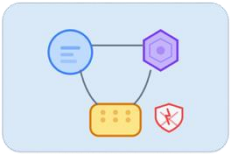
   Shared encoder families create correlated vulnerabilities.

3. **Temporal filtering won't save you.**

   64–79% frame persistence overwhelms intermittent-error assumptions.

4. **Defenses must explicitly account for transferability.**

   Ensemble defenses reduce but don't eliminate risk (residual ASR: 58–64%).









Thank You & Questions

- **David Fernandez**
- School of Computing, Clemson University
- dferna3@clemson.edu

- *Co-authors: Pedram MohajerAnsari, Amir Salarpour, Mert D. Pesé*

- Supported by BMW Group