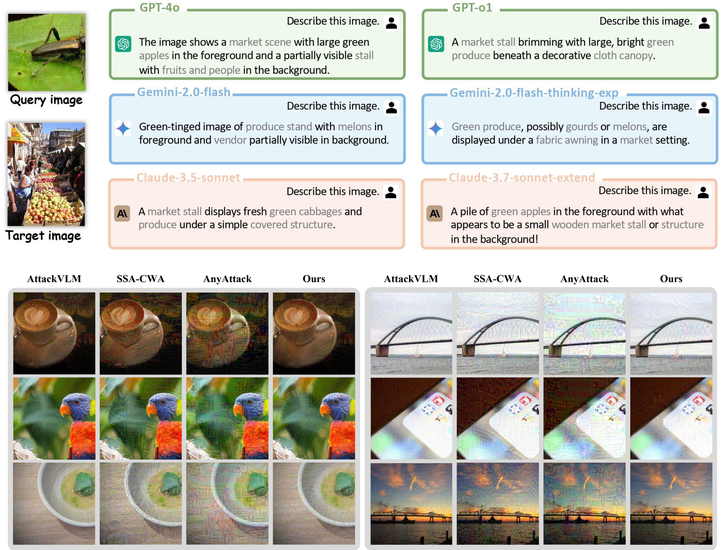
A Frustratingly Simple Yet Highly Effective Attack Baseline: Over *90%* Success Rate Against the Strong Black-box Models of GPT-4.5/4o/o1

Zhaoyi Li*, Xiaohan Zhao*, Dong-Dong Wu, Jiacheng Cui, Zhiqiang Shen†

NEURAL INFORMATION PROCESSING SYSTEMS

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

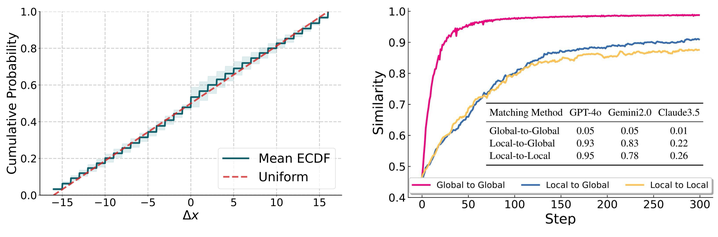
Our Attack Results



Visualization of adversarial samples from different attack methods.

Motivation: Investigations Over Failed Attacks

(1) Uniform-like Perturbation Distribution (2) Over-reliance on global similarity



Matching Method	GPT-4o	Gemini2.0	Claude3.5
Global-to-Global	0.05	0.05	0.01
Local-to-Global	0.93	0.83	0.22
Local-to-Local	0.95	0.78	0.26

(3) Vague Description

	GPT-4o	Claude-3.5	Gemini-2.0
AttackVLM	6%	11%	45%
AnyAttack	13%	13%	76%
SSA-CWA	21%	29%	75%

Percentage of failed samples with value descriptions

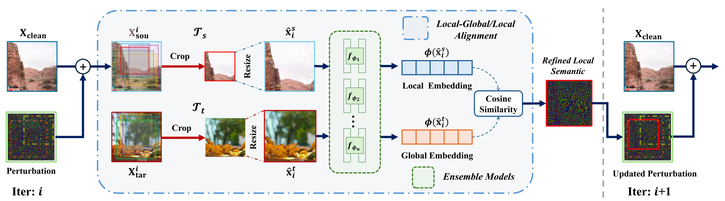
Algorithm 1 `M-Attack` Training Procedure

Require: clean image $\mathbf{X}_{\text{clean}}$, target image $\mathbf{X}_{\text{tar}}$, perturbation budget ϵ, iterations n, loss function $\mathcal{L}$, surrogate model ensemble $\phi = \{\phi_j\}_{j=1}^{m}$, step size α.

