DualTAP: A Dual-Task Adversarial Protector for Mobile MLLM Agents

Fuyao Zhang[1], Jiaming Zhang[1⋆], Che Wang[1], Xiongtao Sun[1], Yurong Hao[1], Guowei Guan[1], Wenjie Li[1], Longtao Huang[2], and Wei Yang Bryan Lim[1]

[1] Nanyang Technological University, Singapore
[2] Alibaba Group, China

Abstract. The reliance of mobile GUI agents on Multimodal Large Language Models (MLLMs) introduces a severe privacy vulnerability: screenshots containing Personally Identifiable Information (PII) are often sent to untrusted, third-party routers. These routers can exploit their own MLLMs to mine this data, violating user privacy. Existing privacy perturbations fail the critical dual challenge of this scenario: protecting PII from the router's MLLM while simultaneously preserving task utility for the agent's MLLM. To address this gap, we propose the **Dual-Task Adversarial Protector (DualTAP)**, a novel framework that, for the first time, explicitly decouples these conflicting objectives. DualTAP trains a lightweight generator using two key innovations: (i) a contrastive attention module that precisely identifies and targets only the PII-sensitive regions, and (ii) a dual-task adversarial objective that simultaneously minimizes a task-preservation loss (to maintain agent utility) and a privacy-interference loss (to suppress PII leakage). To facilitate this study, we introduce PrivScreen, a new dataset of annotated mobile screenshots designed specifically for this dual-task evaluation. Comprehensive experiments on six diverse MLLMs (e.g., GPT-5) demonstrate DualTAP's state-of-the-art protection. It reduces the average privacy leakage rate to 31.7 percentage points (a 2.6× relative improvement) while, critically, maintaining an 80.8% task success rate—a negligible drop from the 83.6% unprotected baseline. DualTAP presents the first viable solution to the privacy-utility trade-off in mobile MLLM agents. Our code is available at `https://github.com/fyzhang1/DualTAP`.

Keywords: mobile MLLM agents; · PII leakage · adversarial protector

1 Introduction

Multimodal large language models (MLLMs) are increasingly the core reasoning engines for Graphical User Interface (GUI) mobile agents [30, 32, 41]. Leveraging these models, such agents can handle a wide range of real-world tasks, including personal assistance, travel planning, and financial operations. As illustrated in Fig. 1 (a), these agents typically operate on a continuous Per-

⋆ Corresponding author.

ceive–Router–MLLM loop. The Router acts as a centralized API scheduler, receiving user screenshots and task instructions and dispatching requests to a curated set of MLLM APIs to obtain an optimal response. This interaction paradigm, while facilitating coherent and efficient task completion [21, 42], introduces significant privacy risks.

The core vulnerability lies in the router, which we assume to be an honest-but-curious adversary. In a typical GUI agent scenario, a single user task is decomposed into multiple subtasks (Fig. 1 (a)), generating a large volume of sequential screenshots. The router gains access to this entire stream, allowing it to correlate information across continuous interactions. This creates a potent attack vector for the router provider: by leveraging its own MLLMs, it can mine the screenshot stream to extract and reconstruct sensitive personally identifiable information (PII). Such unauthorized processing and profiling of personal data directly violates the core principles of the EU General Data Protection Regulation (GDPR) [5] and California Consumer Privacy Act (CCPA) [16], amplifying the privacy threat and exposing users to legal and reputational risks.

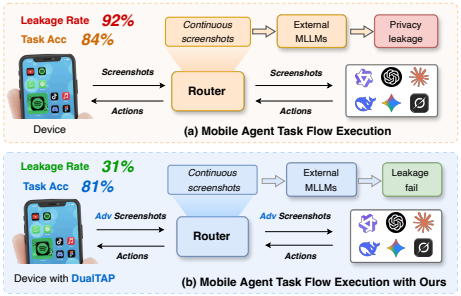

Fig. 1: The Dual Task Completion workflow of the GUI Agent under the Perceive–Router–MLLM paradigm. It shows two task flows: (a) without protection, the agent completes the functional task but leaks private information; (b) with DualTAP protection, privacy leakage is mitigated while normal functionality is preserved.

Existing visual privacy-preserving techniques encompass a broad area of applications, ranging from preventing unauthorized facial recognition [2, 15, 23, 25, 47] to specific visual regions of interest [19, 46]. To achieve robust protection across these domains, adversarial perturbations [12, 19, 44] have emerged as a particularly effective strategy, adding structured noise to inputs to disrupt an MLLM's ability to recognize sensitive content. However, these methods prove inadequate for the mobile agent scenario, as they fail to address the critical *dual challenge*: (i) preserving *task validity* for the agent's own MLLMs, while (ii) protecting *user privacy* against the router's MLLMs. Consequently, when enhancing privacy protection, most existing methods substantially compromise task utility.

To address this gap, we propose **Dual-T**ask **A**dversarial **P**rotector (**DualTAP**), a pluggable module designed to protect PII within the mobile agent ecosystem (Fig. 1 (b)). Our generator first incorporates a contrastive attention module to precisely identify and target regions sensitive to privacy cues. We then optimize the generator using a dual-task adversarial objective, which simultaneously minimizes a task-preservation loss (ensuring agent utility) and maximizes a privacy-interference loss (suppressing sensitive information leakage). To fa-

cilitate this study, we introduce **PrivScreen**[3], a new benchmark dataset for evaluating privacy leakage in mobile MLLM agents. It contains over 500 high-resolution screenshots with synthetically injected PII, sourced from 10 common mobile applications across diverse usage scenarios.

We conduct a comprehensive evaluation of six state-of-the-art MLLMs, spanning open source (Qwen2.5, InternVL3), GUI specialized (Holo1.5, UI-TARS), and commercial systems (GPT-5, Gemini 2.0 Flash). DualTAP provides robust protection: on our primary privacy leakage metric, it achieves an average reduction in leakage rate by 31.7 percentage points (a 2.6× relative improvement over the strongest baseline). Critically, it preserves utility with an 80.8% task success rate, negligible degradation from the 83.6% of the original, unprotected system. These results establish DualTAP as the state-of-the-art for privacy protection in mobile MLLM agents. Our contributions are three-fold:

- We propose **DualTAP**, a novel dual-task adversarial framework that jointly optimizes for task preservation and privacy interference, effectively protecting sensitive information in screenshots without compromising agent performance.
- We introduce **PrivScreen**, a new dataset of 500+ privacy-sensitive screenshots with injected PII across 10 mobile apps, designed to evaluate and benchmark privacy leakage and protection mechanisms in GUI agents.
- We demonstrate through comprehensive experiments that DualTAP achieves state-of-the-art privacy protection, significantly reducing information leakage while maintaining a high task success rate for the mobile agent.

