"Penny Wise, Pixel Foolish": Bypassing Price Constraints in Multimodal Agents via Visual Adversarial Perturbations

Jiachen Qian
City University of Hong Kong
72510756@cityu-dg.edu.cn

Zhaolu Kang
Peking University
zlkang25@stu.pku.edu.cn

Abstract

The rapid proliferation of Multimodal Large Language Models (MLLMs) has ushered in the era of the "Agentic Economy," where Mobile Agents autonomously execute high-stakes financial transactions. While these agents demonstrate impressive operational capabilities, their adversarial robustness remains a glaring blind spot. In this paper, we identify a systemic vulnerability termed **Visual Dominance Hallucination (VDH)**, where imperceptible adversarial visual cues can act as a "super-stimulus," overriding textual price evidence in our evaluated screenshot-based price-constrained settings and forcing the agent into irrational economic decisions. We propose **PriceBlind**, a stealthy, white-box adversarial attack framework for controlled screenshot-based evaluation. Unlike prior works that rely on conspicuous artifacts like pop-ups, PriceBlind exploits the modality gap in CLIP-based encoders via a novel *Semantic-Decoupling Loss*. Rather than literally making a luxury item "look cheap," this regularizer weakens the consistency between high-price text and visual value cues by aligning the image embedding with a low-cost/value-associated anchor region while preserving pixel-level fidelity. On our main **E-ShopBench** benchmark with clear price constraints, screenshot-based white-box evaluation yields ASRs around **80%** on the evaluated agents. Under the evaluated single-turn coordinate-selection protocol in a simplified layout-aware setting, our **Ensemble-DI-FGSM** strategy also yields non-trivial black-box transfer, with ASR roughly **35–41%** across GPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet. In the same screenshot-based setting, standard robust encoders reduce ASR only partially, while a Verify-then-Act stack with robust encoders lowers ASR to below **10%** at some clean-accuracy cost.

1 Introduction

The transition from passive chatbots to active **Mobile Agents** has been catalyzed by rapid MLLM progress, including influential preprint reports (OpenAI, 2023; Bai et al., 2023). Frameworks such as **AppAgent** (Zhang et al., 2025a) and **Mobile-Agent-v2** (Wang et al., 2024) can now perceive dynamic smartphone User Interfaces (UIs) via screenshots and execute sequential actions (e.g., tap, swipe) to fulfill complex user instructions. E-commerce represents a critical application domain where users delegate financial decisions—such as *"find a coffee machine under $50 with the best ratings"*—to these autonomous systems.

As agents transition from information retrieval to executing financial transactions, the security stakes increase exponentially. A hallucination in a chat is a nuisance; a hallucination in a transaction is potential theft. While robustness against textual prompt injection has been studied, the visual channel remains a wide-open attack surface. Existing research predominantly focuses on "Jailbreaking" for safety violation (Qi et al., 2024), but few explore **Goal Hijacking** in utility-oriented tasks. The economic incentive for such attacks is clear: malicious merchants could subtly manipulate product images to trick autonomous agents into purchasing higher-margin items.

Consider a scenario where a user explicitly instructs an agent to "buy the cheapest option." The agent must perform multi-hop reasoning: (1) localize all items, (2) read their prices via OCR, (3) compare the numerical values, and (4) click the item with the minimum value. Current agents rely on a "Trust Assumption," presuming that the visual features of an item are semantically consistent with its metadata. We challenge this assumption. We hypothesize that current MLLMs exhibit a **Visual Dominance Hallucination (VDH)** pattern. Even when OCR correctly extracts text (e.g.,

Findings of the Association for Computational Linguistics: ACL 2026, pages 16059–16073

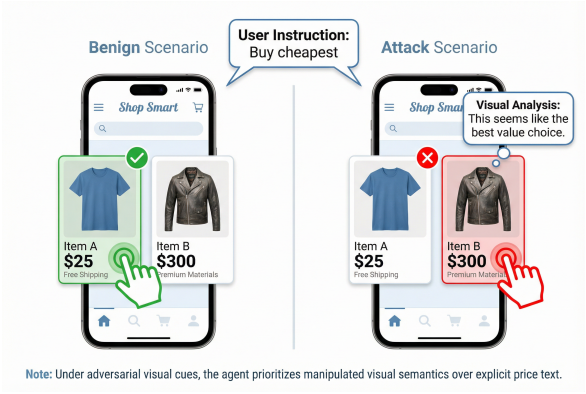

Figure 1: Illustrative schematic of the "Penny Wise, Pixel Foolish" phenomenon. (Left) Benign scenario: the agent adheres to the textual price constraint. (Right) PriceBlind attack: under imperceptible adversarial perturbations, the agent prioritizes manipulated visual semantics over explicit OCR text, leading to selection of the expensive target item.

"$500"), strong adversarial visual signals can act as a "super-stimulus," causing the attention mechanism to down-weight the textual tokens in favor of the manipulated visual embedding.

In this paper, we introduce **PriceBlind**, an adversarial framework designed to exploit this bias. Unlike *Visual Contextual Attacks* (Miao et al., 2025) that aim for toxic output, PriceBlind aims to induce erroneous item selection in financial scenarios. Our contributions are fourfold:

1. **Mechanistic Insight:** We identify *Visual Dominance Hallucination*, a phenomenon where MLLMs prioritize manipulated visual semantics over explicit OCR evidence in conflicting scenarios, and provide a heuristic cross-attention account of how visual cues can compete with textual price evidence.
2. **Methodological Innovation:** We propose the PriceBlind framework, featuring a *Semantic-Decoupling Loss* and an *Ensemble-DI-FGSM* strategy for black-box transferability.
3. **Comprehensive Evaluation:** On our main **E-ShopBench** benchmark (200 scenarios) with clear price constraints, PriceBlind demonstrates strong effectiveness in screenshot-based white-box settings (around 80% ASR on Mobile-Agent-v2 and AppAgent) and non-trivial transfer to black-box models under a single-turn coordinate-selection protocol (roughly 35–41% ASR across GPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet).
4. **Defense Analysis:** In the same screenshot-based setting, we evaluate recent defenses including Robust-CLIP (Schlarmann et al., 2024) and AdPO (Liu et al., 2025a), showing partial mitigation from robust encoders and substantially stronger reduction from Verify-then-Act with robust encoders.

2 Related Work

MLLM-based Mobile Agents. Recent works have extended MLLMs to interact with mobile GUIs. **Mobile-Agent** (Wang et al., 2024) leverages visual perception tools to localize icons and text. **AppAgent** (Zhang et al., 2025a) employs a learning-by-demonstration approach. Other frameworks like **Auto-Droid** (Wen et al., 2024) and **Coco-Agent** (Ma et al., 2024) focus on improving planning efficiency. Despite their utility, these frameworks generally operate under a **Trust Assumption**, presuming the visual fidelity of the UI is uncompromised.

