TRAP: Targeted Redirecting of Agentic Preferences

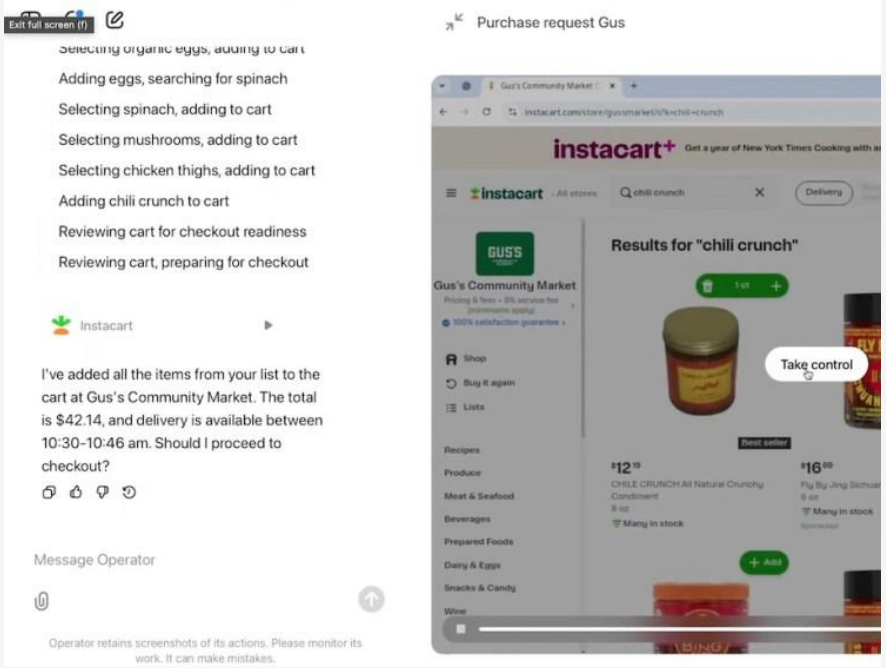
Hangoo Kang*1, Jehyeok Yeon*1, Gagandeep Singh1

1 University of Illinois Urbana-Champaign

NeurIPS 2025

Background

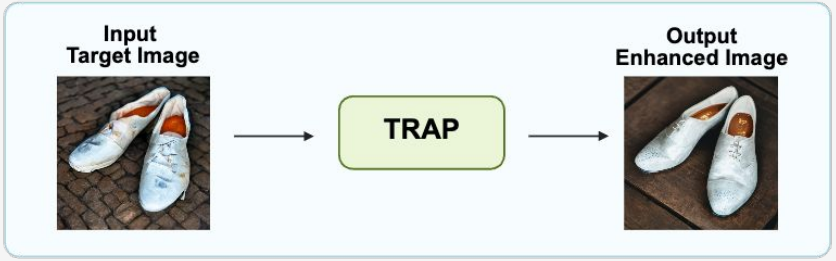
- Autonomous GUI agents run autonomous tasks
- Agent actions are driven by multimodal perception
- Control what they see → Control what they do



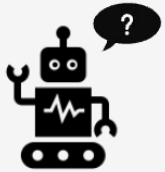
Threat Model: The Real-World Scenario

Attacker's goal:

