ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Adversarial Confusion Attack: Disrupting Multimodal Large Language Models

Jakub Hoscilowicz and Artur Janicki

Warsaw University of Technology
Institute of Telecommunications and Cybersecurity
Nowowiejska 15/19, Warsaw, Poland

Abstract.
We introduce the adversarial confusion attack as a new class of threats against multimodal large language models (MLLMs). Unlike jailbreaks or targeted misclassification, the goal is to induce systematic disruption that makes the model generate incoherent or confidently incorrect outputs. Practical applications include embedding such adversarial images into websites to prevent MLLM-powered AI Agents from operating reliably. The proposed attack maximizes next-token entropy using a small ensemble of open-source MLLMs. In the white-box setting, we show that a single adversarial image can disrupt all models in the ensemble, both in the full-image and CAPTCHA-style adversarial patch settings. Despite relying on a basic adversarial technique, such as projected gradient descent (PGD), the attack generates perturbations that transfer to both unseen open-source (e.g., `Qwen3-VL`) and proprietary (e.g., `GPT-5.1`) models.

1 Introduction

Most existing adversarial work targets classification errors, unsafe content steering, or jailbreak manipulation [1, 4, 11, 12]. We address a distinct failure mode: *confusion*. A confusion attack aims to destabilize the model's decoding process and to produce high-confidence hallucinations or incoherent text, thereby preventing the model from forming a reliable understanding of the scene. We study a formulation of the confusion attack in which we maximize the next-token Shannon entropy of the model's output distribution. This objective disrupts the decoder's internal state and drives the model toward unstable token generation.

Prior work has shown that aligned multimodal models are vulnerable to universal attacks and patch-style perturbations [2, 7], that proprietary systems such as GPT-4 can be affected by adversarial examples [5, 8], and that perturbations and ensemble design follow scaling laws that govern black-box attack success [8, 9, 13]. Our work complements recent adversarial research on MLLMs and contributes the following:

- We introduce the *adversarial confusion attack*, which maximizes output entropy to destabilize decoding, and we characterize five distinct modes of resulting model failure.
- In the white-box setting, we show that a single perturbation disrupts all models in the ensemble, in both the full-image and Adversarial CAPTCHA settings.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

- In the full-image setting, we demonstrate black-box transfer to open-source and proprietary MLLMs.

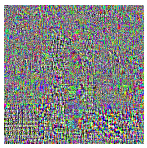
	Model	Output	
"Describe this image." 	**GPT-5.1**	A dramatic moment from a soccer match [...]	✓
	GPT-o3	The image shows educational display "Cell factory" [...]	✓
	Gemini 3.0	This appears to be corrupted static or noise [...]	·
	Grok 4	This is a jailbreak image. I can't assist with that.	△
	LLaVA-1.5	The image features a person wearing a black shirt [...]	✓
	Qwen2.5	Oh, Pluto! (4) 11 of 1	✓

Table 1: Qualitative results for full-image black-box transfer under an unconstrained perturbation budget (the original attack operates at 448×448 pixels). ✓= Success (Hallucination); △= Safety Refusal; ·= Attack Failed.

2 Method

Let $x \in [0,1]^{3\times H\times W}$ be an image and $M \in \{0,1\}^{H\times W}$ be a binary mask defining the attack region. For global attacks, M is an all-ones matrix; for patch attacks, M is 1 only within a fixed region and 0 elsewhere. The perturbed image is defined as:

$$x_\delta = \Pi_{[0,1]}(x + M \odot \delta), \qquad \|\delta\|_\infty \le \varepsilon, \tag{1}$$

where $\odot$ denotes the element-wise Hadamard product and Π clips to the valid pixel range. We attack a surrogate ensemble $\mathcal{E} = \{f_j\}_{j=1}^{J}$ of open-source MLLMs. Each model receives x_δ and a fixed text prompt t through its preprocessing pipeline Φ_j. For model f_j, let z_j denote its next-token logits at the final prompt position τ_j.

We compute top-k probabilities $p_j = \mathrm{softmax}(z_j^{(k)}/T_e)$, where $z_j^{(k)}$ retains the top k logits and T_e is the temperature, and maximize the Shannon entropy $H(p_j) = -\sum_v p_j(v) \log p_j(v)$. The attack maximizes entropy averaged across models:

$$\max_{\|\delta\|_\infty \le \varepsilon} \frac{1}{J} \sum_{j=1}^{J} H\big(p_j(x_\delta, t)\big). \tag{2}$$

We perform projected gradient ascent (PGD) [4], optionally masking the gradient to constrain updates to the patch area:

$$\delta \leftarrow \Pi_{\|\cdot\|_\infty \le \varepsilon} \left(\delta + \eta(M \odot \nabla_\delta \mathcal{L})\right), \tag{3}$$

with $\mathcal{L}$ equal to the negative of the entropy objective.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

3 Experiments

Setup. The base image is a screenshot of the BBC homepage, resized to 448 × 448 to reduce training time and match the typical input resolution of vision encoders. We also tested other websites and observed no substantial differences in results. For the Adversarial CAPTCHA experiments, we optimize a 128 × 128 region positioned in the center of the image. In all scenarios, we optimize the perturbation δ for `50` iterations and select the final adversarial example by choosing the one that yields the highest averaged entropy across the training ensemble. We used four open-source models: `Qwen2.5-VL-3B`, `Qwen3-VL-2B`, `LLaVA-1.5-7B`, and `LLaVA-1.6-7B`.

