Focus Like a Human: Efficient GUI Grounding via Coarse-to-Fine Visual Attention and Parallel Verification

Zhenhua Yang[1] , **Xiachong Feng**[2] , **Weihong Zhong**[1] , **Lunjun Liu**[1] , **Xiaocheng Feng**[1*] , **Bing Qin**[1*]
[1]Harbin Institute of Technology
[2]University of Hong Kong
{zhyang, whzhong, ljliu, xcfeng, qinb}@ir.hit.edu.cn, fengxc@hku.hk

Abstract

Building upon powerful Large Visual Language Models, recent GUI agents have revolutionized autonomous GUI interaction. Given the high information density and structural complexity of GUI layouts, a critical challenge lies in accurately identifying where to focus, i.e., precise GUI grounding. To ensure grounding accuracy, prior approaches predominantly rely on processing inputs at native high resolutions. However, this paradigm inevitably introduces excessive visual redundancy, resulting in increased computational cost and dispersed attention. In our preliminary analysis, we observe that while high-resolution inputs are indispensable for recognizing fine-grained visual details, they paradoxically impair the model's ability to attend to the correct regions. In contrast, low-resolution inputs, which suppress high-frequency details while preserving global layout structure through downsampling, produce significantly more robust and coherent attention patterns that better guide the model toward relevant regions. Interestingly, this observation mirrors the human visual system, which first leverages low-resolution peripheral vision to identify salient areas before conducting a detailed examination via high-resolution foveal focus. Motivated by this insight, we propose a coarse-to-fine GUI grounding framework termed **FastFocus**. Our approach first exploits the stability of low-resolution attention to efficiently generate candidate grounding regions. These candidates are then evaluated by a Parallel Verification module, which selectively zooms into high-resolution views to resolve fine details and filter out false positives. Experiments on ScreenSpot-Pro demonstrate state-of-the-art performance, validating that a hierarchical integration of low-resolution guidance and high-resolution verification constitutes an effective and robust paradigm for GUI grounding.

1 Introduction

Large Visual Language Models enable recent agents to interact with Graphical User Interfaces (GUIs) in a human-like and intuitive manner [Zheng *et al.*, 2024; Cheng *et al.*, 2024; Gou *et al.*, 2025; Wu *et al.*, 2025c; Zhang *et al.*, 2025b; Bai *et al.*, 2025b]. While these agents have demonstrated impressive performance on synthetic or low-resolution benchmarks [Xu *et al.*, 2025a; Toyama *et al.*, 2021; Zhang *et al.*, 2024], operating real-world, information-dense interfaces remains a challenge. GUIs of professional software are often populated with a large number of visually or functionally similar elements, posing a critical bottleneck to *GUI grounding*—the ability to accurately locate functional elements corresponding to natural language instructions [Li *et al.*, 2025; Qian *et al.*, 2024]. To accommodate diverse input resolutions and capture fine-grained details in these scenarios, prior works have proposed various solutions, including resizing to fixed scales, grid-based slicing [Bai *et al.*, 2023; Li *et al.*, 2024; Vasu *et al.*, 2025; Wang *et al.*, 2024b]. Increasingly, the prevailing paradigm has shifted toward processing inputs at native high resolutions, scaling to 4K and beyond. While this strategy benefits visual fidelity, this paradigm incurs excessive visual redundancy, leading to higher computational cost and attention dispersion. As shown in Figure 1, we observe that excessive resolution inputs cause attention dispersion, biasing the model towards trivial pixel-level variations and superficial details instead of the global spatial layout. Consequently, the model lacks the context to anchor its predictions, resulting in coordinate instability and ungrounded hallucinations, where bounding boxes appear in arbitrary locations detached from any actual UI elements. We term this phenomenon *Semantic Fragmentation*, where the excessive density of visual tokens dilutes the model's global receptive field, compromising its representation of the global semantic layout as it fixates on local pixel-level details. In contrast, we identify that low-resolution inputs function as effective *Semantic Aggregators*. By reducing visual token redundancy, downsampling serves as a form of structural abstraction, forcing the model to bypass superficial details and instead align with the global spatial layout.

