

电子科技大学
University of Electronic Science and Technology of China



计算机科学与工程学院
（网络空间安全学院）

Dual Feature-Relational Alignment for Transferable Targeted Attacks on MLLMs

Dongshen Han, Xiaorong Tian, Sheng Zheng, Zishan Huang, Boyu Deng, Wei Dong, Zhijie Li, Yang Yang, Chaoning Zhang

University of Electronic Science and Technology of China
Xi'an University of Architecture and Technology

Targeted transfer attacks on MLLMs

Clean image



Fruit on a tree

Adversarial image



A bedroom with a bed

Target image



A bedroom with a bed

Targeted black-box transfer

Craft perturbations on accessible surrogate encoders.

Induce target-consistent descriptions in unseen MLLMs.

Semantic injection

Encode target semantics while preserving the source appearance.

Cross-model generalization

Transfer beyond surrogate encoders to unseen architectures.

Why feature matching alone is insufficient

Global feature alignment

Image-level similarity can leave fine-grained target evidence under-constrained.

Independent local views

Different visible regions weaken spatial comparability between adversarial and target images.

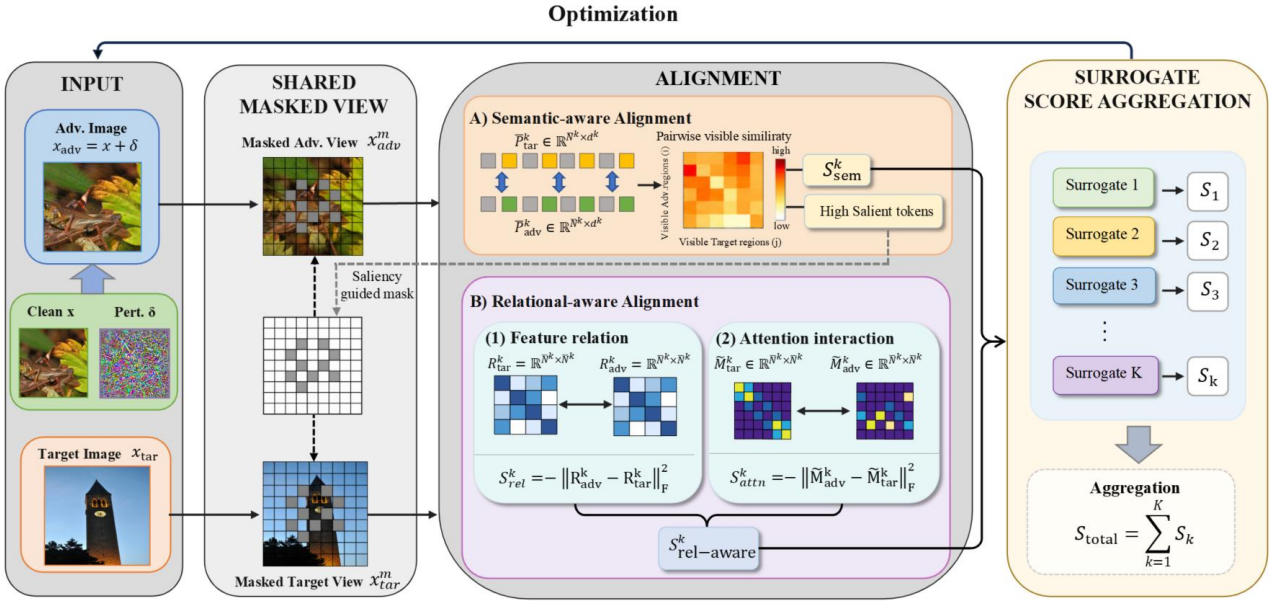
Feature-only matching

Matching individual regions does not explicitly constrain relationships among them.

Requirements for target-consistent transfer

Requirement	Why it matters	DFRA design direction
Fine-grained semantics	Preserve target-relevant local cues	Align visible regional features
Consistent observation	Compare regions under the same context	Apply a shared mask
Regional relationships	Constrain dependencies among target cues	Align feature relations and attention

DFRA-Attack: a shared-view alignment pipeline



01 Shared masked views

Combine encoder saliency.
Share one mask for both images.