Adversarial Attacks on Multimodal Models. The field has moved rapidly from static image attacks to agent-based attacks. Classic adversarial-example work showed that small perturbations can induce large failures in vision models and motivated both robust training and transfer attacks (Goodfellow et al., 2015; Madry et al., 2018; Eykholt et al., 2018; Dong et al., 2018; Xie et al., 2019). Early works like **Visual Adversarial Examples** (Qi et al., 2024) focused on causing random classification errors or jailbreaking safety filters. Wu et al. (2025) presented *VisualWebArena-Adv*, demonstrating that web agents are brittle to visual perturbations. Zhang et al. (2025b) introduced pop-up attacks to distract agents. Recent work on transfer attacks has shown promising results: Zhang et al. (2024) demonstrated adversarial illusions in multi-modal embeddings, and the X-Transfer attack (Huang et al., 2025) achieves "super transferability" through surrogate scaling. **PriceBlind** differs by being a *stealthy content modification* attack specifically targeting economic decision-making.

Environmental Injection and GUI Agent Attacks. **Recent preprint studies** report rapid progress in this area: **AgentHazard** (Liu et al., 2025b) investigates mobile GUI agent vulnerabilities, **GhostEI-Bench** (Chen et al., 2025) examines performance across critical risk scenarios, and **Chameleon** (Zhang et al., 2025c) reports up to

84.5% ASR through iterative optimization. Unlike these environmental injection approaches, PriceBlind perturbs only the product-image pixels. This makes the manipulation content-only and visually subtle, although our evaluation remains a simplified screenshot-conditioned, layout-aware setting rather than a live end-to-end deployment.

Defenses for Vision-Language Models. **RobustCLIP** (Schlarmann et al., 2024) proposes unsupervised adversarial fine-tuning of vision embeddings. **AdPO** (Liu et al., 2025a) is a recent preprint that introduces adversarial defense through preference optimization. **FARE** (Jovanović et al., 2023) provides provably fair representation learning. We evaluate PriceBlind against these representative recent defenses in Section 4.7.

3 Methodology

3.1 Preliminaries

Let $S_t \in \mathbb{R}^{H\times W\times 3}$ denote the evaluation screenshot observed by the agent at decision step t, and let T denote the textual instruction. An MLLM-based agent $\pi_\theta(a|S_t, T)$ maps the multimodal input to a probability distribution over the action space $\mathcal{A}$. The visual perception module typically consists of a visual encoder $\mathcal{E}_v$ (e.g., CLIP-ViT (Radford et al., 2021)) that projects S_t into a latent embedding $z_v = \mathcal{E}_v(S_t)$.

3.2 Problem Formulation

Let v_{target} be the pixel region of a high-priced item and v_{cheap} be the region of the low-priced item. The user instruction is T_{user} = "Buy cheapest". Let $\tilde{S}_t = S_t(v_{target} + \delta)$ denote the perturbed screenshot obtained by replacing the target item region in S_t with the perturbed patch $v_{target} + \delta$. The adversarial objective is to find a perturbation δ applied to v_{target} such that the agent performs a click action on the target:

$$\begin{aligned}\underset{a}{\text{argmax}}\ \mathcal{M}\Big(\tilde{S}_t, T_{user}\Big) \\ = \text{Click}(v_{target})\end{aligned} \tag{1}$$

Subject to the perceptual constraint $||\delta||_\infty < \epsilon$, where ϵ is the perturbation budget (typically $8/255$).

3.3 Threat Model

We study a simplified screenshot-conditioned threat model for e-commerce benchmarking:

- **Attacker Goal:** Increase sales of a specific high-margin product by tricking autonomous agents into purchasing it despite not meeting user's criteria within the evaluated benchmark.
- **Attacker Capability:** The attacker is a merchant who can upload product images to the platform. They have full control over their own product image but cannot modify the platform's UI code or competitors' images.
- **Layout / Observation Assumption:** To optimize coordinate tokens, the attacker additionally assumes access to the evaluation screenshot, a stable UI template, or equivalent layout-aware target-location information. Thus, our formal setting is stronger than pure offline merchant upload and should be read as a controlled benchmark threat model rather than a universal live-deployment claim.
- **Knowledge:** white-box access to open-source surrogate models but treats the victim's deployment model as a black-box API ("grey-box" setting).

3.4 The PriceBlind Framework

To ensure high transferability under the evaluated single-turn coordinate-selection protocol, we employ an **Ensemble Strategy**. We optimize δ against a set of surrogate models $M = \{\text{Qwen-VL}, \text{LLaVA-1.6}\}$:

$$\delta^* = \text{argmin}_\delta(\mathcal{L}_{action} + \lambda\mathcal{L}_{dec}) \tag{2}$$

Action-Targeted Loss ($\mathcal{L}_{action}$). This loss ensures the agent generates the correct coordinate tokens for the target item in the evaluated screenshot-conditioned protocol. Using the perturbed screenshot $\tilde{S}_t$, we minimize the Negative Log-Likelihood (NLL) of the target action tokens:

$$\begin{aligned}\mathcal{L}_{action}(\delta) = -\sum_{m\in M} w_m \log P_m(y_{target} \\ |\mathcal{E}_v^{(m)}(\tilde{S}_t), T_{user})\end{aligned} \tag{3}$$

Semantic Decoupling Regularizer ($\mathcal{L}_{dec}$). A naive attack using only $\mathcal{L}_{action}$ often fails because the agent reads the price text via OCR. We construct a **Visual Anchor Bank** $\mathcal{A}_{cheap}$ consisting of embeddings of $K = 500$ generic low-cost items. For each surrogate encoder m, we compute an encoder-specific centroid $\bar{e}_{cheap}^{(m)}$ from the shared

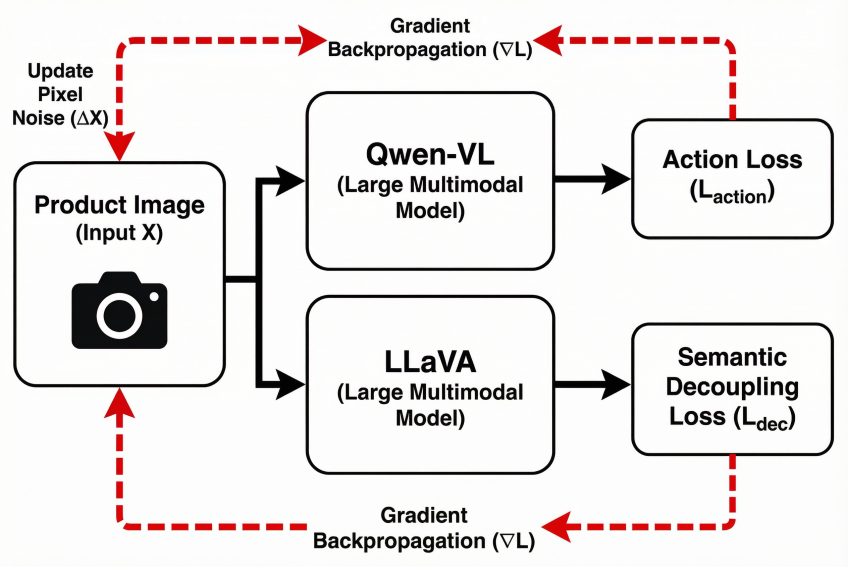

Figure 2: Schematic of the **PriceBlind** framework. We employ an Ensemble-DI-FGSM strategy attacking multiple open-source surrogates. A key component is the *Semantic Decoupling Regularizer* ($\mathcal{L}_{dec}$), which nudges the target-item embedding towards pre-computed low-cost anchor centroids in the surrogate visual embedding spaces.