2 Related Work

Mobile GUI Agents Mobile GUI agents have emerged as a research hotspot for automating complex tasks [8, 30, 41]. Early approaches often relied on structured data or accessibility services. However, recent advances are dominated by MLLMs acting directly on visual input. These agents employ visual grounding to map natural-language instructions to specific UI elements [24] or integrate it with chain-of-thought (CoT) reasoning to improve trustworthy visual-language reasoning in task execution [17, 35]. A standard workflow involves capturing device screenshots and routing them via APIs to MLLMs for contextual awareness and decision-making [6, 11, 38]. This paradigm enables end-to-end task execution without direct programmatic access to app internals [7, 43]. Modern systems build on this by incorporating element-level metadata to enhance robustness [27,31] or adopting CoT prompting to improve interpretability and mitigate errors in long-horizon tasks [35, 39].

PII Extraction by MLLMs Early PII extraction relied on static scanning with keywords or regular expressions, enhanced by OCR [4] to parse screen data and match predefined sensitive patterns [13, 20, 34]. These methods depend on fixed

[3] `https://huggingface.co/datasets/fyzzzzzz/PrivScreen`

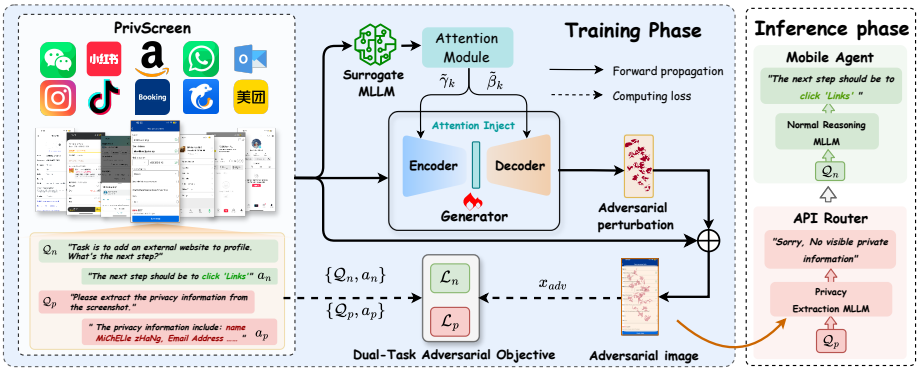

Fig. 2: Overview of the proposed **DualTAP** framework. **(a) Training Phase**: A perturbation generator $\mathcal{G}_\phi$ is trained using a dual-task objective. It learns to produce a perturbation δ based on a contrastive attention map $\mathcal{A}(x)$, guided by a surrogate MLLM $\mathcal{M}_\theta$. The objective is to simultaneously minimize the task-preservation loss $\mathcal{L}_\text{n}$ (for normal tasks $\mathcal{Q}_n$) and the privacy-interference loss $\mathcal{L}_\text{p}$ (for privacy tasks $\mathcal{Q}_p$). **(b) Inference Phase**: The pre-trained generator is deployed on the mobile device. It adds the optimized perturbation δ to the raw screenshot x to create x_adv. This x_adv is sent to the untrusted router, which blocks the router's Privacy Extraction MLLM ($\mathcal{M}_e$) from extracting PII, while still allowing the agent's Normal Task MLLM ($\mathcal{M}_a$) to function correctly.

signatures, lack contextual insight, and struggle to detect natural-language PII or link entities across screenshots for user profiling [14, 18]. Advanced MLLMs integrate visual encoders with language models, boosting the semantic understanding of on-screen content [26]. State-of-the-art models like GPT-5 [10] and Gemini-2.5 [28] interpret document semantics, UI details, and cross-frame associations, enabling comprehensive PII inference [37, 40] and uncovering hidden sensitive cues in complex interfaces.

Privacy Protection via Adversarial Perturbations As a promising privacy mechanism, adversarial perturbations inject structured noise into inputs to disrupt MLLM recognition [9, 33, 36]. Recent advances include Co-Attack [45], which breaks image-text alignment via fusion-stage gradients; AnyAttack [44], using a self-supervised generator; and FOA-Attack [12], which enhances transferability through global and local feature optimization. VIP [19] protects visual privacy by using targeted perturbations to mask sensitive regions in an image selectively. AVIH [25] proposes an end-to-end method that uses perturbations to make images unrecognizable to human eyes for privacy protection. Despite these advances, existing methods primarily focus on general tasks like image captioning or VQA. Their applicability to mobile agents is limited, as they are not designed for the specific dual objective of selectively suppressing PII while preserving the UI-grounded task execution capabilities of the agent.

3 Method

3.1 Preliminary

Threat Model We consider a privacy threat through the screenshot channel. The agent periodically captures screenshots x and forwards them to remote MLLM APIs via a routing service (e.g., OpenRouter). We assume an honest-but-curious router, which faithfully relays requests but also analyzes the data it processes. The router employs a *Privacy Extraction MLLM* $\mathcal{M}_e$ to analyze screenshots and extract sensitive PII for user profiling. It does not modify content or interfere with the agent's *Normal Task MLLM* $\mathcal{M}_a$.

Problem Formulation Our protector targets two black-box MLLMs: the router's $\mathcal{M}_e$ and the agent's $\mathcal{M}_a$. Both process the same input screenshot $x \in [0,1]^{3\times H\times W}$. Given a privacy-related question-answer (QA) set $(\mathcal{Q}_p, a_p)$, $\mathcal{M}_e$ aims to extract the private text a_p. Conversely, for a normal task QA set $(\mathcal{Q}_n, a_n)$, $\mathcal{M}_a$ uses chain-of-thought reasoning $\mathcal{C} = (c_1, \ldots, c_k)$ to derive the benign answer a_n. Our objective is to train a generator $\mathcal{G}_\phi$ that synthesizes an L_∞-bounded adversarial perturbation δ, yielding $x_{\text{adv}} = \text{clip}_0^1(x + \delta)$ where $\|\delta\|_\infty \leq \varepsilon$. This achieves a dual goal: (i) disrupting $\mathcal{M}_e$'s privacy extraction, while (ii) preserving $\mathcal{M}_a$'s utility for normal tasks. Due to the black-box nature of $\mathcal{M}_e$ and $\mathcal{M}_a$, we employ a white-box surrogate model $\mathcal{M}_\theta$, which approximates the conditional distribution $p_\theta(a \mid q, x)$, to optimize $\mathcal{G}_\phi$. We rely on the transferability of perturbations from $\mathcal{M}_\theta$ to the target models.

3.2 Dual-Task Adversarial Protector

Overview As shown in Fig. 2, **DualTAP** formulates the privacy-utility balance as a dual-objective optimization problem. We first introduce a contrastive attention module to generate a spatial map $\mathcal{A}(x)$ that isolates regions pertinent to privacy cues. This map is integrated into a U-Net-style generator $\mathcal{G}_\phi$ to guide the allocation of perturbations. Second, we optimize $\mathcal{G}_\phi$ using a dual-task adversarial objective, which explicitly trains the generator to trade off task usability and privacy security.