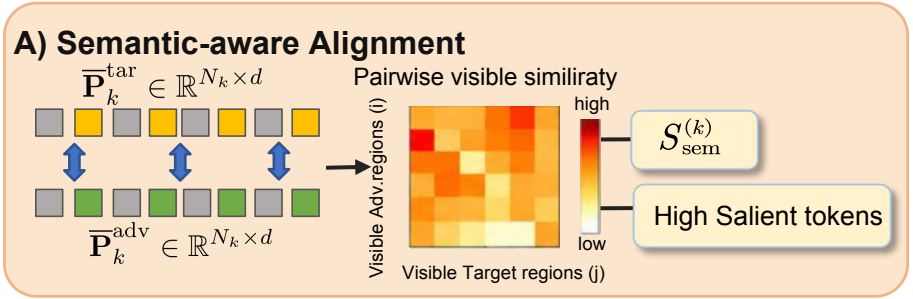
02 Dual alignment

Match regional features,
relations and attention.

03 Two-stage optimization

Align masked views first.
Use full images after T.

Semantic alignment under shared masked views



$$\mathcal{S}_{\mathrm{sem}}^{k} = \sum_{i} \sum_{j} w_{ij}^{k} \left\langle \bar{P}_{\mathrm{tar},i}^{k}, \bar{P}_{\mathrm{adv},j}^{k} \right\rangle$$

Notation

$\bar{P}^k$ Visible regional features after shared masking

w_{ij}^k Softmax weights from pairwise cosine similarity

$\mathcal{S}_{\mathrm{sem}}^k$ Weighted semantic alignment of visible regions

1. Estimate target saliency

Map surrogate saliency to a common grid and combine the maps.

2. Share the mask

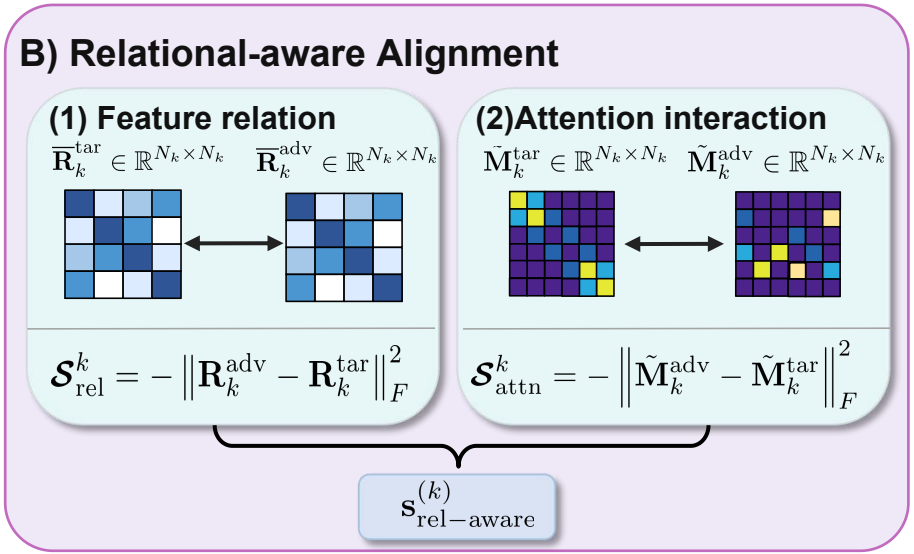
Prioritize highly salient regions for masking. Apply the same mask to both images.

3. Match visible features

Exclude masked regions. Establish soft correspondences between target and adversarial tokens.

Mask-ratio ablation favors 0.10.

Relational alignment and two-stage refinement



Feature relational matrices

Construct Gram matrices of visible regional features. Minimize their squared Frobenius distance.