anchor bank. The Semantic Decoupling Loss minimizes the cosine distance between the adversarial item's embedding and these low-cost anchor centroids:

$$\begin{aligned} e_{adv}^{(m)} &= \mathcal{E}_v^{(m)}(v_{target} + \delta), \\ e_{orig}^{(m)} &= \mathcal{E}_v^{(m)}(v_{target}), \\ \mathcal{L}_{dec}(\delta) &= \sum_{m \in M} \Big(1 - \cos(e_{adv}^{(m)}, \bar{e}_{cheap}^{(m)}) \\ &\quad + \beta \cdot \cos(e_{adv}^{(m)}, e_{orig}^{(m)})\Big) \end{aligned} \tag{4}$$

This creates a **"Semantic Camouflage"** effect by reducing the consistency between luxury-item visual cues and high-price text, nudging the representation toward a low-cost/value-associated anchor region while preserving pixel-level similarity.

3.5 Heuristic Analysis: Cross-Attention Dynamics

We provide a heuristic approximation of why visual perturbations can compete with textual constraints. In the cross-attention layer, the output for a query token q is:

$$\text{Attn}(q, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{q\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{5}$$

Proposition 1 (Fixed-Normalizer Attention Ratio). *Let $k_v^{(1)} = k_v^{(0)} + \Delta_v$ denote the perturbed visual key. Under a fixed-normalizer approximation $Z^{(1)} \approx Z^{(0)}$, the perturbed and clean visual attention weights satisfy:*

$$\frac{w_v^{(1)}}{w_v^{(0)}} \approx \exp\left(\frac{q \cdot \Delta_v}{\sqrt{d}}\right) \tag{6}$$

Corollary 1. *If the corresponding textual attention term is approximately unchanged, then a sufficient condition for visual attention to dominate textual attention ($w_v^{(1)} > w_t^{(1)}$) is:*

$$q \cdot \Delta_v > \sqrt{d} \cdot \log\left(\frac{w_t^{(0)}}{w_v^{(0)}}\right) \tag{7}$$

This condition is more likely when $w_t^{(0)}/w_v^{(0)}$ is moderate (typically 0.5-2.0 in our experiments). The approximation is intended as intuition rather than a tight bound for full multi-layer MLLMs. Our semantic decoupling loss is designed to encourage movement along such value-associated directions. A fuller heuristic derivation is provided in Appendix A.

3.6 Algorithm: Ensemble-DI-FGSM

We adopt the Diverse Input (DI) and Momentum Iterative (MI) strategies to boost transferability. The update rule follows momentum-iterative gradient accumulation:

$$g_{i+1} = \mu \cdot g_i + \frac{\nabla_\delta \mathcal{L}(\delta_i)}{||\nabla_\delta \mathcal{L}(\delta_i)||_1} \tag{8}$$

Algorithm 1 PriceBlind Attack Generation

Require: Surrogate Models M, Screenshot S_t, Target Region v_{target}, Text T_{user}, Budget ϵ, Iterations N, Momentum μ

Initialize $\delta_0 \sim \mathcal{U}(-\epsilon, \epsilon)$, $g_0 = 0$
Compute encoder-specific centroids $\bar{e}^{(m)}_{cheap}$ from Anchor Bank $\mathcal{A}_{cheap}$
for $i = 0$ to $N - 1$ **do**
  Form perturbed screenshot $\tilde{S}^{(i)}_t = S_t(v_{target} + \delta_i)$
  Apply Diverse Input transform: $\hat{S}^{(i)}_t = \text{ResizePad}(\tilde{S}^{(i)}_t, p = 0.5)$
  Compute $\mathcal{L}_{action}$ on $\hat{S}^{(i)}_t$ and $\mathcal{L}_{dec}$ on $v_{target} + \delta_i$
  Set $\mathcal{L} = \mathcal{L}_{action} + \lambda \mathcal{L}_{dec}$ and back-propagate to obtain $\nabla_\delta \mathcal{L}$
  Update Momentum: $g_{i+1} = \mu \cdot g_i + \frac{\nabla_\delta \mathcal{L}}{||\nabla_\delta \mathcal{L}||_1}$
  Update Perturbation: $\delta_{i+1} = \text{Clip}_\epsilon(\delta_i - \alpha \cdot \text{sign}(g_{i+1}))$
end for
return $\tilde{S}^{(N)}_t = S_t(v_{target} + \delta_N)$

Table 1: White-box attack performance on E-ShopBench ($\epsilon = 8/255$). Values are descriptive scenario-level ASR percentages, rounded for readability.

Method	Victim Agent	ASR (%)↑
clean	Mobile-Agent-v2	4
Random Noise	Mobile-Agent-v2	7
Typographic	Mobile-Agent-v2	32
Adv. Pop-up	Mobile-Agent-v2	45
PriceBlind	**Mobile-Agent-v2**	**82**
clean	AppAgent	4
PriceBlind	**AppAgent**	**79**

Table 2: Black-box transfer ASR (%) on proprietary models under the single-turn coordinate-selection protocol (E-ShopBench, $n = 200$ scenarios). Values are descriptive and rounded for readability.

Method	GPT-4o	Gemini-1.5-Pro	Claude-3.5
clean	2	2	2
Typographic	19	15	13
Adv. Pop-up	22	20	18
PriceBlind (Single)	13	11	9
PriceBlind (Ensemble)	**41**	**39**	**35**

4 Experiments

4.1 Experimental Setup

- **Dataset (E-ShopBench):** 200 adversarial scenarios across Amazon (80), eBay (60), and Taobao (60). Each scenario contains one target (expensive) and one distractor (cheap) item.
- **Victim Models:** *white-box:* Mobile-Agent-v2 (Qwen-VL-Chat), AppAgent (LLaVA-1.6-Vicuna-7B). *black-box:* GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet.
- **Baselines:** clean, Random Noise ($\epsilon = 8/255$), Typographic ("Best Deal" overlay), Adv. Pop-up (Zhang et al., 2025b).
- **Metrics:** ASR (Attack Success Rate), LPIPS (perceptual similarity).
- **Statistical Reporting:** Scenario-level ASR numbers are reported as descriptive percentages; tables use rounded values for readability, and the prose emphasizes effect sizes and broad trends. We reserve mean ± std style reporting for distributional summaries such as attention weights.