Contrastive Attention Module To isolate regions that are uniquely sensitive to privacy cues, we introduce an attention map $\mathcal{S}(\cdot)$. This map selectively highlights these responsive areas while occluding surrounding input regions, thereby distinguishing them from semantically salient features unrelated to privacy risks. In doing so, it enables precise, targeted generation of adversarial perturbations:

$$\mathcal{S}(x; \mathcal{Q}) = \left| \frac{\partial \mathcal{L}_{\text{nll}}(x; \mathcal{Q})}{\partial x} \right| \in \mathbb{R}^{3\times H\times W}, \tag{1}$$

where $\mathcal{L}_{\text{nll}}(x; \mathcal{Q}) = \sum_{(q,a)\in\mathcal{Q}} \big(-\log p_\theta(a \mid q, x) \big)$ is the total negative log-likelihood (NLL) loss over a QA set $\mathcal{Q}$, using the frozen surrogate model $\mathcal{M}_\theta$. This map quantifies the aggregate degradation in model confidence for each

image x from the dataset $\mathcal{D}$. To disentangle privacy-sensitive regions from task-salient regions, we propose a contrastive attention module:

$$\mathcal{A}(x) = \mathrm{ReLU}(\mathcal{S}_{\mathrm{p}}(x) - \mathcal{S}_{\mathrm{n}}(x)), \tag{2}$$

where $\mathcal{S}_{\mathrm{p}}(x) = \mathcal{S}(x; \mathcal{Q}_{\mathrm{p}})$ is the saliency map for privacy-oriented QA pairs and $\mathcal{S}_{\mathrm{n}}(x) = \mathcal{S}(x; \mathcal{Q}_{\mathrm{n}})$ is for normal task pairs. The ReLU function ensures that all values in the saliency map remain non-negative. This formulation isolates activations dominated by privacy-specific attention, yielding a spatially adaptive mask $\mathcal{A}(x)$ for subsequent perturbation allocation.

Dual-Task Adversarial Objective We employ a U-Net-style network $\mathcal{G}_\phi$ as the perturbation generator, which takes an input image x and the contrastive attention map $\mathcal{A}(x)$ to produce a targeted perturbation. The objective is to generate noise that selectively shields privacy-sensitive regions while preserving overall task utility. To achieve this dual objective, we integrate the contrastive attention module into the generation process via affine modulation at each resolution level k of the generator's decoder. Specifically, at level k, the attention map $\mathcal{A}(x)$ is downsampled to the current resolution, yielding $\mathcal{A}_k$. A lightweight convolutional network $\mathbf{L}_{\mathrm{cnn}}^{(k)}$ (part of $\mathcal{G}_\phi$) takes $\mathcal{A}_k$ as input and predicts two modulation coefficient maps: γ_k and β_k. These are mapped to bounded scale/shift coefficients:

$$\tilde{\gamma}_k = 1 + s\,\gamma_k, \qquad \tilde{\beta}_k = s\,\beta_k, \tag{3}$$

where s is a fixed strength hyperparameter. Attention is injected by affine-modulating the k-th feature map $\mathcal{F}_k$ to emphasize attended regions and add signal where attention is high:

$$\hat{\mathcal{F}}_k = \tilde{\gamma}_k \odot \mathcal{F}_k + \tilde{\beta}_k, \tag{4}$$

with $\odot$ denoting element-wise multiplication. Let $\mathcal{G}'_\phi$ denote $\mathcal{G}_\phi$ the equipped with this injection at every layer; its output is squashed by tanh to yield a per-pixel perturbation in $[-1, 1]$:

$$\tilde{\delta}(x, \mathcal{A}) = \varepsilon\,\mathcal{G}'_\phi(x; \mathcal{A}), \quad \|\tilde{\delta}(x, \mathcal{A})\|_\infty \le \varepsilon. \tag{5}$$

To explicitly balance task utility and privacy protection through our novel dual-task mechanism, we define two complementary losses: the task-preservation loss $\mathcal{L}_{\mathrm{n}}(x_{\mathrm{adv}})$, which encourages high confidence in responses to non-private queries $\mathcal{Q}_n$ by minimizing negative log-likelihood, and the privacy-interference loss $\mathcal{L}_{\mathrm{p}}(x_{\mathrm{adv}})$, which penalizes fidelity on privacy-oriented ones:

$$\begin{cases} \mathcal{L}_{\mathrm{n}}(x_{\mathrm{adv}}) = \displaystyle\sum_{(q,a)\in\mathcal{Q}_{\mathrm{n}}} -\log p_\theta(a \mid q, x_{\mathrm{adv}}), \\ \mathcal{L}_{\mathrm{p}}(x_{\mathrm{adv}}) = \displaystyle\sum_{(q,a)\in\mathcal{Q}_{\mathrm{p}}} -\log p_\theta(a \mid q, x_{\mathrm{adv}}). \end{cases} \tag{6}$$

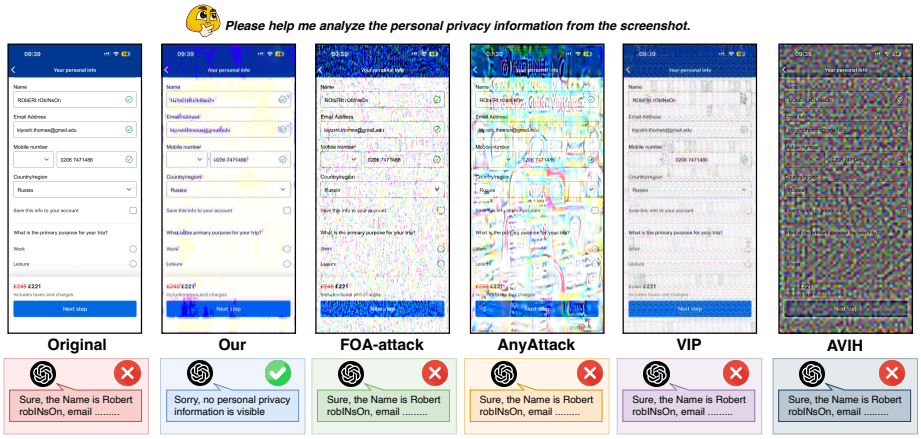

Fig. 3: Comparison of adversarial images generated by different baselines after adding perturbations to the same privacy-sensitive image.