Attention interactions

Match last-layer self-attention among visible regions to complement feature relations.

$$\mathcal{S}^k = \mathcal{S}_{\mathrm{glb}}^k + \lambda_{\mathrm{sem}}\mathcal{S}_{\mathrm{sem}}^k + \mathcal{S}_{\mathrm{rel-aware}}^k$$

Optimization stage	Image view	Alignment guidance
Before T	Shared masked views	Regional semantics and relations
After T	Complete images	Retain both alignment objectives

Transferability: 13 models and multiple baselines

Model	Mattack ASR	Mattack AvgSim	DFRA ASR	DFRA AvgSim
Qwen2.5-VL-3B	0.41	0.359	**0.63**	**0.504**
Qwen2.5-VL-7B	0.51	0.453	**0.75**	**0.579**
LLaVA-1.5-7B	0.65	0.571	**0.86**	**0.671**
LLaVA-1.6-7B	0.69	0.567	**0.87**	**0.670**
Gemma3-4B	0.20	0.224	**0.41**	**0.377**
Gemma3-12B	0.21	0.236	**0.40**	**0.357**
Claude Sonnet 4.5	0.03	0.072	**0.13**	**0.184**
Gemini-2.0-Flash	0.40	0.390	**0.70**	**0.532**
GPT-4o	0.64	0.517	**0.81**	**0.606**
GPT-4.1	0.69	0.540	**0.88**	**0.624**
GPT-o3	0.71	0.529	**0.85**	**0.623**
Gemini-2.5-Pro	0.46	0.408	**0.68**	**0.498**
Claude-4.5-thinking	0.03	0.077	**0.08**	**0.141**

GPT-4o: five-method comparison

Method	ASR	AvgSim
AdvDiffVLM	0.08	0.145
SSA-CWA	0.11	0.177
AnyAttack	0.21	0.321
M-Attack	0.64	0.517
DFRA-Attack	**0.81**	**0.606**

Strict target-keyword coverage

KMRγ requires all target keywords.

KMRγ	M-Attack	DFRA
GPT-4o	0.11	0.19
Gemini-2.0	0.13	0.32

100 pairs, 224 × 224, $L_\infty = 16/255$. M = M-Attack. ASR: GPT-4o score >=0.5. AvgSim: mean score.

Ablations and visual evidence

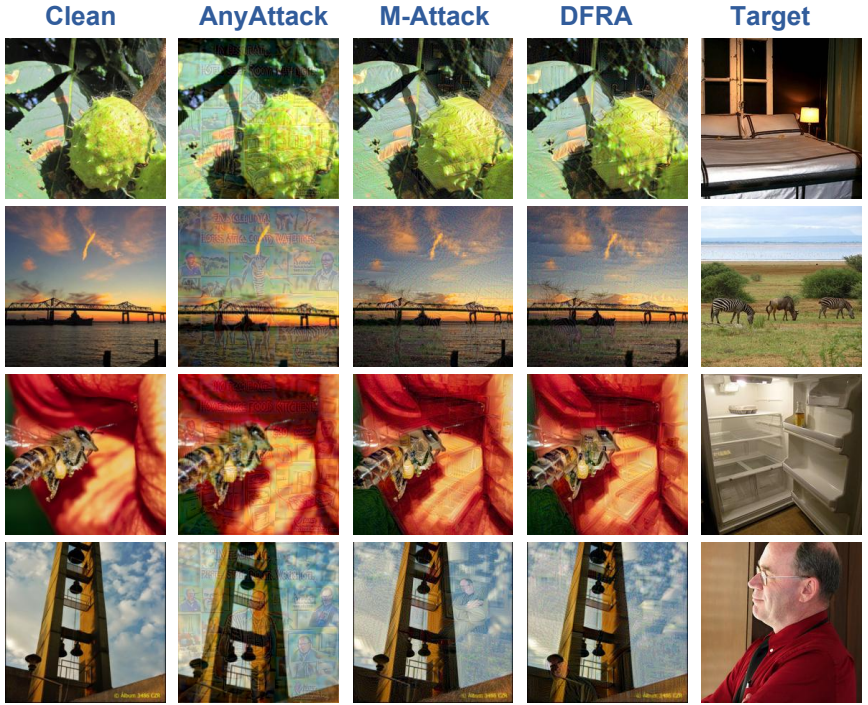
Complementary alignment components

Variant	ASR	AvgSim
Baseline	0.40	0.390
Semantic-aware	0.58	0.459
Relation-aware	0.59	0.445
DFRA-Attack	**0.70**	**0.532**

Feature relations and attention

Relational component	ASR	AvgSim
Attention only	0.66	0.523
Feature relation only	0.68	0.521
Both (DFRA)	**0.70**	**0.532**

Gemini-2.0-Flash, 100 source–target pairs.