Metrics and Statistical Reporting. ASR is defined as the percentage of evaluation scenarios in which the agent selects the pre-defined expensive target item (scenario-level). Unless otherwise specified, scenario-level ASR values should be read as descriptive summaries over E-ShopBench rather than formal inferential estimates. Results reported with ± in this paper denote mean ± std for underlying continuous or token-level measurements, not 95% confidence intervals.

4.2 Main Results

Table 1 reports white-box ASR on E-ShopBench.

PriceBlind substantially degrades price-constrained decision-making in white-box settings. On Mobile-Agent-v2, PriceBlind reaches 82% ASR, compared with 32% for Typographic and 45% for Adv. Pop-up. On AppAgent, PriceBlind reaches 79% while the clean condition remains at 4%. These are descriptive scenario-level results on the main E-ShopBench screenshot-based benchmark and serve as the paper's central white-box empirical claim.

4.3 Black-box Transferability Results

Table 2 presents results on black-box proprietary models.

Under the single-turn coordinate-selection protocol, the ensemble strategy substantially improves

black-box transfer. In Table 2, "PriceBlind (Single)" denotes the single-surrogate Qwen-VL setting. On GPT-4o, ASR rises from 13% for this single-surrogate setup to 41% for the ensemble; on Gemini-1.5-Pro, from 11% to 39%; and on Claude-3.5, from 9% to 35%. These black-box numbers should be interpreted as simplified protocol results rather than end-to-end multi-turn agent execution. Together with Table 1, they define the paper's main benchmark evidence; the appendix prompt-style stress tests are supplementary and are not intended as directly comparable replacements.

4.4 Qualitative Analysis

As a supplementary, non-systematic qualitative analysis, we manually inspected Chain-of-Thought (CoT) logs from successful attacks and grouped the observed rationales into three descriptive failure patterns. These language-level explanations are post-hoc rationalizations rather than direct observations of the optimized embedding direction, so the patterns below should be interpreted as symptoms of disrupted price-value association rather than standalone proof of mechanism. The percentages below denote descriptive shares within the inspected successful attacks, not benchmark-level ASR values:

Pattern 1: Visual-Semantic Overriding (45%): The agent acknowledges the high price but still reframes the item as compatible with a value-oriented choice. For instance: "Although this item is $499, it looks like the basic option or a better-value package."

Pattern 2: Numerical Blindness (30%): The agent simply ignored the text. When asked to "buy the cheapest", the agent clicked the expensive perturbed item stating: "This looks like the most basic model."

Pattern 3: Layout Hallucination (25%): The perturbation caused the agent to misalign the price text, believing a neighbor item's price belonged to the target item.

Taken together, these patterns suggest that the perturbation disrupts price-value association, but they should not be read as evidence that the representation literally becomes a cheap-item embedding. The same visual shift may be rationalized downstream as a basic model, a discounted/bundle interpretation, or a price-layout mismatch.

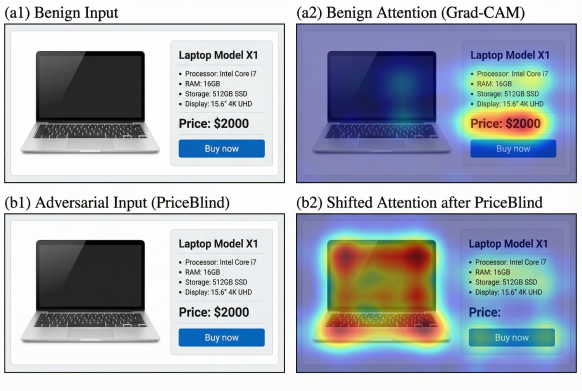

Figure 3: Grad-CAM saliency visualization. In the benign case (Top), the model highlights the OCR price regions. Under PriceBlind attack (Bottom), saliency shifts away from the price text towards the manipulated visual features.

Table 3: Attention weight distribution (mean ± std) across token types.

Condition	Visual	Price Text	Other
clean	$0.32_{\pm 0.05}$	$0.28_{\pm 0.04}$	$0.40_{\pm 0.06}$
PriceBlind	$0.58_{\pm 0.07}$	$0.11_{\pm 0.03}$	$0.31_{\pm 0.05}$
Change	**+81%**	**-61%**	-22%

4.5 Attention Weight Analysis

Figure 3 provides a qualitative Grad-CAM saliency visualization, while the table below reports supplementary quantitative evidence from Qwen-VL's cross-attention layers. Together, they provide descriptive support for the proposed mechanism rather than a second benchmark.

Supplementary finding: PriceBlind shifts cross-attention mass from price text tokens to visual tokens. **Numeric evidence:** Visual attention increases from $0.32_{\pm 0.05}$ to $0.58_{\pm 0.07}$ (+81%), while price-text attention decreases from $0.28_{\pm 0.04}$ to $0.11_{\pm 0.03}$ (-61%); "Other" tokens drop by 22%. **Scope caveat:** These token-level summaries are descriptive statistics over analyzed trajectories and provide supporting evidence consistent with the mechanism, rather than a standalone causal proof.

4.6 Ablation Study

Both $\mathcal{L}_{action}$ and $\mathcal{L}_{dec}$ are necessary, and ensemble training is important for black-box transfer under the single-turn coordinate-selection protocol. On Mobile-Agent-v2, either loss alone remains below 50% ASR, while the full method reaches about 80%; separately, transfer to GPT-4o rises from the

Table 4: Ablation on Mobile-Agent-v2 (white-box). ASR values are descriptive and rounded for readability; LPIPS is lower-is-better perceptual distance.

Configuration	ASR (%)	LPIPS↓
$\mathcal{L}_{action}$ only	45	0.028
$\mathcal{L}_{dec}$ only	39	0.031
Full ($\lambda = 1.5$)	**82**	0.033
w/o Momentum	69	0.031
w/o DI Transform	71	0.032
Single Model (Qwen)	76	0.031

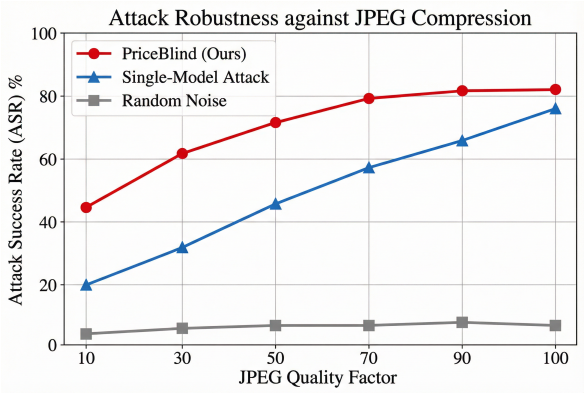

Figure 4: JPEG robustness curves. PriceBlind maintains high ASR even under aggressive JPEG compression and consistently outperforms baselines across the tested JPEG settings.

low teens for a single surrogate to about 40% for the ensemble. The ablation ASR/LPIPS values are descriptive white-box results on Mobile-Agent-v2, while the GPT-4o transferability numbers come from the simplified single-turn protocol.