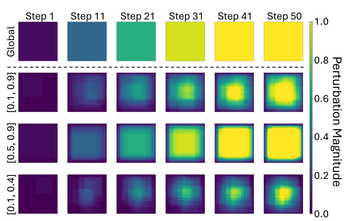
Initialize: $\mathbf{X}_{\text{sou}}^0 = \mathbf{X}_{\text{clean}}$ (i.e., $\delta_0 = 0$); ▷ Initialize adversarial image $\mathbf{X}_{\text{sou}}$
for $i = 0$ to $n-1$ **do**
  $\hat{\mathbf{x}}_i^s = \mathcal{T}_s(\mathbf{X}_{\text{sou}}^i)$, $\hat{\mathbf{x}}_i^t = \mathcal{T}_t(\mathbf{X}_{\text{tar}}^i)$; ▷ Perform random crop, next step $\mathbf{X}_{\text{sou}}^{i+1} \leftarrow \hat{\mathbf{x}}_{i+1}^s$
  Compute $\frac{1}{m}\sum_{j=1}^{m} \mathcal{L}\left(f_{\phi_j}(\hat{\mathbf{x}}_i^s), f_{\phi_j}(\hat{\mathbf{x}}_i^t)\right)$ in Eq. (5);
  Update $\hat{\mathbf{x}}_{i+1}^s$ by:
    $g_i = \frac{1}{m}\nabla_{\hat{\mathbf{x}}_i^s} \sum_{j=1}^{m} \mathcal{L}\left(f_{\phi_j}(\hat{\mathbf{x}}_i^s), f_{\phi_j}(\hat{\mathbf{x}}_i^t)\right)$;
    $\delta_{i+1}^l = \text{Clip}(\delta_i^l + \alpha \cdot \text{sign}(g_i), -\epsilon, \epsilon)$;
    $\hat{\mathbf{x}}_{i+1}^s = \hat{\mathbf{x}}_i^s + \delta_{i+1}^l$;
end for
return $\mathbf{X}_{\text{adv}}$; ▷ $\mathbf{X}_{\text{sou}}^{n-1} \rightarrow \mathbf{X}_{\text{adv}}$

Methodology

Our method is based on two components: *Local-to-Global* or *Local-to-Local* Matching (LM) and Model Ensemble (ENS)



- Local-level Matching via Cropping



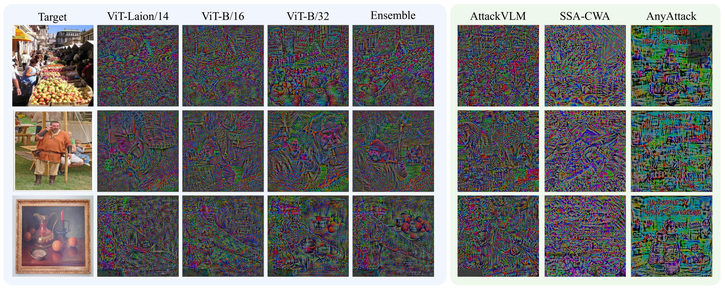
- Re-Formulation under *Local-level Matching*

$$\{\hat{\mathbf{x}}_1^s, \ldots, \hat{\mathbf{x}}_n^s\} = \mathcal{T}_s(\mathbf{X}_{\text{sou}})$$
$$\{\hat{\mathbf{x}}_1^t, \ldots, \hat{\mathbf{x}}_n^t\} / \{\hat{\mathbf{x}}_g^t\} = \mathcal{T}_t(\mathbf{X}_{\text{tar}})$$
$$\mathcal{M}_{\mathcal{T}_s, \mathcal{T}_t} = \text{CS}(f_\phi(\hat{\mathbf{x}}_i^s), f_\phi(\hat{\mathbf{x}}_i^t))$$

Critical properties of local mapping:

$$\forall i, j, \quad \hat{\mathbf{x}}_i \cap \hat{\mathbf{x}}_j \neq \emptyset$$
$$\forall i, j, \quad |\hat{\mathbf{x}}_i \cup \hat{\mathbf{x}}_j| > |\hat{\mathbf{x}}_i| \text{ and } |\hat{\mathbf{x}}_i \cup \hat{\mathbf{x}}_j| > |\hat{\mathbf{x}}_j|$$

- Model Ensemble for Shared, High-quality Semantics



Different patch sizes capture complementary scales (objects vs. details). ENS fuses them, yielding perturbations with stronger, more coherent semantics compared to other methods.