The generator $\mathcal{G}'_\phi$ is optimized via stochastic gradient descent to minimize the composite loss:

$$\min_{\phi} \quad \alpha \mathbb{E}\big[\mathcal{L}_{\mathrm{n}}(x_{\mathrm{adv}})\big] - \beta \mathbb{E}\big[\mathcal{L}_{\mathrm{p}}(x_{\mathrm{adv}})\big], \tag{7}$$

where $\alpha > 0$ and $\beta > 0$ empirically trades off task fidelity against leakage suppression. This dual-loss formulation ensures that minimizing $\mathcal{L}_{\mathrm{n}}$ preserves agent performance on normal tasks $\mathcal{Q}_{\mathrm{n}}$, while maximizing $\mathcal{L}_{\mathrm{p}}$ actively suppresses confidence on privacy extractions $\mathcal{Q}_{\mathrm{p}}$ (as $\log p_\theta < 0$ for $p_\theta < 1$). By integrating the attention-guided modulation, it achieves decoupled task control, ultimately yielding state-of-the-art privacy protection.

4 Experiments

4.1 Experimental Setup

Baselines. We evaluate three categories of visual privacy protection methods: general-purpose approaches (AnyAttack [44] and FOA-Attack [12]), a region-specific approach (VIP [19]), and an end-to-end approach (AVIH [25]). Specifically, for AnyAttack and FOA-Attack, we treat the clean image as a screenshot containing sensitive information and the target image as a non-sensitive screenshot. We then generate adversarial perturbations to obscure potential privacy leaks. For VIP, we first mark the text locations within the image and rewrite the target question for optimization as *"What is the private information in the image?"*. For AVIH, we utilize the encrypted image. Additionally, we include a baseline representing the undefended scenario, in which clean screenshots are processed directly without the addition of any noise.

Table 1: Main experimental results comparing four baseline methods across seven metrics on 2 Open-Source MLLMs (O-MLLMs), 2 Commercial MLLMs (C-MLLMs), and 2 Specialized GUI Agents. Best results are in **bold**. ↑ indicates metrics where higher values are better; ↓ indicates metrics where lower values are better.

	Model	Attack Method	Acc ↑	LR ↓	MS ↓	BertScore ↓	CS ↓	BLEU ↓	ROUGE-L ↓
O-MLLMs	InternVL3-5-8B	Original	83.00	96.19	92.96	0.7172	0.8866	0.5586	0.7770
		AnyAttack	78.00	90.48	86.38	0.5689	0.7671	0.3512	0.5702
		FOA-Attack	70.00	78.57	78.54	0.5706	0.7198	0.4229	0.6012
		VIP	80.00	85.31	83.52	0.5801	0.8089	0.4022	0.6273
		AVIH	77.00	72.56	71.31	0.5215	0.6941	0.3221	0.4836
		Ours	**83.00**	**24.29**	**38.16**	**0.1905**	**0.3996**	**0.0841**	**0.1275**
	Qwen2.5-VL-7B	Original	89.00	97.14	96.99	0.8675	0.9011	0.7830	0.8991
		AnyAttack	70.00	95.24	91.86	0.6721	0.8705	0.5164	0.7317
		FOA-Attack	73.00	83.81	82.84	0.6307	0.8150	0.5343	0.7309
		VIP	81.00	94.76	92.67	0.7556	0.8960	0.5932	0.7856
		AVIH	83.00	93.57	90.82	0.7116	0.8882	0.5389	0.7522
		Ours	**88.00**	**32.86**	**38.82**	**0.1998**	**0.4202**	**0.0908**	**0.1660**
C-MLLMs	GPT-5	Original	93.00	97.14	97.15	0.9260	0.9627	0.8584	0.9246
		AnyAttack	90.00	91.43	87.49	0.6757	0.8490	0.4978	0.7446
		FOA-Attack	80.00	80.95	79.72	0.6293	0.7532	0.4967	0.7004
		VIP	86.00	80.00	79.33	0.6662	0.7715	0.5284	0.6846
		AVIH	88.00	81.21	80.28	0.6632	0.7673	0.5019	0.6731
		Ours	**91.00**	**23.33**	**24.43**	**0.1589**	**0.2379**	**0.0941**	**0.1628**
	Gemini-2.0 flash	Original	87.00	96.67	96.72	0.8651	0.9574	0.8030	0.9302
		AnyAttack	82.00	97.14	96.66	0.7993	0.9400	0.6960	0.9001
		FOA-Attack	79.00	94.76	94.13	0.7401	0.9045	0.6362	0.8342
		VIP	84.00	94.79	92.76	0.7311	0.8830	0.5746	0.7828
		AVIH	77.00	89.56	87.35	0.6947	0.8331	0.5529	0.7005
		Ours	**87.00**	**57.62**	**61.41**	**0.3805**	**0.5193**	**0.1686**	**0.2592**
GUI Agents	Holo1.5-7B	Original	70.00	94.76	94.07	0.8351	0.9036	0.7182	0.8232
		AnyAttack	61.00	81.43	79.07	0.5711	0.7447	0.3886	0.5801
		FOA-Attack	58.00	78.10	78.29	0.5590	0.7571	0.3964	0.5852
		VIP	57.00	89.05	86.19	0.6559	0.8263	0.4633	0.6583
		AVIH	65.00	79.02	77.98	0.5667	0.7271	0.3712	0.5628
		Ours	**69.00**	**17.14**	**30.75**	**0.1327**	**0.3620**	**0.0302**	**0.0949**
	UI-TARS-7B	Original	80.00	93.81	93.10	0.7846	0.9079	0.6721	0.8363
		AnyAttack	**76.00**	81.43	80.11	0.5533	0.7615	0.3748	0.6057
		FOA-Attack	64.00	76.67	75.60	0.5349	0.7187	0.4226	0.6178
		VIP	66.00	89.05	85.83	0.6739	0.8288	0.5232	0.7223
		AVIH	61.00	71.93	70.10	0.5244	0.6981	0.3563	0.5980
		Ours	67.00	**34.76**	**38.35**	**0.2278**	**0.4535**	**0.1139**	**0.2034**

Datasets. This work introduces PrivScreen, a dual-task QA-style privacy protection dataset derived from real application screenshots. To construct PrivScreen, we first collect privacy-sensitive interfaces from 10 high-frequency mobile applications, following task flows from SPA-Bench [1] and Mobile-Agent-V3 [41]. We then inject semantically realistic but fictitious PII from [26] into corresponding UI elements and record fine-grained annotations, including the PII type and exact value. Although the concrete PII values are synthetic for safety and ethical reasons, they preserve the structural format, semantic meaning, and UI context of real-world PII, such as names, phone numbers, addresses, and profile information. The dataset comprises over 500 screenshots, augmented with over 1000 synthetic PII, and includes two QA annotations per image: (i) a privacy-

focused QA querying sensitive items (to assess leakage) and (ii) a utility-focused QA on general screen content (to evaluate functionality). Both QAs share the same screenshot to maintain identical visual distributions. An in-app split (80% training, 20% evaluation) minimizes distribution bias, while an additional held-out-app split is used to evaluate cross-app generalization.