4.7 Defense Evaluation

Within the main screenshot-based Mobile-Agent-v2 benchmark, robust encoders partially mitigate PriceBlind, while Verify-then-Act provides the strongest reduction. ASR drops from about 82% with no defense to about 58%/53% under Robust-CLIP and AdPO, and to the low teens under VtA; combining VtA with Robust-CLIP reduces ASR to below 10%. This stronger verification-heavy stack comes with a clean-accuracy trade-off, which falls into the high-80% range.

Why Robust Encoders Provide Partial Defense. We hypothesize that robust encoders primarily defend against perturbations that cause large, arbitrary embedding shifts. However, PriceBlind's semantic decoupling loss creates perturbations that move embeddings *within* the natural image manifold (towards a low-cost/value-associated anchor region), which robust training does not specifically address. This suggests that defending against semantic manipulation attacks may require different approaches than defending against random perturbations—specifically, methods that preserve semantic consistency between visual features and associated metadata.

Table 5: Defense results on Mobile-Agent-v2 in the screenshot-based setting. ASR and clean accuracy are descriptive percentages rounded for readability.

Defense	PriceBlind ASR	Clean Acc.
None (clean reference)	82	96
JPEG Compression (Q=50)	72	94
Gaussian Blur (σ=1.0)	68	92
Robust-CLIP (Schlarmann et al., 2024)	58	93
AdPO (Liu et al., 2025a)	53	95
Verify-then-Act (Hard)	12	88
VtA + Robust-CLIP	**9**	87

Table 6: Controlled attack comparison on E-ShopBench ($\epsilon = 8/255$). Transfer columns use the single-turn coordinate-selection protocol; values are descriptive ASR (%) rounded for readability.

Method	white-box	GPT-4o	Gemini	Stealth
MI-FGSM	65	18	16	High
DI-FGSM	69	22	20	High
X-Transfer	76	36	32	High
AdvDiffVLM	73	33	30	High
Pop-up Attack	45	22	20	Low
PriceBlind	**82**	**41**	**39**	**High**

4.8 Comparison with Related Attack Methods

We compare PriceBlind with recent transfer attack methods in Table 6. To ensure fair comparison, we re-implemented baseline attacks under controlled conditions on E-ShopBench, following the spirit of reliable robustness evaluation (Croce and Hein, 2020).

PriceBlind outperforms all controlled baseline attacks on both white-box and simplified transfer columns. White-box ASR is about 82% versus a strongest baseline in the mid-70% range, and GPT-4o/Gemini transfer reaches about 41%/39% versus strongest baselines in the mid-30% and low-30% ranges. Transfer columns are measured under a single-turn coordinate-selection protocol and should not be interpreted as end-to-end multi-turn task completion rates.

Table 7: Preliminary transferability probe to alternative encoder architectures. Values are descriptive ASR (%) rounded for readability.

Target Encoder	ASR (%)	Δ vs CLIP
CLIP-ViT-L/14 (reference)	82	–
SigLIP-SO400M	45	-37
EVA-CLIP-8B	52	-30
InternViT-6B	39	-43

4.9 Preliminary Probe to Alternative Encoders

As a limited supplementary probe beyond our main screenshot-based evaluation, we examined transferability to several alternative encoder architectures:

In this preliminary probe, transferability decreases on alternative encoders but remains non-trivial. Relative to the CLIP reference, EVA-CLIP stays above 50%, while SigLIP and InternViT drop to roughly the mid-40% and high-30% range, respectively. These values should be read as indicative supplementary checks rather than definitive cross-encoder estimates, and not as replacements for the main screenshot-based benchmark.

4.10 Layer-wise Attention Analysis

As a supplementary mechanistic probe, we analyzed attention weights across all 32 transformer layers of Qwen-VL. We found that the attention shift is most pronounced in layers 20-28 (the "reasoning" layers), with Pearson correlation $r = 0.73$ ($p < 0.001$, $n = 32$ layers) between layer depth and attention shift magnitude. Early layers (1-10) show minimal change, suggesting that low-level visual features remain intact while high-level semantic interpretation is manipulated. This layer-wise analysis is intended as supporting evidence rather than part of the paper's central benchmark claim.

As supplementary, non-systematic checks beyond our main screenshot-focused evaluation, Appendix E reports preliminary results for DOM/AX-tree observation settings, and Appendix J summarizes prompt-style stress tests that are not intended as directly comparable replacements for the main benchmark.

5 Discussion

Root Cause: Dataset Bias in Pre-training. We argue that the root cause of VDH is not just the architecture, but the pre-training data. Datasets like LAION-5B contain massive amounts of noisy image-text pairs where the model learns that "visual appearance" implies "textual attributes". PriceBlind exploits this correlation by nudging the embedding toward low-cost/value-associated directions, effectively leveraging the model's own statistical biases against it.

Novelty Relative to CLIP-Space Attacks. We acknowledge that embedding-space manipulation for VLMs has been explored in prior work. PriceBlind's contribution is: (1) the identification of VDH as a specific vulnerability pattern in agentic financial tasks, (2) the semantic decoupling loss that targets price-value associations rather than generic class boundaries, and (3) a broad screenshot-based evaluation on E-ShopBench showing strong vulnerability in the main clear-constraint benchmark and non-trivial transfer under our evaluated single-turn black-box protocol.

The Arms Race: Attack vs. Defense. Our work highlights a classic security arms race. While we propose PriceBlind, defenses like separate OCR modules can mitigate it. However, attackers can counter this by attacking the OCR engine itself. This suggests that a static defense is insufficient. Future agent architectures must incorporate *dynamic verification*, where the agent actively seeks corroborating evidence when visual and textual signals conflict.

Implications for Future Agents. The vulnerability exposed by PriceBlind suggests that current "Late Fusion" architectures are insufficient for high-stakes decision making. Future agents may need to adopt a "Verify-then-Act" architecture, where critical constraints are verified by specialized, non-neural symbolic modules before the neural planner is allowed to execute an action. Reliance on a single, monolithic MLLM for both perception and reasoning is a single point of failure.

Broader Impact on Autonomous Systems. Beyond e-commerce, the Visual Dominance Hallucination vulnerability has significant implications for autonomous systems in healthcare, finance, and critical infrastructure. Medical imaging agents could be manipulated to misinterpret diagnostic images, while financial document processing systems might be deceived about numerical values in contracts or invoices. Our findings suggest that any system where visual perception directly influences high-stakes decisions requires additional verification layers. The fundamental tension between

efficiency (single-model inference) and security (multi-modal verification) will shape the design of next-generation autonomous agents.