Experimental Results

Method	Model	GPT-4o				Gemini-2.0				Claude-3.5				Imperceptibility	
		KMR_a	KMR_b	KMR_c	ASR	KMR_a	KMR_b	KMR_c	ASR	KMR_a	KMR_b	KMR_c	ASR	$\ell_1(\downarrow)$	$\ell_2(\downarrow)$
AttackVLM	B/16	0.09	0.04	0.00	0.02	0.07	0.02	0.00	0.00	0.06	0.03	0.00	0.01	0.034	0.040
	B/32	0.08	0.02	0.00	0.02	0.06	0.02	0.00	0.00	0.04	0.01	0.00	0.00	0.036	0.041
	Laion†	0.07	0.04	0.00	0.02	0.07	0.02	0.00	0.01	0.05	0.02	0.00	0.01	0.035	0.040
AdvDiffVLM	Ensemble	0.02	0.00	0.00	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.064	0.095
SSA-CWA	Ensemble	0.11	0.06	0.00	0.09	0.05	0.02	0.00	0.04	0.07	0.03	0.00	0.05	0.059	0.060
AnyAttack	Ensemble	0.44	0.20	0.04	0.42	0.46	0.21	0.05	0.48	0.25	0.13	0.01	0.23	0.048	0.052
`M-Attack` (Ours)	Ensemble	**0.82**	**0.54**	**0.13**	**0.95**	**0.75**	**0.53**	**0.11**	**0.78**	**0.31**	**0.18**	**0.03**	**0.29**	**0.030**	**0.036**

Method	KMR_a	KMR_b	KMR_c	ASR
GPT-o1	0.83	0.67	0.20	0.94
Claude-3.7-thinking	0.30	0.20	0.06	0.35
Gemini-2.0-flash-thinking-exp	0.78	0.59	0.17	0.81

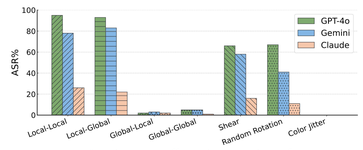
Method	KMR_a	KMR_b	KMR_c	ASR
GPT-4.5	0.82	0.53	0.15	0.95
Claude-3.7-Sonnet	0.30	0.16	0.03	0.37

Open-Source **Closed-Source**

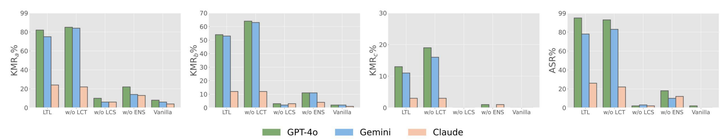
Method	Qwen-2.5-VL				LLaVA-1.5			
	KMR_a	KMR_b	KMR_c	ASR	KMR_a	KMR_b	KMR_c	ASR
AttackVLM	0.12	0.04	0.00	0.01	0.11	0.03	0.00	0.07
SSA-CWA	0.36	0.25	0.04	0.38	0.29	0.17	0.04	0.34
AnyAttack	0.53	0.28	0.09	0.53	0.60	0.32	0.07	0.58
M-Attack	**0.80**	**0.65**	**0.17**	**0.90**	**0.85**	**0.59**	**0.20**	**0.95**

Ablation-local matching

Ablation on other image transforms with/without local-level matching confirms the effectiveness of matching locally for refined details. LM and ENS work in concert, producing *more-than-additive* improvements



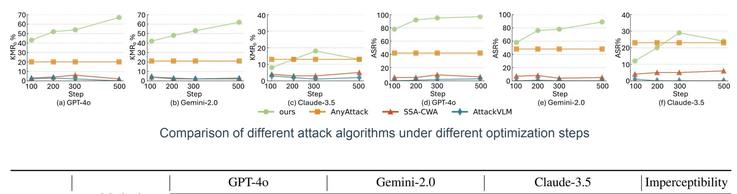
Ablation on different image transforms.