This computational insight finds a profound theoretical parallel in cognitive neuroscience, specifically the coarse-to-

*Co-corresponding authors.

Figure 1: The Resolution-Attention Conflict in GUI Grounding. Contrary to the intuition that "higher is better", processing GUIs at native high resolution (Right) often leads to *Semantic Fragmentation*, where visual attention is distracted by details (e.g., text strokes, edges), losing the object's center. In contrast, FastFocus (Middle) leverages low-resolution inputs as a "Semantic Aggregator", functioning as a filter to form a cohesive, robust attention prior for precise localization.

fine processing dynamics of the human visual system [Schyns and Oliva, 1994; Hochstein and Ahissar, 2002]. To manage the immense bandwidth of visual information [Carrasco, 2011], the biological system employs an efficient division of labor: Peripheral Vision operates at low resolution to rapidly extract the global spatial layout and structural gist [Larson and Loschky, 2009], while Foveal Vision is serially deployed for the high-acuity identification of specific details [Bar, 2004]. Our finding that low-resolution tokens effectively encode global semantics aligns perfectly with the functional role of peripheral vision. Driven by this biological plausibility, we propose **FastFocus**, a coarse-to-fine framework that mimics this dual-process strategy of human visual system. Specifically, FastFocus comprises two synergistic modules: (1) Global Proposal Module: leveraging downsampled whole-image views to act as a high-recall semantic aggregator, filtering out background noise to identify potential regions of interest; and (2) Local Verification Module: processing these candidate regions at native resolution, utilizing high-fidelity visual features to perform fine-grained verification and precise localization. FastFocus allows the model to concentrate on fine-grained details within relevant semantic regions, effectively decoupling the semantic understanding from pixel-level localization. Extensive experiments on the ScreenSpot-Pro benchmark demonstrate that FastFocus establishes a new state-of-the-art GUI grounding accuracy. Analysis confirms that our method maintains robust focus in visually cluttered environments, effectively resolving the trade-off between grounding precision and inference efficiency. Our main contributions are summarized as follows:

- We identify "Semantic Fragmentation" as a primary failure mode in high-resolution GUI grounding and empirically demonstrate that low-resolution perception serves as a superior cognitive prior by effectively aggregating semantic features.
- We propose FastFocus, a biomimetic framework that decouples grounding into attention-based proposal generation and parallel saccadic verification. This design effectively breaks the serial latency bottleneck while mitigating the visual hallucinations common in direct regression paradigms.
- We validate FastFocus on the challenging ScreenSpot-Pro benchmark, where it achieves state-of-the-art and demonstrates exceptional robustness in high-density professional GUIs.

2 Related Work

2.1 GUI Agent

The development of GUI agents has transitioned from relying on structured metadata to leveraging direct visual perception. Early approaches predominantly utilized text-based reasoning, parsing HTML, DOM trees, or accessibility XMLs