Models. We benchmark three model categories: (1) open-source MLLMs, including *InternVL3* [48] and *Qwen2.5* [29]; (2) GUI-specialized MLLMs, such as *Holo1.5* [3] and *UI-TARS* [22]; and (3) commercial MLLMs, like *GPT-5* and *Gemini-2.0*. This diverse evaluation provides a robust benchmark across architectures and optimizations.

Metrics. To comprehensively evaluate our method and baselines, we assess normal task execution and privacy-protection efficacy metrics. For the normal task, which measures the agent's ability to correctly complete intended operations on screenshot content. We report Accuracy (Acc) as the success rate. For the privacy information task—aimed at suppressing leakage of sensitive fields, we evaluate at two complementary levels: (i) character level, using Match Score (MS) for exact text matching, Leakage Rate (LR) defined as the proportion of samples with MS > 0.6, BLEU for n-gram precision with word-order sensitivity, and ROUGE-L for longest-common-subsequence similarity; and (ii) semantic level, using BERTScore to compare contextual embeddings and Cosine Similarity (CS) to measure vector-based semantic equivalence.

Implementation Details. In this work, supervision targets only the answer span, excluding question and visual placeholder tokens from the loss. Perturbations are generated using a U-Net-style network, with its output passed through a tanh activation and constrained by an L_∞ norm ($\varepsilon = 128/255$) before being added to the original image and clipped to $[0, 1]$. This level of perturbations aims to suppress PII extraction while maintaining key UI semantics and spatial layout for robust GUI reasoning. The surrogate MLLM, *InternVL3-5-2B*, serves solely as the gradient target and remains frozen during training. Loss weights are $\alpha = \beta = 1.0$. Training uses Adam with a learning rate of 1×10^{-4}, batch size of 4, and 20 epochs. All experiments are run on a single NVIDIA L20 GPU.

4.2 Main Results

As shown in Table 1, the Original setting confirms the severe privacy risks inherent in MLLM-based GUI agents. Across all six evaluated models, Original exhibits catastrophic leakage, with LR consistently above 93% and the semantic metric BertScore indicating a high degree of information exposure. This establishes a critical need for robust protection. The experimental results mainly demonstrate the following key aspects:

– **SOTA Privacy Protection.** Our DualTAP framework achieves state-of-the-art privacy protection, outperforming all competing baselines (AnyAttack, FOA-Attack, VIP, and AVIH). This is evident in the results from SOTA

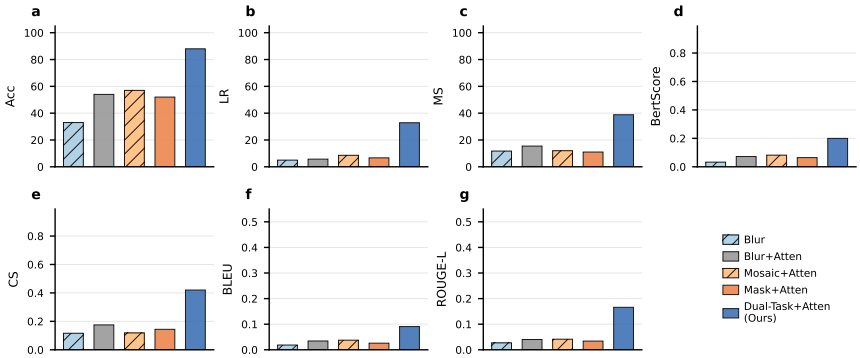

Fig. 4: Comparison of the performance of five different module combinations on the Mobile GUI Agent.

OpenSource models. For instance, on *Qwen-VL-7B-Instruct*, DualTAP reduces the LR to 32.86%, a massive reduction from the Original (97.14%) and superior to the baseline FOA-Attack (83.81%). This robust protection extends beyond simple keyword leakage to the semantic level: DualTAP's BertScore of 0.1998 is approximately 4.3× lower than the Original 0.8675 and 3.8× lower than VIP 0.7556, demonstrating a profound reduction in semantic information leakage.

- **Task Utility.** Crucially, DualTAP achieves this exceptional privacy protection without sacrificing task utility. It consistently maintains the highest Acc among all defense methods, performing on par with the unprotected Original model. For example, on *Qwen-VL-7B*, DualTAP's accuracy (88.00%) is higher than all other defenses. On *InternVL3-5-8B*, DualTAP achieves 83.00% and even matches the Original accuracy exactly. This demonstrates DualTAP's ability to effectively decouple the objectives of privacy protection and task preservation, resolving the critical privacy-utility trade-off.
- **Transferability.** To explore the transferability of DualTAP, we tested it on six SOTA models across three types (Open Source, Specialized GUI, and Commercial). The framework's robustness is evident across different architectures. For instance, on the commercial model *GPT-5*, DualTAP cuts the LR to a mere 23.33% and reduces the BertScore to 0.1589, a 5.8× reduction from the Original's 0.9260. DualTAP consistently establishes the best privacy protection and highest task utility (Acc) across all evaluated models, proving its effective transferability.

The results highlight the superiority of our dual-task adversarial objective and contrastive attention module. As shown in Fig. 3, this design targets adversarial perturbations to privacy-sensitive regions, effectively decoupling privacy interference from task preservation. Overall, experiments confirm our method achieves state-of-the-art privacy protection across diverse MLLM agents while maintaining strong practical utility.

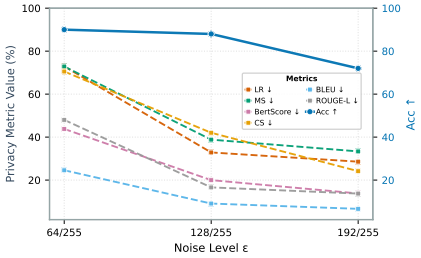

Fig. 5: Impact of the different bounds.

Table 2: Impact of the dual-loss.

Metrics	Privacy_only	Normal_only	**DualTAP**
Acc↑	64.00	89.00	88.00
LR↓	23.80	94.19	32.86
MS↓	20.48	92.52	38.82
CS↓	0.1706	0.8558	0.4202
BLEU↓	0.0570	0.5340	0.0908

4.3 Ablation Study

The Role of Attention. As shown in Fig. 4 (a), when only Blur is used, the accuracy on normal tasks is only about 33%, indicating that simple global blurring damages task-relevant key information. After introducing attention (Blur+Atten), the accuracy increases to about 54%, which demonstrates that concentrating the perturbation more on privacy regions can significantly reduce interference with normal tasks while maintaining a similar level of privacy strength. However, because the MLLM still encodes the entire image in a unified manner, attention cannot be perfectly precise, and a certain amount of perturbation inevitably remains in non-privacy regions. For strong occlusion methods such as mosaic and mask, which directly overwrite pixels, the perturbation is more likely to spill over into task areas, causing a catastrophic impact on normal-task performance. Among them, Mask+Atten achieves only 52% accuracy on normal tasks, even though it provides more thorough privacy masking.