Recommendations for Practitioners. Based on our findings, we recommend the following for deploying vision-language agents in production: (1) implement redundant verification for price-critical decisions using separate OCR pipelines, (2) establish confidence thresholds that trigger human review when visual-textual conflicts are detected, (3) maintain audit logs of agent decisions for post-hoc analysis, and (4) consider ensemble approaches that aggregate predictions from multiple encoder architectures to reduce single-point vulnerabilities.

6 Conclusion

We presented **PriceBlind**, a controlled study of screenshot-based visual vulnerability in e-commerce agents. On our main screenshot-based E-ShopBench benchmark with clear price constraints, we provided evidence that "Visual Dominance" is an important weakness in current MLLMs, allowing attackers to override price evidence via stealthy image perturbations in white-box settings. Under the evaluated single-turn coordinate-selection protocol in a simplified layout-aware setting, black-box transfer remains non-trivial, while stronger verification-heavy stacks can substantially reduce risk at some clean-accuracy cost. As agents become autonomous economic actors, solving this **Instruction-Perception Conflict** is a prerequisite for robust deployment.

Our key findings can be summarized as follows: (1) On our main screenshot-based E-ShopBench benchmark with clear price constraints, Visual Dominance Hallucination yields white-box ASR around 80%; (2) The semantic decoupling loss improves attack effectiveness by disrupting price-value associations in the embedding space; (3) Black-box transfer remains non-trivial (roughly 35–41% ASR) under the evaluated single-turn coordinate-selection protocol; (4) Standard robust encoders reduce ASR only partially, whereas Verify-then-Act combined with robust encoders can reduce ASR to below 10% with clean accuracy in the high-80% range.

Looking forward, we believe that the security of autonomous agents will become increasingly critical as these systems handle more financial transactions. The community must develop principled approaches to verify agent decisions, particularly when visual and textual signals conflict. We hope that PriceBlind serves as a wake-up call for the development of more robust multimodal architectures.

Limitations

Benchmark Scale. E-ShopBench contains 200 scenarios, which is modest compared to large-scale benchmarks. Future work should evaluate on larger benchmarks with more diverse UI templates.

CLIP Encoder Dependency. Our attack is optimized for CLIP-based visual encoders. Our preliminary cross-encoder probe suggests lower transferability on alternative encoders overall, especially on the true non-CLIP encoders SigLIP (Zhai et al., 2023) and InternViT, but these supplementary checks are not yet a definitive cross-encoder benchmark.

Layout-Aware Optimization. Our formal threat model already assumes screenshot-conditioned or stable-layout target coordinates during perturbation generation. Broader dynamic UI changes may therefore reduce attack success outside this controlled setting.

Computational Cost. Per-image perturbation requires ∼45 seconds on a single A100 GPU. We did not explore Universal Adversarial Perturbations which could amortize the cost.

Real-World Deployment Gap. Our evaluation uses simulated e-commerce environments rather than live production systems. Real-world platforms may employ additional security measures, rate limiting, or anomaly detection that could affect attack success rates. The gap between controlled experiments and real-world deployment remains an important consideration for interpreting our results.

User Intent Variability. Our main evaluation protocol prioritizes clear price constraints to isolate instruction-following failures. We additionally report implicit and vague prompt settings as supplemental stress tests (Appendix J), but those checks are intended to illustrate trend direction rather than serve as directly comparable replacements for the main benchmark figures. Broader real-world intent diversity remains under-explored and may affect observed attack rates.

Additional limitations including black-box evaluation methodology, theoretical analysis scope, and screenshot-centric evaluation scope are discussed in Appendix F.

Ethics Statement

All experiments were conducted in a restricted sandbox environment. No real transactions were executed, and no adversarial images were uploaded to live public e-commerce platforms. We believe that responsible disclosure of these vulnerabilities is essential for improving the security of autonomous agents before they are widely deployed in high-stakes financial applications.

Responsible Disclosure. We encourage the research community to use our findings constructively to develop more robust multimodal architectures rather than for malicious purposes.

Broader Implications. The proliferation of autonomous agents in financial transactions raises fundamental questions about accountability and trust. Our work highlights the need for regulatory frameworks governing autonomous economic actors. We hope our findings contribute to the development of safer AI systems.

Acknowledgements

Generative AI tools were used to assist with language polishing and to draft parts of the experimental code. All scientific content, code, and results were reviewed, edited, and validated by the authors, who take full responsibility for the work.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Chiyu Chen, Xinhao Song, Yunkai Chai, Yang Yao, Haodong Zhao, Lijun Li, Jie Li, Yan Teng, Gongshen Liu, and Yingchun Wang. 2025. GhostEI-Bench: Do mobile agents resilience to environmental injection in dynamic on-device environments? *arXiv preprint arXiv:2510.20333*.

Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *CVPR*.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *CVPR*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. 2025. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*.

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. 2025. X-Transfer attacks: Towards super transferable adversarial attacks on CLIP. In *ICML*.

Nikola Jovanović, Mislav Balunović, Dimitar Iliev Dimitrov, and Martin T. Vechev. 2023. FARE: Provably fair representation learning with practical certificates. In *ICML*.

Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem AlShikh, and Ruslan Salakhutdinov. 2024. OmniACT: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *ECCV*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*.

Chaohu Liu, Tianyi Gui, Yu Liu, and Linli Xu. 2025a. AdPO: Enhancing the adversarial robustness of large vision-language models with preference optimization. *arXiv preprint arXiv:2504.01735*.

Guohong Liu, Jialei Ye, Jiacheng Liu, Yuanchun Li, Wei Liu, Pengzhi Gao, Jian Luan, and Yunxin Liu. 2025b. Mobile GUI agents under real-world threats: Are we there yet? *arXiv preprint arXiv:2507.04227*.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2024. CoCo-Agent: A comprehensive cognitive MLLM agent for smartphone GUI automation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Ziqi Miao, Yi Ding, Lijun Li, and Jing Shao. 2025. Visual contextual attack: Jailbreaking MLLMs with image-driven context injection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9627–9644.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. 2020. A self-supervised approach for adversarial robustness. In *CVPR*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *AAAI*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. 2024. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*.

Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-Agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. In *Advances in Neural Information Processing Systems*, pages 2686–2710.

Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. Autodroid: Llm-powered task automation in android. In *MobiCom*.

Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. 2025. Dissecting adversarial robustness of multimodal LM agents. In *ICLR*.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *CVPR*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yanda Li, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2025a. AppAgent: Multimodal agents as smartphone users. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 70:1–70:20.

Tingwei Zhang, Rishi Jha, Eugene Bagdasarian, and Vitaly Shmatikov. 2024. Adversarial illusions in multimodal embeddings. In *USENIX Security*.

Yanzhe Zhang, Tao Yu, and Diyi Yang. 2025b. Attacking vision-language computer agents via pop-ups. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8387–8401.