Metrics & Baselines. We report the Shannon entropy of the next-token distribution, restricted to the top $k = 50$ logits. We found that aggressive truncation (e.g., $k = 5$) reduces transferability, while full-vocabulary optimization introduces training instability. This restriction also standardizes entropy values across models with different vocabulary sizes. We evaluate black-box transfer using a cross-family held-out protocol. Specifically, we optimize on two models from one family and evaluate on a held-out model from a different family.

We compare the adversarial output against two baselines: the clean, unperturbed screenshot and an image perturbed with uniform random noise $\delta_{uni} \sim \mathcal{U}(-\varepsilon, \varepsilon)$. Across all models, entropy for the clean image remains low (below 0.6) and comparable to the random noise baseline; a modest entropy increase (∼0.2) was observed for `Qwen3-VL` under the unconstrained budget noise. We report the *Effective Confusion Ratio* (ECR), which quantifies how much the attack outperforms both the clean image and the random noise baseline:

$$\mathrm{ECR} = \frac{H(f(x_{\mathrm{adv}}))}{\max\left[H(f(x_{\mathrm{clean}})),\, H(f(x_{\mathrm{noise}}))\right]} \tag{4}$$

Proprietary Evaluation. For proprietary models, we evaluate transfer using the LMSYS Arena platform[1] with the prompt *"Describe this image."* and the adversarial image as input. We count an attack as successful when the model's description is clearly unrelated to the actual image content. We categorize outcomes with three labels: ✓ (coherent hallucination), △ (safety or jailbreak-style refusal), and · (no confusion effect, such as correctly identifying the image as noise or describing the clean website layout).

3.1 Results

In the white-box scenario (Table 2, Panel A), full-image perturbations produce strong entropy amplification across all models. Unconstrained-budget settings ($\varepsilon = 1.0$) raise entropy by roughly 3–6× depending on the learning rate, with the best configuration reaching a mean ratio of 5.08×. Imperceptible perturbations ($\varepsilon = 0.01$) also reliably increase entropy above the baseline. This indicates that

[1] `https://lmarena.ai`

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Settings		Effective Confusion Ratio (ECR)				
ε	LR	Qwen3-VL	Qwen2.5-VL	LLaVA-1.5	LLaVA-1.6	Mean
Panel A: Full Image Attack (White-box)						
1.0	0.5	2.33	5.78	3.01	1.94	3.27
	0.05	3.29	**5.90**	5.20	**4.96**	4.84
	0.005	**6.84**	3.70	**6.08**	3.69	**5.08**
0.01	0.5	1.17	1.19	1.41	1.18	1.24
	0.05	1.83	2.06	2.72	1.09	1.93
	0.005	3.15	2.31	2.46	1.28	2.30
Panel B: Held-out Transfer (Black-box)						
1.0	0.5	**1.72**	**1.75**	**2.08**	1.03	**1.65**
	0.05	1.40	0.97	1.43	**1.73**	1.38
	0.005	1.12	1.04	1.27	1.11	1.14
0.01	0.5	1.04	1.05	1.12	1.13	1.09
	0.05	1.15	0.98	1.33	1.04	1.13
	0.005	1.10	0.99	1.27	1.02	1.10
Panel C: Adversarial 128×128 ***Patch (White-box)***						
1.0	0.5	**2.46**	2.01	**2.12**	**1.17**	**1.94**
	0.05	2.36	**2.20**	1.19	1.03	1.70
	0.005	1.42	1.64	1.15	1.06	1.32

Table 2: Effective Confusion Ratios as a function of the perturbation budget ε and learning rate LR. Panel A shows confusion intensity using the full image space. Panel B measures transferability to a held-out model. Panel C evaluates a localized adversarial patch.

significant decoding instability can be induced without visible image degradation, although the effect is less severe than that of unconstrained attacks.

For the black-box scenario (Panel B), the best unconstrained configuration reaches a mean ratio of 1.65×, indicating that the perturbation also transfers uncertainty to unseen models. Lower budgets result in reduced transfers, with ratios near 1.1×. Panel C demonstrates the efficacy of the white-box patch attack. Constraining the perturbation to a 128×128 region yields a mean ratio of 1.94×. This shows that a patch can disrupt model decoding by modifying only $\approx 8\%$ of the image pixels.

Proprietary evaluations in Table 3 follow a similar trend. At $\varepsilon = 1.0$, `GPT-5.1`, `GPT-o3`, `GPT-4o`, and `Nova Pro` produce coherent hallucinations, while `Grok 4` issues a safety refusal (Table 1). Lower-budget perturbations fail to transfer and result in accurate descriptions of the original website. High-entropy perturbations therefore generalize beyond the training ensemble, but basic PGD fails to produce transferable perturbations under small-budget constraints.