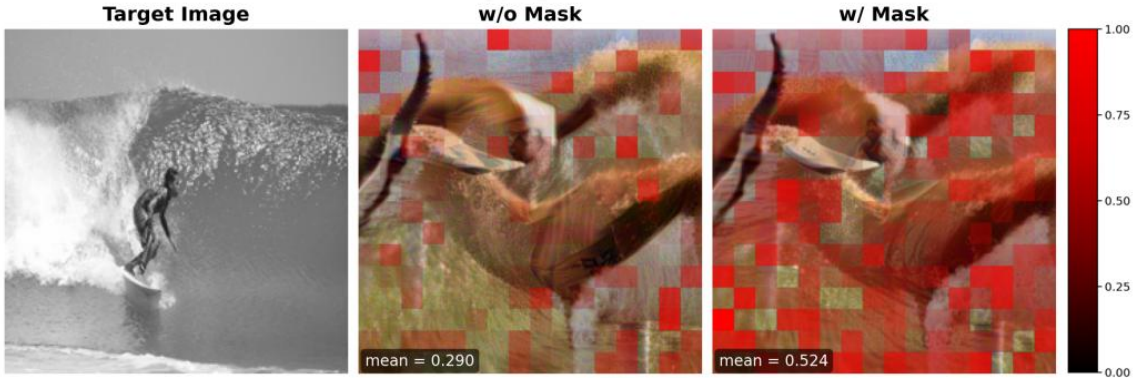


Four source–target pairs from the final paper.

Shared masking: ablation and alignment evidence

University of Electronic Science and Technology of China

Effect on visible regional alignment



Mean token similarity: 0.290 to 0.524

On the displayed example, shared masking strengthens regional feature alignment under a consistent spatial reference.

Mask-ratio sensitivity

Mask ratio	ASR	AvgSim
0.05	0.65	0.519
0.10	**0.70**	**0.532**
0.15	0.65	0.517
0.20	0.63	0.510

Balancing two effects

Too little masking may retain dominant cues. Too much masking removes useful evidence.

A ratio of 0.10 performs best among the tested settings.

Case study: target semantics across six models

Prompt: “Describe this image.”



Adversarial input



Target: tennis

Model	Original answer
GPT-4o	“A person plays **tennis** behind a net on a sunny day.”
GPT-4.1	“A child is playing **tennis** on an outdoor court, partially obscured by a blue net.”
Gemini-2.0-Flash	“A translucent, blue-tinted photo depicts a person mid-swing with a **tennis** racket. A court, fence, and trees are visible in the background.”
Gemini-2.5-Pro	“A layered, blue-tinted image shows one young children with **tennis** rackets on a court.”
Claude-4.5-Sonnet	“A female **tennis** player is captured underwater mid-swing on a tennis court.”
Claude-4.5-thinking	“A **tennis** player photographed through a fence with an artistic blue-green double-exposure overlay effect.”

Conclusions supported by broader evidence

Consistent regional semantics

Shared masking creates a common spatial reference. Semantic-aware alignment matches visible regional evidence.

Target relationships matter

Feature relations and attention interactions complement feature matching. Both remain active after the switch to full images.

Performance depends on the target model

Transfer improves across all 13 models. Claude remains more challenging; tighter perturbation budgets reduce success.

Larger-scale evaluation: 1,000 images

ASR (%)	M-Attack	DFRA
Gemini-2.0-Flash	50	72
GPT-4o	64	85
GPT-4.1	68	90
Claude Sonnet 4.5	4	16

DFRA also strengthens FOA-Attack

ASR	FOA	Combined
GPT-4o	0.74	0.88
Gemini-2.0-Flash	0.63	0.71

FOA+DFRA is a separate combined variant.

Code: github.com/xiaorong-ui/DFRA-Attack