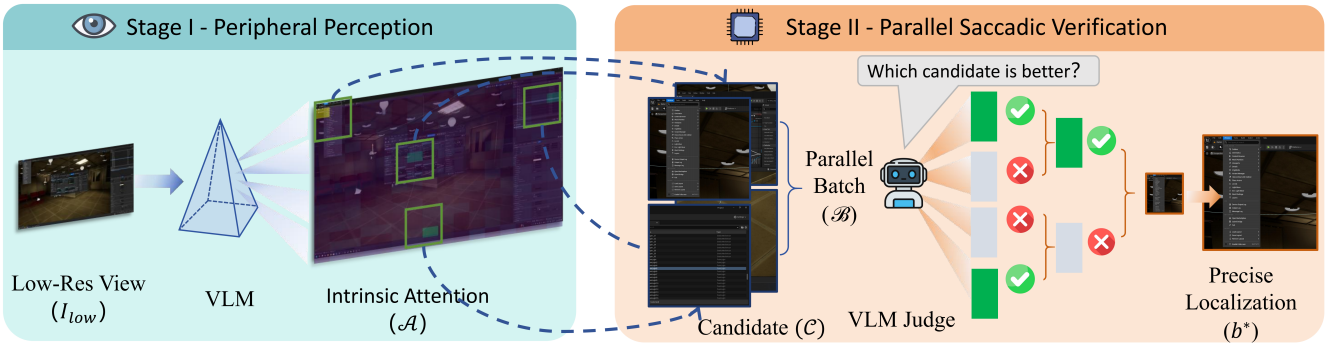

Figure 2: The biomimetic framework of FastFocus. The pipeline decouples grounding into two synergistic stages: (I) Global Proposal Module (Left): A low-resolution view (The longest side does not exceed 448px) drives the VLM to generate an intrinsic attention map. To capture diverse regions, we employ iterative peak identification with spatial suppression, sequentially selecting salient centers while masking visited areas to generate high-recall proposals. (II) Recursive Pairwise Verification (Right): High-resolution crops are organized into a parallel tournament structure. Instead of independent scoring, the VLM performs discriminative pairwise selection, recursively eliminating weaker matches to robustly isolate the target from hard negatives.

to interact with Web and mobile environments [Zhang *et al.*, 2025b; Wang *et al.*, 2024a; Yan *et al.*, 2023]. While efficient for structured data, these methods struggle with "visual-centric" applications where accessibility metadata is often missing or incomplete. Recent advances in Multimodal Large Language Models (MLLMs) have enabled a shift towards vision-based agents that perceive raw pixels directly [Liu *et al.*, 2023; Chen *et al.*, 2024; Wang *et al.*, 2024b; Bai *et al.*, 2025b]. By bridging the visual perception gap, these agents can operate in open-ended desktop environments, achieving end-to-end visual grounding and action prediction [Cheng *et al.*, 2024; Qin *et al.*, 2025; Hong *et al.*, 2024; Gou *et al.*, 2025; Xu *et al.*, 2025b]. However, capturing fine-grained details in dense interfaces remains challenging. To mitigate these issues, recent works have introduced domain-specific pre-training and modality-specific finetuning [Wu *et al.*, 2025c; Sun *et al.*, 2024], as well as reinforcement learning approaches to improve agent performance [Lu *et al.*, 2025; Zhou *et al.*, 2025; Liu *et al.*, 2025]. Nevertheless, these methods often require extensive training data and computational resources. Unlike prior works that attempt to optimize the backbone itself, our framework focuses on a generic perception strategy to resolve the conflict between high-resolution accuracy and inference latency.

2.2 GUI Grounding

Precise element localization is the fundamental capability underpinning vision-based agents [Qian *et al.*, 2024]. Existing methodologies can be broadly categorized into three paradigms:

Supervised Regression. The dominant paradigm involves fine-tuning VLMs on large-scale datasets to directly regress bounding box coordinates [Cheng *et al.*, 2024; Wu *et al.*, 2025c; Gou *et al.*, 2025]. Models like SeeClick and OS-Atlas [Wu *et al.*, 2025c] have demonstrated strong performance by learning from massive interaction data. However, these regression-based methods often suffer from hallucinations in visually cluttered regions and lack interpretability. Furthermore, their performance is heavily dependent on the scale and quality of annotated data.

Visual Test-Time Scaling (TTS). To handle high-resolution details without extensive retraining, recent works have introduced test-time visual refinement strategies. Methods such as RegionFocus and Iterative Narrowing adopt a "crop-and-zoom" mechanism, iteratively updating the model's visual focus based on history or initial predictions. While effective, these methods inherently suffer from a "Serial Accumulation Bottleneck". Since each refinement step requires a sequential forward pass, the inference latency scales linearly with the number of refinement iterations. In contrast, FastFocus employs a parallel verification mechanism, assessing multiple candidate regions simultaneously to achieve fine-grained precision without the latency penalty of serial interactions.

