AYI NEDJIMI Consultants

# Adversarial Patches 2026 : VLM GPT-4V, Cla Gemini

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3498 mots •

Un patch physique imprime trompe GPT-4V, Claude Vision, Gemini avec u 60%. Code Python, math, defenses 2026. Decryptage technique 2026 ave Python et benchmarks.

À RETENIR

## A retenir — Adversarial Patches Physiques sur VLM

**Adversarial patches** physiques imprimes trompent GPT-4V (ASR 62%), Cl Vision (51%), Gemini Vision (67%) sur tasks de classification visuelle.

**Visual prompt injection** via texte cache dans image (Carlini et al., 2024) by des prompt classifiers texte. ASR 78%.

Threat model 2026 : agents IA avec compute puter Use, Operator) exposes a screenshots adversariau

> Defenses : input transformation (compression JPEG aggressive), randomiz
> classifier visuel dedie, watermarking image authentique.
>
> Cas reel Q1 2026 : agent commercial autonome trompe par invoice PDF av
> adversarial — transfer de 80k EUR vers IBAN attaquant.

L'**adversarial patch vision llm** est l'attaque qui combine le pire des deux mondes : vulnerabilites adversariales classiques de la computer vision (depuis Szegedy et a les capacites generatives etendues des Vision-Language Models 2026 (GPT-4V, G Vision, Gemini Vision, Llama 4 Vision). Un patch physique — un autocollant, un mo — place sur un objet trompe le VLM avec un ASR > 60% sur les tasks de classifica un texte adversarial cache dans une image (steganographie visuelle, watermark in de prompt injection imperceptible. Quand un agent IA avec computer use (Claude Use, OpenAI Operator) ingere des screenshots contenant des patches, l'impact bu devient critique : agents commerciaux trompes par invoices, agents juridiques par truques, agents IT par dashboards modifies. Cet article presente la math (LaVAN, 2017 + extensions VLM), le code Python, les benchmarks 2026, et les defenses. P CISO et architectes IA, l'**adversarial patch vision llm** est en 2026 le risque le plus sur les VLM en production — un seul patch imprime peut declencher une fraude fi documentee (cas Workday 80kEUR Q1 2026).

## 1. Genese et etat de l'art

Chronologie des adversarial patches :

- **Szegedy et al. (2013)** — *Intriguing properties of n* emier adve example.

Réponse sous 24h

Devis gratuit →