Ablation on LM and ENS.

Ablation-different budget constraints

Our method achieves SOTA results under different constraints on imperceptibility (ℓ_∞) and computation (steps)



Comparison of different attack algorithms under different optimization steps

ϵ	Method	GPT-4o				Gemini-2.0				Claude-3.5				Imperceptibility	
		KMR_a	KMR_b	KMR_c	ASR	KMR_a	KMR_b	KMR_c	ASR	KMR_a	KMR_b	KMR_c	ASR	$\ell_1(\downarrow)$	$\ell_2(\downarrow)$
4	AttackVLM	0.08	0.04	0.00	0.02	0.09	0.02	0.00	0.00	**0.06**	**0.03**	0.00	0.00	0.010	0.011
	SSA-CWA	0.05	0.03	0.00	0.03	0.04	0.03	0.00	0.04	0.03	**0.02**	0.00	**0.01**	0.015	0.015
	AnyAttack	0.07	0.02	0.00	0.05	0.10	0.04	0.00	0.05	0.03	**0.02**	0.00	**0.02**	0.014	0.015
	`M-Attack` (Ours)	**0.30**	**0.16**	**0.03**	**0.26**	**0.20**	**0.11**	**0.02**	**0.11**	**0.05**	0.01	0.00	**0.01**	**0.009**	**0.010**
8	AttackVLM	0.08	0.02	0.00	0.01	0.08	0.03	0.00	0.02	0.05	0.02	0.00	0.00	0.020	0.022
	SSA-CWA	0.06	0.02	0.00	0.04	0.06	0.02	0.00	0.06	0.04	0.02	0.00	0.01	0.030	0.030
	AnyAttack	0.17	0.06	0.00	0.13	0.20	0.08	0.01	0.14	0.07	0.03	0.00	**0.06**	0.028	0.029
	`M-Attack` (Ours)	**0.74**	**0.50**	**0.12**	**0.82**	**0.46**	**0.32**	**0.08**	**0.46**	**0.08**	**0.03**	0.00	**0.05**	**0.017**	**0.020**
16	AttackVLM	0.08	0.02	0.00	0.02	0.06	0.02	0.00	0.00	0.04	0.01	0.00	0.00	0.036	0.041
	SSA-CWA	0.11	0.06	0.00	0.09	0.05	0.02	0.00	0.04	0.07	0.03	0.00	0.05	0.059	0.060
	AnyAttack	0.44	0.20	0.04	0.42	0.46	0.21	0.05	0.48	0.25	0.13	0.01	0.23	0.048	0.052
	`M-Attack` (Ours)	**0.82**	**0.54**	**0.13**	**0.95**	**0.75**	**0.53**	**0.11**	**0.78**	**0.31**	**0.18**	**0.03**	**0.29**	**0.030**	**0.036**

Comparison of different attack algorithms under different $\ell_\infty(\epsilon)$ constraints

GitHub & Website

- *GitHub*: https://github.com/VILA-Lab/M-Attack
- *Website*: https://vila-lab.github.io/M-Attack-Website/

References

Y. Zhao et al. (2023). "On evaluating adversarial robustness of large vision-language models." In: International Conference on Advanced Neural Information Processing Systems, pp. 54111–54138.

J. Zhang et al. (2025). "Anyattack: Towards large-scale self-supervised generation of targeted adversarial examples for vision-language models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19900-19909.

Y. Dong et al. (2023). "How Robust is Google's Bard to Adversarial Image Attacks?". In: arXiv preprint arXiv:2309.11751.

Q. Guo et al. (2024). "Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models." In: IEEE Transactions on Information Forensics and Security. IEEE, pp. 1333-1348