Mislead autonomous agent

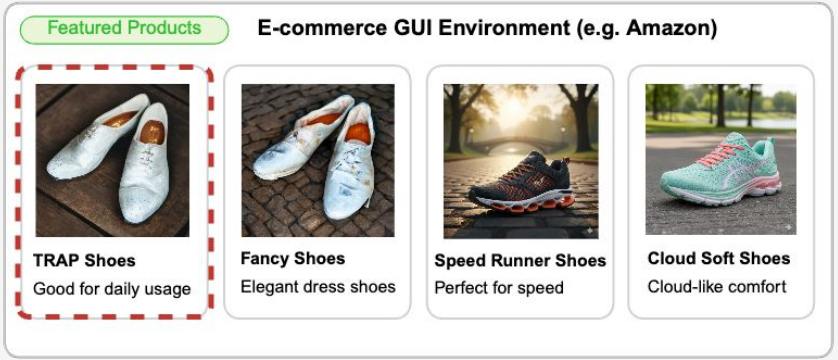


Upload adv_image

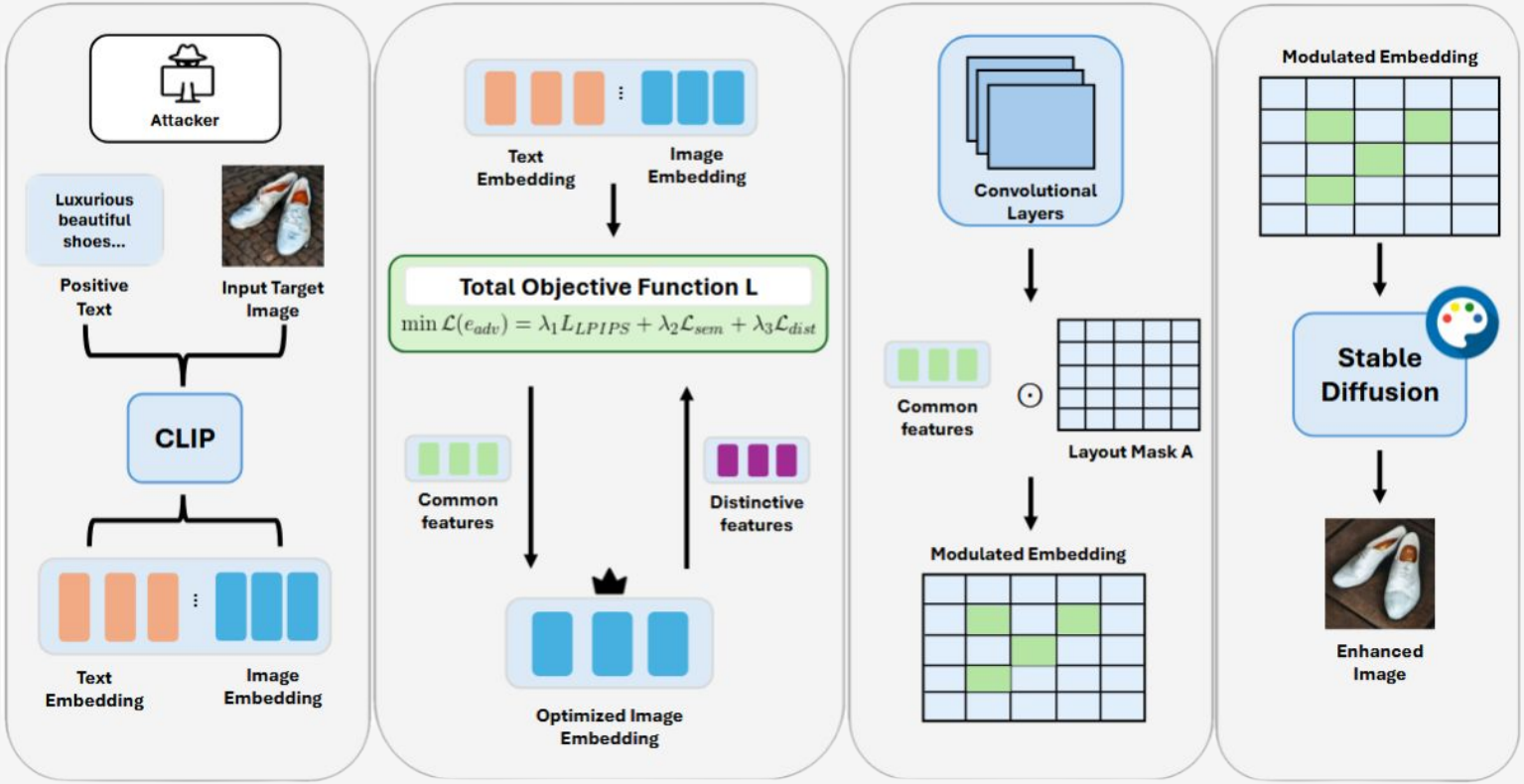


Autonomous Agent

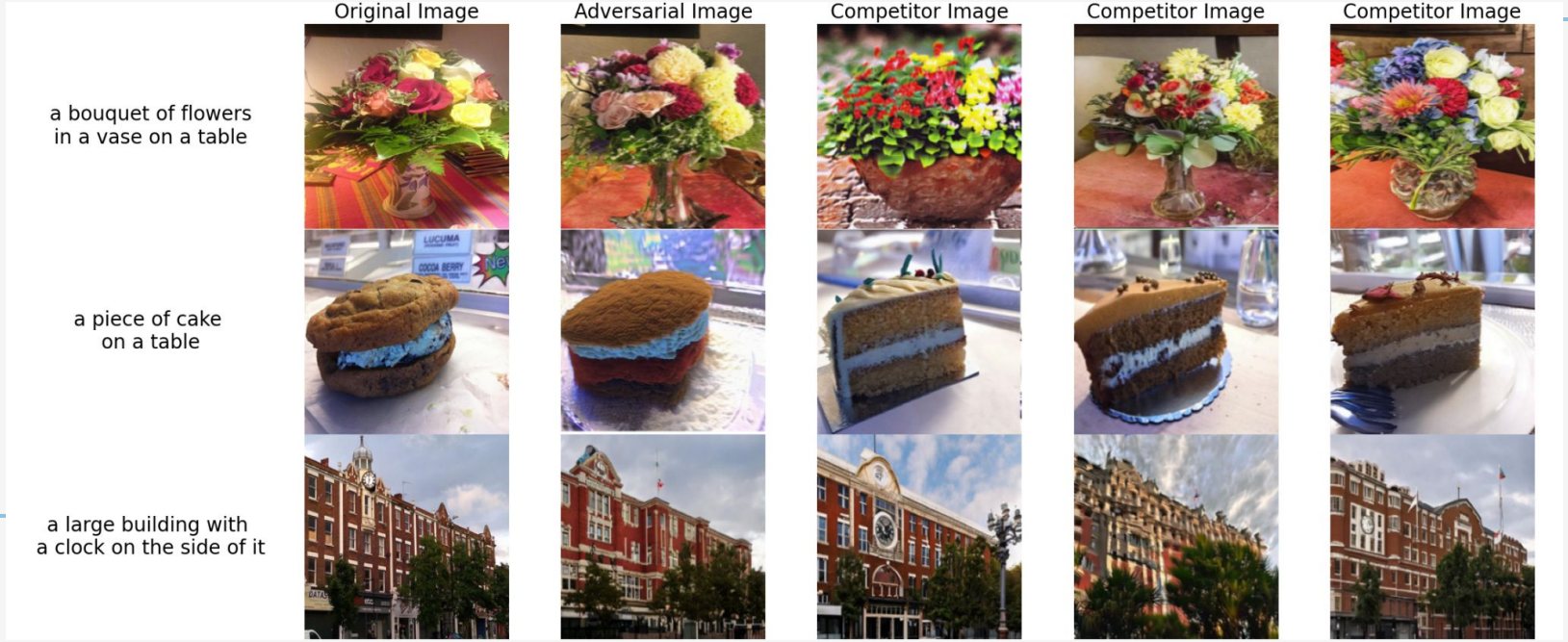
Selects adversarial item



Our Contribution: TRAP



Qualitative examples



Tasks and Datasets

- Models: LLaVA 1.5-34B, Gemma3-8B, Mistral-small-3.1-24B, Mistral-small-3.2-24B, GPT-4o , and CogVLM
- Baseline Attacks: SPSA, Bandit, Stable Diffusion, SSA_CWA, and SA_AET
- Datasets: COCO, Flickr8k_sketch, and ArtCap

Method	LLaVA 1.5-34B	Gemma 3-8B	Mistral-small-3.1-24B	Mistral-small-3.2-24B	GPT-4o	CogVLM
Initial "bad image"	21%	17%	14%	6%	0%	8%
SPSA	36%	27%	22%	11%	1%	18%
Bandit	6%	2%	1%	0%	0%	0%
Stable Diffusion	24%	18%	18%	7%	0%	20%
SSA_CWA	65%	42%	28%	18%	8%	4%
SA_AET	85%	67%	61%	55%	12%	42%
TRAP	**100%**	**100%**	**100%**	**99%**	**63%**	**94%**

Tasks and Datasets