Visual Attention and Priors. Attention mechanisms have long been fundamental to visual understanding. Recent methods such as TAG [Xu *et al.*, 2025a] and GUI-Actor [Wu *et al.*, 2025b] avoid coordinate regression by leveraging internal attention signals from VLMs. However, they typically utilize attention primarily for coarse localization without analyzing failure causes under challenging conditions. Diverging from the convention of operating on high-resolution features—which we identify as susceptible to "Semantic Fragmentation"—we revisit the potential of low-resolution attention. We demonstrate that low-resolution inputs serve as a superior "Semantic Aggregator," providing a robust, training-free prior for proposal generation that aligns better with object-level layouts than high-resolution counterparts.

3 Methodology

In this section, we present FastFocus, a biomimetic coarse-to-fine framework designed for efficient and precise GUI grounding. Unlike prevailing methods that rely on computationally expensive native-resolution processing, FastFocus mimics the efficient attention mechanism of the human eye.

The overall pipeline is illustrated in Figure 2.

3.1 Biological Plausibility

The architectural design of FastFocus is theoretically rooted in the *Reverse Hierarchy Theory* of visual perception [Hochstein and Ahissar, 2002]. Neuroscientific evidence indicates that the human visual system avoids processing the entire visual field with uniform high fidelity, a strategy that would otherwise overwhelm the optic nerve's transmission bandwidth [Carrasco, 2011]. Instead, it employs a resource-efficient dual-process strategy:

Peripheral Processing as Semantic Aggregation: The human peripheral visual field, despite its low acuity, is highly sensitive to spatial layout and global statistics (the "gist" of the scene) [Oliva and Torralba, 2006]. It acts as a rapid scanning mechanism that filters out high-frequency noise to identify potential regions of interest [Larson and Loschky, 2009]. In our framework, this maps to the Global Proposal Module, where low-resolution visual tokens serve as semantic aggregators to stabilize localization anchors.

Foveal Processing as Detail Verification: Once a salient region is detected, the eyes perform a saccade to bring that region into the fovea, which processes high-frequency details for precise identification [Bar, 2004]. This biological saccade parallels our Local Verification Module, which crops and processes high-resolution patches only where necessary.

Driven by this biological plausibility, we operationalize the Reverse Hierarchy Theory into the FastFocus framework. Specifically, we decouple the grounding process into two synergistic stages that mirror the human visual system: a Global Proposal Module acts as the peripheral processing to rapidly identify potential regions of interest from low-resolution inputs; and a Parallel Local Verification Module functions as the foveal processing, executing high-resolution saccades to resolve fine-grained details.

3.2 Global Proposal Module

Problem Formulation. The primary objective of this stage is to serve as a Coarse-Grained Proposal Generator. Formally, given an image $I_{native} \in \mathbb{R}^{H\times W\times 3}$ and a natural language instruction T, our goal is to generate a set of candidate regions $\mathcal{P} = \{b_1, b_2, \ldots, b_K\}$, where each b_k represents a potential target location. We prioritize recall over precision here, aiming to capture the target within $\mathcal{P}$ with a low computational cost.

Intrinsic Attention Extraction. We leverage the intrinsic cross-attention mechanism of the VLM as a spatial prior. As shown by Zhang *et al.* and Xu *et al.*, the attention weights in intermediate transformer layers often exhibit a strong correlation with the model's focus on relevant image regions. Building on this insight, we resize I_{native} to a lower resolution I_{low} and input it into the VLM. We extract the attention maps averaged across the intermediate layers (layers l_{start} to l_{end}). Let $\mathbf{A} \in \mathbb{R}^{h\times w}$ denote the aggregated attention map, where each scalar $\mathbf{A}_{i,j}$ approximates the semantic relevance of patch (i, j) to the instruction T:

$$\mathbf{A} = \frac{1}{|L|}\sum_{l\in L} \text{AttnMap}(I_{low}, T; \theta_l) \tag{1}$$