The Importance of the Dual-Task Method. In contrast, our proposed Dual-Task+Atten method exhibits a more balanced performance across all metrics. On the one hand, as shown in Fig. 4 (f)(g), all privacy-protection scores are below 0.2, meaning that at the semantic level, the text is no longer sufficient to reconstruct private content, indicating that the privacy protection effect is not weakened. On the other hand, our method achieves over 80% accuracy on normal tasks. By jointly optimizing the objectives of privacy protection and task preservation during training, our method effectively constrains the spread of perturbations into normal-task regions and, while maintaining strong privacy protection, significantly alleviates the performance degradation on normal tasks.

Impact of the perturbation bound ε. As shown in Fig. 5, to evaluate the impact of noise levels ε at 64/255, 128/255, and 192/255, experiments show that escalating ε bolsters privacy by reducing privacy metrics from 80 $\sim$ 100% to 20 $\sim$ 40%, yet impairs accuracy Acc from 88% to 72%. We primarily employ $\varepsilon = 128/255$ because it offers accuracy near 64/255 with only a 5 $\sim$ 10% reduction, while providing privacy protection nearly on par with 192/255, which gains little despite greater utility degradation from intensified noise.

Table 3: Sensitivity to loss weights.

(α, β)	Acc↑	LR↓	MS↓	CS↓	BLEU↓
$(1.75, 0.25)$	84.00	45.71	52.10	0.5138	0.1410
$(1, 1)$	83.00	24.29	38.16	0.3996	0.0841
$(0.25, 1.75)$	70.00	18.01	30.33	0.3179	0.0896

Table 4: Efficiency comparison.

Method	E2E	No pretrain	Time
AnyAttack	✔	✘	< 0.3s
FOA-Attack	✘	✔	≈ 120s
AVIH	✔	✘	≈ 120s
VIP	✘	✔	≈ 600s
Ours	✔	✔	< 0.3s

Impact of dual-loss and loss weights. As shown in Table 2 and Table 3, we conduct an ablation study to validate the dual-loss (DualTAP) design. Removing the privacy loss leads to severe privacy leakage, while removing the task-preserving loss harms task performance (37.5% drop vs. DualTAP). These results confirm that a single loss cannot jointly ensure privacy protection and task effectiveness. α and β control the privacy–utility trade-off. Increasing α favors task accuracy but results in higher leakage, while increasing β further suppresses leakage at the cost of normal-task performance. We therefore use the balanced setting $(\alpha, \beta) = (1, 1)$ as the default.

4.4 Efficiency Comparison

Table 4 compares the computational efficiency and simplicity of different baselines. The proposed method demonstrates clear advantages in both end-to-end integration and real-time inference efficiency. Unlike FOA-Attack and VIP, which rely on multi-stage optimization and extensive offline pre-training, DualTAP operates as a fully end-to-end pipeline without additional pre-training, achieving inference times under 0.3s per image, comparable to lightweight methods like Anyattack but with a more stable and task-preserving design. In contrast, AVIH, FOA-Attack, and VIP require approximately 120s and 600s per image, making them impractical for real-time or mobile deployment. These results highlight that DualTAP achieves an optimal trade-off between privacy protection and computational cost, allowing efficient execution on edge devices or within mobile MLLM agents. The low latency and independence from pre-training make it particularly suitable for on-device inference scenarios, where energy efficiency, responsiveness, and user privacy must be simultaneously preserved.

4.5 Cross-Analysis of Multi-surrogate Models

As shown in Table 5, we investigate the impact of different attention surrogates and multi-surrogate configurations. The results demonstrate that DualTAP maintains a stable and consistent performance across all settings. Specifically, whether trained on a single attention surrogate (Qwen2b or InternVL2b) or an ensemble (Q+InternVL), the Acc exhibits only minor fluctuations, generally remaining robust between 77.00 and 91.00 across the target models. Similarly, the LR metric stays relatively stable, ranging from 17.43 to 33.14. The marginal performance gaps among these three configurations indicate that DualTAP is highly stable, ensuring effective transferability regardless of the chosen attention surrogate models.

Table 5: Multi-surrogate analysis.

Config	Metric	Qwen	Intern	GPT
Qwen2b	Acc↑	83.00	80.00	85.00
	LR↓	33.14	25.24	29.02
InternVL2b	Acc↑	88.00	83.00	91.00
	LR↓	32.86	24.29	23.33
Q+InternVL	Acc↑	82.00	77.00	81.00
	LR↓	20.95	17.43	19.65

Table 6: Robustness evaluation.

Scenario	Attack	LR↓	MS↓	CS↓	BLEU↓
Transform	Zoom	26.67	25.33	0.2562	0.1032
	Crop	30.46	26.91	0.2665	0.1193
	Resample	35.24	28.09	0.2734	0.1362
	Prompt	30.77	28.11	0.2615	0.1266
Adaptive	LoRA-InternVL	26.51	38.85	0.4027	0.1291
	LoRA-QwenVL	34.96	39.92	0.4310	0.1031
Purification	JPEG Q=75	24.29	26.96	0.2687	0.1051

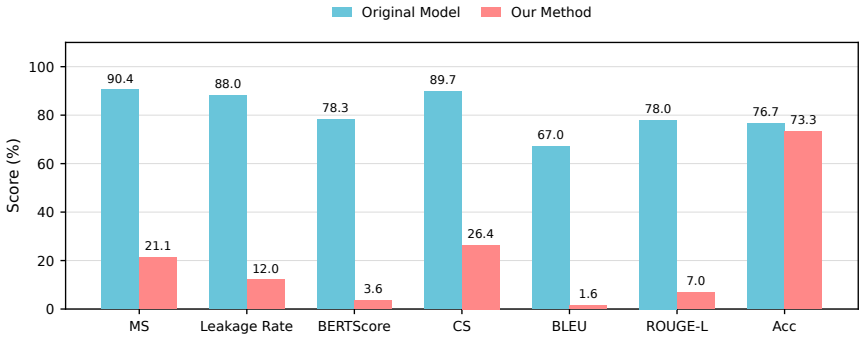

Fig. 6: Comparison of the performance of privacy protection in cross-app settings

4.6 Evaluation of stronger attacker strategies

To evaluate the robustness of **DualTAP** against adaptive attackers bypassing our protection, we analyze its performance under stronger extraction strategies. Table 6 shows **DualTAP** consistently suppresses these attempts, keeping quantitative extraction metrics securely low. Specifically, under geometric transformations, the exact LR is strictly contained between 26.67% (Zoom) and 35.24% (Resample). Furthermore, under prompt variations, extraction success remains heavily restricted, yielding an LR of 30.77% and a low ROUGE-L score of 0.1902. We further evaluate stronger adaptive settings, including LoRA fine-tuning of the attacker-side PII extractor and JPEG-based input purification. The results show only marginal leakage rebound and stable task accuracy, suggesting that **DualTAP** does not rely on fragile high-frequency artifacts but suppresses privacy-relevant visual-semantic cues. These results confirm that **DualTAP** maintains effectiveness against geometric transformations, prompt variations, adaptive extraction, and input purification, protecting PII from practical attacker strategies.