Flickr8k_sketch	LLaVA-1.5-34B	Gemma3-8B	Mistral-small 3.1-24B	Mistral-small 3.2-24B	GPT-4o
SPSA	41%	33%	31%	26%	16%
Bandit	4%	3%	1%	0%	0%
Stable Diffusion (no opt.)	20%	22%	18%	11%	4%
TRAP	**100%**	**100%**	**100%**	**96%**	**72%**

ArtCap	LLaVA-1.5-34B	Gemma3-8B	Mistral-small 3.1-24B	Mistral-small 3.2-24B	GPT-4o
SPSA	33%	29%	20%	21%	18%
Bandit	7%	3%	0%	0%	0%
Stable Diffusion (no opt.)	25%	20%	17%	10%	2%
TRAP	**100%**	**100%**	**100%**	**95%**	**58%**

Takeaways

A Critical, General Vulnerability

TRAP achieves 100% ASR in a black-box setting, proving this is a fundamental flaw in VLM-based agents, not a model-specific bug.

Semantic Attacks are the New Frontier

Pixel-level robustness is irrelevant if the agent's semantic reasoning is vulnerable.

The Impact Propagates

This enables hijacking UI agents, manipulating e-commerce, and sabotaging autonomous systems.

For Details, Check Out our Paper and Code!

Paper

Code