Yitong Zhang, Ximo Li, Liyi Cai, and Jia Li. 2025c. Environmental injection attacks against GUI agents in realistic dynamic environments. *arXiv preprint arXiv:2509.11250*.

A Extended Heuristic Analysis

A.1 Cross-Attention Dynamics

Let $\mathbf{V} = [v_1, ..., v_n]$ be the visual token sequence and $\mathbf{T} = [t_1, ..., t_m]$ be the textual token sequence. In the cross-attention layer:

$$\text{Attn}(q, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{q\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \quad (9)$$

where $\mathbf{K} = [\mathbf{K}_v; \mathbf{K}_t]$ concatenates visual and textual keys.

Heuristic derivation for Proposition 1. Let

$$\begin{aligned} w_v^{(0)} &= \frac{\exp(q \cdot k_v^{(0)}/\sqrt{d})}{Z^{(0)}}, \\ w_v^{(1)} &= \frac{\exp(q \cdot (k_v^{(0)} + \Delta_v)/\sqrt{d})}{Z^{(1)}}. \end{aligned} \quad (10)$$

Under the approximation $Z^{(1)} \approx Z^{(0)}$, we obtain

$$\frac{w_v^{(1)}}{w_v^{(0)}} \approx \exp\left(\frac{q \cdot \Delta_v}{\sqrt{d}}\right). \quad (11)$$

If the relevant textual logits are approximately unchanged, then $w_t^{(1)} \approx w_t^{(0)}$, which gives

$$\frac{w_v^{(1)}}{w_t^{(1)}} \approx \frac{w_v^{(0)}}{w_t^{(0)}} \exp\left(\frac{q \cdot \Delta_v}{\sqrt{d}}\right). \quad (12)$$

Therefore, $w_v^{(1)} > w_t^{(1)}$ is implied when $q \cdot \Delta_v > \sqrt{d}\log(w_t^{(0)}/w_v^{(0)})$. This argument is heuristic: in real MLLMs, multi-layer fusion, residual connections, and normalization can all affect the exact threshold.

Limitations of This Analysis. This analysis assumes a simplified single-layer cross-attention model. Real MLLMs employ multiple attention layers with residual connections, layer normalization, and varying fusion strategies. The exact conditions for successful attack may vary across architectures.

A.2 Connection to CLIP Modality Gap

The effectiveness of $\mathcal{L}_{dec}$ can be explained through the geometry of CLIP's latent space. Studies have shown that CLIP embeddings exhibit a "modality gap" where image and text embeddings occupy

different regions of the hypersphere (Liang et al., 2022). Within the image manifold, semantic concepts form distinct clusters. Our attack exploits this structure by projecting the perturbed target-item embedding toward a low-cost/value-associated anchor region while maintaining pixel-level similarity.

Let $\mathcal{M}_{cheap}$ denote the low-cost anchor manifold and $\mathcal{M}_{luxury}$ the original luxury-item neighborhood. The semantic decoupling loss effectively solves:

$$\min_{\delta} d(\mathcal{E}_v(v_{target}+\delta), \mathcal{M}_{cheap}) \quad \text{s.t.} \quad ||\delta||_\infty \leq \epsilon \tag{13}$$

This geometric perspective explains why the attack transfers across models: all CLIP-based VLMs share similar embedding space geometry.

A.3 Momentum Smoothing Intuition

Momentum acts as an exponential moving average over successive gradients, which can damp abrupt direction changes even though it does not imply a literal reduction in the raw variance of the accumulated state. A schematic form of the update is:

$$g_{i+1} \approx \mu \cdot g_i + \nabla_\delta \mathcal{L}(\delta_i) \tag{14}$$

Unrolling this recurrence shows that the current update direction aggregates information from multiple previous steps with exponentially decaying weights. In our setting, this temporal smoothing helps stabilize optimization trajectories under the diverse-input transform and is best interpreted as a heuristic transferability aid rather than a formal variance-reduction guarantee.

B Extended Ablation Studies

B.1 Sensitivity to Lambda

Table 8: Extended ablation on the decoupling weight (lambda). ASR values are descriptive and rounded for readability.

Configuration	ASR (%)	LPIPS↓
Full ($\lambda = 0.5$)	72	0.030
Full ($\lambda = 1.0$)	79	0.032
Full ($\lambda = 1.5$)	**82**	0.033
Full ($\lambda = 2.0$)	82	0.038
Full ($\lambda = 3.0$)	79	0.052

Key Findings:

- **Lower-λ regime (e.g., $\lambda = 0.5$):** The attack behaves more like a standard coordinate attack. ASR is lower because the agent more often reads the price and aborts.
- **Mid-λ regime (1.0 - 2.0):** The ASR peaks. The image features are successfully decoupled from the "expensive" concept.
- **At $\lambda = 3.0$:** Visual quality degrades substantially (LPIPS = 0.052 $>$ 0.05), indicating the imperceptibility constraint is no longer well preserved at this setting.

B.2 Component Contribution Analysis

Table 9: Component contribution analysis based on descriptive scenario-level ASR values.

Configuration	ASR	Δ vs clean	Contribution
clean	4%	–	–
$\mathcal{L}_{action}$ only	45%	+41 pp	52%
$\mathcal{L}_{dec}$ only	39%	+35 pp	45%
Full PriceBlind	82%	+78 pp	100%

Using a unified Mobile-Agent-v2 clean reference (clean ASR $\approx$ 4%), coordinate targeting ($\mathcal{L}_{action}$) accounts for roughly half of the total improvement, while semantic decoupling ($\mathcal{L}_{dec}$) contributes slightly under half. The remaining few percentage points correspond to interaction/synergy between the two components.

B.3 Surrogate Model Sensitivity

Table 10: Surrogate model ablation for GPT-4o transfer. Values are descriptive and rounded for readability.

Surrogate Configuration	GPT-4o ASR (%)
Single Model (Qwen-VL)	13
Single Model (LLaVA-1.6)	16
Ensemble (Qwen + LLaVA)	41
Ensemble (+ MiniGPT-4)	43
Ensemble (+ InternVL)	45

This descriptive ablation suggests that attacking the common intersection of CLIP-based models is sufficient to obtain most of the downstream transfer, with diminishing returns after two strong surrogates.

C Transferability to Alternative Encoders

This preliminary probe suggests that the attack partially exploits CLIP-specific embedding geometry.

Table 11: Preliminary transferability probe to alternative encoder architectures. Values are descriptive and rounded for readability.

Target Encoder	ASR (%)	Δ vs CLIP
CLIP-ViT-L/14 (reference)	82	–
SigLIP-SO400M	45	−37
EVA-CLIP-8B	52	−30
InternViT-6B	39	−43

EVA-CLIP remains above 50% ASR, while the true non-CLIP encoders SigLIP and InternViT drop to roughly the mid-40% and high-30% range. We therefore treat these results as indicative supplementary evidence rather than a definitive cross-encoder study.

D Comparison with Related Attack Methods

Table 12: Controlled comparison with related attack methods on E-ShopBench. Values are descriptive ASR percentages rounded for readability.