4.7 Cross-app generalization

As shown in Fig. 6, we further evaluate the generalization capability of DualTAP in cross-app scenarios. We split the dataset into 7 training apps and 3 completely zero-shot apps. The training phase exclusively uses samples from the training apps to learn the noise strategy, and the inference phase directly applies the trained generator to all interfaces of the remaining three apps. We adopted

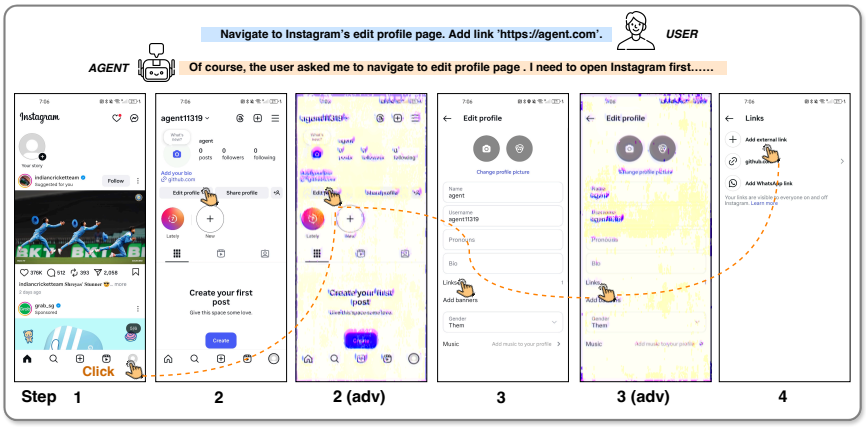

Fig. 7: The task execution process on Mobile-Agent-V3, shown as a full sequence of screenshots.

the same evaluation metrics as the main experiment. The experimental results show that the model, even on unseen apps, can still significantly suppress the privacy leakage rate, maintain a low field matching degree, and preserve the effectiveness of the normal task, thereby verifying the method's transfer potential in complex GUI scenarios. This indicates that the attention-guided noise strategy learned by the generator possesses strong layout or theme independence, laying the foundation for subsequent deployment on more real-world applications.

4.8 Real-World Case

We evaluated our approach through real-device task execution on a HarmonyOS smartphone using Mobile-Agent-V3 [41]. Across a range of representative interaction workflows, the agent successfully perceives the interface, issues actions, and completes tasks end-to-end with our privacy protection module enabled. As illustrated in Fig. 7, even after applying our protection mechanisms, the tasks continue to function normally, with no noticeable degradation in success rate or interaction quality. Moreover, our generator is lightweight and plug-and-play, which is approximately 300 MB, making it feasible for on-device deployment on commodity smartphones. Taken together, these findings indicate that our method can be deployed on real devices while preserving both privacy protection and the agent's practical usability.

5 Conclusion

In this work, we reveal and address a critical privacy leakage risk inherent in mobile MLLM agents: external API routers can exploit MLLMs to extract private information from user screenshots. Existing privacy protection methods fail

to balance task performance and privacy protection. To tackle this, we propose **DualTAP**, an adversarial protector that innovatively integrates a contrastive attention module focused on sensitive regions into the generator and, for the first time, introduces a dual-task adversarial objective, enabling explicit disentanglement of normal and privacy tasks. By jointly optimizing task-preservation and privacy-interference losses, **DualTAP** effectively suppresses PII exposure without compromising the agent's functionality. To further support this goal, we contribute PrivScreen, a multi-application mobile screenshot dataset for privacy and interface understanding. Experimental results across 6 MLLMs demonstrate the significant advantage of our method, substantially reducing privacy leakage while maintaining normal task execution. This work establishes a new paradigm for privacy protection in MLLM agents, shifting the research focus from traditional perception-level anomaly detection to adversarial disentanglement of task utility and privacy information at the inference level.

6 Acknowledgements

This research is supported by the RIE2025 Industry Alignment Fund –Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).This research is also supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award MOE-T2EP20125-0005).

References

1. Chen, J., Yuen, D., Xie, B., Yang, Y., Chen, G., Wu, Z., Yixing, L., Zhou, X., Liu, W., Wang, S., et al.: Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In: NeurIPS 2024 Workshop on Open-World Agents (2024)
2. Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., Goldstein, T.: Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. arXiv preprint arXiv:2101.07922 (2021)
3. Company, H.: Holo1.5 - open foundation models for computer use agents (2025), https://huggingface.co/collections/Hcompany/holo15-68c1a5736e8583a309d23d9b
4. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., Zhang, Y., Lv, W., Huang, K., Zhang, Y., Zhang, J., Zhang, J., Liu, Y., Yu, D., Ma, Y.: Paddleocr 3.0 technical report (2025), https://arxiv.org/abs/2507.05595
5. European Commission: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) (2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj
6. Guo, Z., Xu, B., Zhu, C., Hong, W., Wang, X., Mao, Z.: Mcp-agentbench: Evaluating real-world language agent performance with mcp-mediated tools. arXiv preprint arXiv:2509.09734 (2025)
7. Hu, Q.J., Bieker, J., Li, X., Jiang, N., Keigwin, B., Ranganath, G., Keutzer, K., Upadhyay, S.K.: Routerbench: A benchmark for multi-llm routing system. arXiv preprint arXiv:2403.12031 (2024)
8. Hu, X., Xiong, T., Yi, B., Wei, Z., Xiao, R., Chen, Y., Ye, J., Tao, M., Zhou, X., Zhao, Z., et al.: Os agents: A survey on mllm-based agents for general computing devices use. arXiv preprint arXiv:2508.04482 (2025)
9. Huang, H., Erfani, S., Li, Y., Ma, X., Bailey, J.: X-transfer attacks: Towards super transferable adversarial attacks on clip (2025), https://arxiv.org/abs/2505.05528
10. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
11. Jia, H., Liao, J., Zhang, X., Xu, H., Xie, T., Jiang, C., Yan, M., Liu, S., Ye, W., Huang, F.: Osworld-mcp: Benchmarking mcp tool invocation in computer-use agents (2025), https://arxiv.org/abs/2510.24563
12. Jia, X., Gao, S., Qin, S., Pang, T., Du, C., Huang, Y., Li, X., Li, Y., Li, B., Liu, Y.: Adversarial attacks against closed-source mllms via feature optimal alignment. arXiv preprint arXiv:2505.21494 (2025)
13. Kuang, Z., Sun, H., Li, Z., Yue, X., Lin, T.H., Chen, J., Wei, H., Zhu, Y., Gao, T., Zhang, W., Chen, K., Zhang, W., Lin, D.: Mmocr: A comprehensive toolbox for text detection, recognition and understanding (2021), https://arxiv.org/abs/2108.06543
14. Kulkarni, P., Cauvery, N.: Personally identifiable information (pii) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique. International Journal of Advanced Computer Science and Applications **12**(9) (2021)