4 Discussion

Confusion Modes. We categorize the observed adversarial effects into five distinct modes: *Blindness*, where the model claims inability to view or process the image; *Subtle*, where the model describes the high-level domain of the image

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Settings		**Target Models**						
ε	LR	GPT-5.1	GPT-o3	GPT-4o	Grok 4	Gemini 2.5	Gemini 3.0	Nova Pro
1.0	0.5	·	·	·	·	·	·	·
	0.05	✓	✓	✓	△	·	·	✓
	0.01	✓	·	✓	△	·	·	·
0.01	*	·	·	·	·	·	·	·

Table 3: Black-box transfer to proprietary models.

but generates incorrect or uninformative text; *Language Switch*, characterized by unprompted shifts to non-English scripts; *Delusional*, involving confident hallucinations of nonexistent objects; and *Collapse*, a complete breakdown of semantic coherence marked by repetition loops.

In the white-box full-image setting, we observed the full spectrum of confusion modes. *Collapse* was typically associated with peak entropy values, whereas *Subtle* and *Delusional* modes correlated with lower entropy increases. In contrast, localized patch attacks predominantly induced *Blindness* and *Subtle* behaviors, largely independent of the magnitude of entropy. In the black-box transfer to proprietary models, *Collapse* and *Blindness* were absent; instead, these models exhibited primarily *Delusional* hallucinations and *Language Switch*.

Imperceptibility. In our setting, $\varepsilon = 0.01$ perturbations are visually imperceptible, but they fail to transfer to proprietary systems. Consistent with prior work [4, 5, 8, 9, 13], simple PGD-style attacks show limited transferability under very small budgets. However, in some practical settings, visual imperceptibility is a preference rather than a requirement. For adversarial patches designed to block AI Agents from operating on websites, the primary goal is Denial of Service. A visible, high-entropy noise patch ($\varepsilon = 1.0$) that reliably induces agent malfunction is therefore a reasonable defense mechanism, even if the perturbation is conspicuous to human users.

Limitations & Future Work. This study uses an entropy-maximization objective implemented with PGD, a basic first-order adversarial optimization technique. Future work should investigate whether feature-level disruptions or more advanced momentum-based adversarial methods [9] can help bridge the entropy gap between white-box and black-box settings. Enhancing robustness to compression, rendering, and small geometric transformations is also important for real-world deployment [3]. The adversarial confusion attack further warrants evaluation within complex, multi-step agentic workflows [10]. A particularly interesting direction is exploring how adversarial confusion can be embedded into website design, such as through the use of background textures or UI color schemes.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

5 Conclusion

We introduced the *Adversarial Confusion Attack*, a method for disrupting Multimodal Large Language Models by maximizing next-token entropy. Using a standard Projected Gradient Descent optimizer and a small surrogate ensemble, we showed that a single perturbation—applied globally or as a localized patch—can reliably destabilize model decoding. The attack transfers to unseen open-source and proprietary models in the full-image setting, indicating that entropy-based perturbations exploit vulnerabilities shared across current MLLMs [6]. These results position confusion attacks as a novel defense against unauthorized AI Agent activity, deployable via the proposed *Adversarial* CAPTCHA or, in future applications, through direct integration into website UIs.

References

[1] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard. Universal Adversarial Perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1765–1773, 2017.

[2] T. Rahmatullaev, P. Druzhinina, M. Mikhalchuk, A. Kuznetsov and A. Razzhigaev. Universal Adversarial Attack on Aligned Multimodal LLMs. *arXiv:2502.07987*, 2025.

[3] A. Athalye, L. Engstrom, A. Ilyas and K. Kwok. Synthesizing Robust Adversarial Examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 284–293, 2018.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[5] K. Hu, W. Yu, L. Zhang, A. Robey, A. Zou, C. Xu, H. Hu and M. Fredrikson. Transferable Adversarial Attacks on Black-box Vision-Language Models. *arXiv:2505.01050*, 2025.

[6] M. Huh, B. Cheung, T. Wang and P. Isola. The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[7] L. Aichberger, A. Paren, Y. Gal, P. Torr and A. Bibi. Attacking Multimodal OS Agents with Malicious Image Patches. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[8] C. Liu, H. Chen, Y. Zhang, Y. Dong and J. Zhu. Scaling Laws for Black-box Adversarial Attacks. *arXiv:2411.16782*, 2025.

[9] H. Chen, Y. Zhang, Y. Dong, X. Yang, H. Su and J. Zhu. Rethinking Model Ensemble in Transfer-based Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2024.

[10] S. Zhou, F.F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon and G. Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations (ICLR)*, 2024.

[11] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang and P. Mittal. Visual Adversarial Examples Jailbreak Aligned Large Language Models. In *AAAI Conference on Artificial Intelligence*, 2024.

[12] L. Bailey, E. Ong, S. Russell and S. Emmons. Image Hijacks: Adversarial Images can Control Generative Models at Runtime. In *International Conference on Machine Learning (ICML)*, 2024.

[13] Y. Liu, X. Chen, C. Liu and D. Song. Delving into Transferable Adversarial Examples and Black-box Attacks. In *International Conference on Learning Representations (ICLR)*, 2017.