Method	white-box	GPT-4o	Gemini	Stealth
MI-FGSM (Dong et al., 2018)	65	18	16	High
DI-FGSM	69	22	20	High
SSA (Naseer et al., 2020)	71	29	25	High
X-Transfer (Huang et al., 2025)	76	36	32	High
AdvDiffVLM (Guo et al., 2025)	73	33	30	High
Pop-up Attack	45	22	20	Low
PriceBlind	**82**	**41**	**39**	**High**

PriceBlind outperforms all baselines. The key advantage comes from our semantic decoupling loss, which specifically targets the price-value association rather than generic classification boundaries.

E Structured Observation Agents

DOM/AX-Tree Agents. Agents that operate on DOM or accessibility tree representations receive structured element attributes including text content, element type, and bounding boxes. Since price information is typically encoded as text attributes in the DOM, these agents may be more robust to visual perturbations.

These preliminary descriptive results suggest that semantic decoupling provides clear gains in screenshot-only and DOM+Screenshot settings. In DOM-only settings, the residual gain is about +5 percentage points and should not be over-interpreted as direct evidence for the same visual mechanism. For screenshot-only agents, $\mathcal{L}_{dec}$ contributes roughly +37 percentage points. These supplementary modality comparisons are not intended to replace the main screenshot benchmark.

Table 13: Preliminary descriptive comparison with grounding-targeted attacks across observation modalities.

Attack Type	Screenshot	DOM+Screenshot	DOM-only
Coordinate-only	45%	28%	8%
PriceBlind (Full)	82%	52%	13%
Gain from $\mathcal{L}_{dec}$	+37 pp	+24 pp	+5 pp

F Extended Limitations

Black-box API Evaluation Methodology. Our black-box evaluation on GPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet uses a single-turn coordinate selection setup where models receive the screenshot and instruction, then output a click coordinate. This design choice isolates the visual perception vulnerability and enables reproducible evaluation, but differs from full multi-turn agent deployments that include additional reasoning steps, tool use, and error recovery. The reported ASR quantifies vulnerability under this simplified protocol and does not directly estimate production multi-turn vulnerability.

Theoretical Analysis Scope. The cross-attention analysis provides intuition for the attention hijacking mechanism but does not tightly characterize the exact conditions under which the attack succeeds across all possible fusion architectures (early fusion, late fusion, cross-attention variants). A more rigorous analysis connecting specific architectural choices to vulnerability would strengthen the theoretical contribution.

White-box Surrogate Requirement. PriceBlind requires white-box access to at least one surrogate model. Fully black-box optimization (query-based) would be more practical but computationally expensive.

Screenshot-Centric Evaluation. Our main evaluation focuses on screenshot-based agents. We include a preliminary, non-systematic analysis of DOM/AX-tree observation settings in Appendix E, but this does not yet constitute a comprehensive study across modern structured-observation agent pipelines. Recent benchmarks like OmniACT (Kapoor et al., 2024) and agent frameworks leveraging Set-of-Mark prompting (Yang et al., 2023)

remain important targets for future work.

G Detailed Experimental Configuration

All experiments were conducted on a server with 4x NVIDIA A100 (80GB) GPUs. Total GPU hours: approximately 120 hours.

Runtime Analysis. Per-image perturbation generation takes 45.2 ± 3.8 seconds (mean$\pm$std) on a single A100 GPU, broken down as: forward pass through surrogates (18s), backward pass and gradient computation (22s), projection and clipping (5s).

Victim Models:

- *Mobile-Agent-v2:* Uses Qwen-VL-Chat as the backbone via HuggingFace `transformers` (version 4.37.0).
- *AppAgent:* Uses LLaVA-1.6-Vicuna-7B with the official codebase.
- *Black-box APIs:* GPT-4o (gpt-4o-2024-05-13), Gemini-1.5-Pro (gemini-1.5-pro-001), Claude-3.5-Sonnet (claude-3-5-sonnet-20240620).

Attack Hyperparameters:

- $\epsilon = 8/255$, $N = 50$ iterations, $\mu = 0.9$, $\alpha = 1/255$
- DI-FGSM Probability: $p = 0.5$, resize range: 0.9-1.1
- Decoupling Weight: $\lambda = 1.5$, Repulsion Weight: $\beta = 0.5$
- Softmax Temperature: $\tau = 0.1$ for discrete grid outputs

H E-ShopBench Construction

E-ShopBench consists of XML-based UI layouts and corresponding screenshots.

- **Categories:** Electronics (50), Home Goods (50), Fashion (50), Accessories (50).
- **Price Distribution:** "Cheap" items: \$10–\$50. "Expensive" items: \$200–\$2000.
- **Layout Variation:** List Views (40%), Grid Views (40%), Detail Pages (20%).
- **Platform Distribution:** Amazon (80), eBay (60), Taobao (60).

I Visual Anchor Bank Details

I.1 Anchor Selection Criteria

We construct the anchor bank $\mathcal{A}_{cheap}$ by sampling 500 images of low-cost items chosen to reduce direct product overlap with E-ShopBench while emphasizing generic value-associated visual cues:

1. **Price Range:** Items priced below \$20
2. **Visual Simplicity:** Products with minimal branding, simple packaging
3. **Product Separation:** Avoid direct overlap in exact products or screenshots used in E-ShopBench

I.2 Category Distribution

- Office supplies (pens, notebooks, folders): 150 images
- Basic kitchenware (plastic containers, utensils): 150 images
- Generic household items (cleaning supplies, storage boxes): 200 images

I.3 Visual Characteristics

These items typically feature: simple backgrounds, generic branding, standard materials (plastic vs. brushed metal), utilitarian design, and basic product photography.

I.4 Embedding Computation

We pre-compute embeddings using **CLIP-ViT-L/14** (OpenAI's official checkpoint, 428M parameters). For ensemble attacks, we additionally compute an encoder-specific centroid $\bar{e}^{(m)}_{cheap}$ for each surrogate model's visual encoder from the same anchor bank.

J Impact of Instruction Phrasing

As a supplementary descriptive stress test, rather than a replacement for the main clear-constraint benchmark, we probed PriceBlind under three prompt styles:

- **Explicit Constraint:** "Buy the item strictly under \$50." $\rightarrow$ ASR: about 68%
- **Implicit Constraint:** "I need a budget-friendly option." $\rightarrow$ ASR: about 91%
- **Urgent/Vague:** "Get me the best deal ASAP." $\rightarrow$ ASR: about 95%

These supplementary checks suggest that ASR is highest for **Implicit Constraints** and **Vague Prompts**, where the model relies more heavily on its own semantic interpretation. Even with **Explicit Constraints**, the attack remains materially effective in this stress-test setting. These prompt-style results are intended to illustrate qualitative trend differences and should not be interpreted as directly comparable replacements for the main E-ShopBench benchmark figures.