15. Laishram, L., Shaheryar, M., Lee, J., Jung, S.: Toward a privacy-preserving face recognition system: A survey of leakages and solutions. ACM Computing Surveys **57**(6), 1–38 (2024). https://doi.org/10.1145/3673224
16. Legislature, C.S.: California consumer privacy act (ccpa). California Civil Code **1798100** (2018)
17. Lin, J., Zhu, C., Kneuertz, P.J., Bai, Y., Xue, Y.: When models learn to ask why: Adaptive causal reasoning for trustworthy medical vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5556–5568 (2026)
18. Mainetti, L., Elia, A.: Detecting personally identifiable information through natural language processing: A step forward. Applied System Innovation **8**(2), 55 (2025)
19. Meftah, H.F.Z.B., Hamidouche, W., Fezza, S.A., Déforges, O.: Vip: Visual information protection through adversarial attacks on vision-language models (2025), https://arxiv.org/abs/2507.08982
20. Nazeem, M., Anitha, R., Navaneeth, S., et al.: Open-source ocr libraries: A comprehensive study for low resource language. In: Proceedings of the 21st International Conference on Natural Language Processing (ICON). pp. 416–421 (2024)
21. Nguyen, D., Chen, J., Wang, Y., Wu, G., Park, N., Hu, Z., Lyu, H., Wu, J., Aponte, R., Xia, Y., et al.: Gui agents: A survey. arXiv preprint arXiv:2412.13501 (2024)
22. Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., et al.: Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326 (2025)
23. Shamshad, F., Naseer, M., Nandakumar, K.: Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20595–20605 (2023)
24. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)
25. Su, Z., Zhou, D., Wang, N., Liu, D., Wang, Z., Gao, X.: Hiding visual information via obfuscating adversarial perturbations. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4333–4343 (2023). https://doi.org/10.1109/ICCV51070.2023.00402
26. Sun, X., Li, H., Zhang, J., Yang, Y., Liu, K., Feng, R., Tan, W.J., Lim, W.Y.B.: Multipriv: Benchmarking individual-level privacy reasoning in vision-language models. arXiv preprint arXiv:2511.16940 (2025)
27. Tang, F., Xu, H., Zhang, H., Chen, S., Wu, X., Shen, Y., Zhang, W., Hou, G., Tan, Z., Yan, Y., et al.: A survey on (m) llm-based gui agents. arXiv preprint arXiv:2504.13865 (2025)
28. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
29. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
30. Wang, J., Xu, H., Ye, J., Yan, M., Shen, W., Zhang, J., Huang, F., Sang, J.: Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. arXiv preprint arXiv:2401.16158 (2024)
31. Wang, L., Deng, Y., Zha, Y., Mao, G., Wang, Q., Min, T., Chen, W., Chen, S.: Mobileagentbench: An efficient and user-friendly benchmark for mobile llm agents. arXiv preprint arXiv:2406.08184 (2024)

32. Wang, S., Liu, W., Chen, J., Zhou, Y., Gan, W., Zeng, X., Che, Y., Yu, S., Hao, X., Shao, K., et al.: Gui agents with foundation models: A comprehensive survey. arXiv preprint arXiv:2411.04890 (2024)
33. Wang, X., Bloch, J., Shao, Z., Hu, Y., Zhou, S., Gong, N.Z.: Webinject: Prompt injection attack to web agents (2025), https://arxiv.org/abs/2505.11717
34. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)
35. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
36. Wu, H., Ou, G., Wu, W., Zheng, Z.: Improving transferable targeted adversarial attacks with model self-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24615–24624 (2024)
37. Wu, J., Zhong, M., Xing, S., Lai, Z., Liu, Z., Chen, Z., Wang, W., Zhu, X., Lu, L., Lu, T., et al.: Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. Advances in Neural Information Processing Systems **37**, 69925–69975 (2024)
38. Yan, Y., Wang, S., Du, J., Yang, Y., Shan, Y., Qiu, Q., Jia, X., Wang, X., Yuan, X., Han, X., et al.: Mcpworld: A unified benchmarking testbed for api, gui, and hybrid computer use agents. arXiv preprint arXiv:2506.07672 (2025)
39. Yang, X., Chen, J., Luo, J., Fang, Z., Dong, Y., Su, H., Zhu, J.: Mla-trust: Benchmarking trustworthiness of multimodal llm agents in gui environments. arXiv preprint arXiv:2506.01616 (2025)
40. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Chen, C., Li, H., Zhao, W., et al.: Efficient gpt-4v level multimodal large language model for deployment on edge devices. Nature Communications **16**(1), 5509 (2025)
41. Ye, J., Zhang, X., Xu, H., Liu, H., Wang, J., Zhu, Z., Zheng, Z., Gao, F., Cao, J., Lu, Z., et al.: Mobile-agent-v3: Fundamental agents for gui automation. arXiv preprint arXiv:2508.15144 (2025)
42. Yue, Y., Zhang, G., Liu, B., Wan, G., Wang, K., Cheng, D., Qi, Y.: Masrouter: Learning to route llms for multi-agent systems. arXiv preprint arXiv:2502.11133 (2025)
43. Zhang, D., Rama, B., Ni, J., He, S., Zhao, F., Chen, K., Chen, A., Cao, J.: Litewebagent: The open-source suite for vlm-based web-agent applications. arXiv preprint arXiv:2503.02950 (2025)
44. Zhang, J., Ye, J., Ma, X., Li, Y., Yang, Y., Chen, Y., Sang, J., Yeung, D.Y.: Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19900–19909 (2025)
45. Zhang, J., Yi, Q., Sang, J.: Towards adversarial attack on vision-language pretraining models. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5005–5013 (2022)
46. Zhao, R., Zhang, Y., Wang, T., Wen, W., Xiang, Y., Cao, X.: Visual content privacy protection: A survey. ACM Computing Surveys **57**(5), 1–36 (2025)
47. Zhong, Y., Deng, W.: Opom: Customized invisible cloak towards face privacy protection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3590–3603 (2022)
48. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)