Algorithm 1 FastFocus Inference Process

Require: Native Image I_{native}, Instruction T, Number of Proposals K
Ensure: Target Bounding Box b^*

Stage I: Global Proposal Module
$I_{low} \leftarrow \text{Resize}(I_{native}, 448)$
$\mathbf{A} \leftarrow \text{ExtractIntrinsicMap}(I_{low}, T)$
$\mathcal{P} \leftarrow \emptyset$
for $k \leftarrow 1$ **to** K **do**
  $c_k \leftarrow \arg\max_{(i,j)} \mathbf{A}_{i,j}$ *// Peak identification*
  $\mathcal{P}.\text{append}(\text{GetProposalBox}(c_k))$
  $\mathbf{A} \leftarrow \mathbf{A} \odot (1 - \text{SpatialMask}(c_k))$ *// Spatial suppression*
end for
Stage II: Parallel Local Verification
$\mathcal{C} \leftarrow \text{CropHighFidelity}(I_{native}, \mathcal{P})$
while $|\mathcal{C}| > 1$ **do**
  $\mathcal{B}_{pairs} \leftarrow \text{FormPairs}(\mathcal{C})$
  $\mathcal{W}_{indices} \leftarrow \text{VLM}_{\text{compare}}(\mathcal{B}_{pairs}, T_{pair})$ *// Parallel tournament round*
  $\mathcal{C} \leftarrow \text{SelectWinners}(\mathcal{C}, \mathcal{W}_{indices})$ *// Filter survivors for next round*
end while
$C_{winner} \leftarrow \mathcal{C}[0]$
Stage III: Precise Localization
$b_{local} \leftarrow \text{VLM}_{\text{loc}}(C_{winner}, T)$ *// Regression on optimal patch*
$b^* \leftarrow \text{MapToGlobal}(b_{local}, C_{winner})$
return b^*

where L is the set of selected intermediate layers. This heatmap $\mathbf{A}$ serves as a dense probability distribution of the target's existence.

Greedy Proposal Generation. To select Region of Interest candidates from the attention map $\mathbf{A}$, we design an iterative peak extraction strategy with spatial suppression. This ensures the model captures diverse salient regions without redundancy. We initialize the proposal set $\mathcal{P} = \emptyset$. For $k = 1$ to K:

Peak Identification. We locate the global maximum in the current heatmap as the proposal center c_k:

$$c_k = \arg\max_{(i,j)} \mathbf{A}^{(k)}_{i,j} \tag{2}$$

Proposal Extraction. We generate a bounding box b_k centered at c_k with a predefined anchor scale and project it back to the coordinate space of I_{native}.

Spatial Suppression. To prevent repeated selection of the same region, we apply a hard spatial mask around c_k. The heatmap for the next iteration is updated as:

$$\mathbf{A}^{(k+1)}_{i,j} = \mathbf{A}^{(k)}_{i,j} \cdot \mathbb{I}\left(\|(i, j) - c_k\|_2 > r\right) \tag{3}$$

where r is the suppression radius and $\mathbb{I}(\cdot)$ is the indicator function, which returns 0 if the coordinate lies within the radius and 1 otherwise.

This greedy process efficiently yields a ranked list of candidates $\mathcal{P}$ based on intrinsic semantic saliency, without requiring additional training or external bounding box supervision.

Model	Type	Development		Creative		CAD		Scientific		Office		OS		Avg.
		Text	Icon	Text	Icon	Text	Icon	Text	Icon	Text	Icon	Text	Icon	
Generalist & Proprietary Models														
GPT-4o	Closed	1.3	0.0	1.0	0.0	2.0	0.0	2.1	0.0	1.1	0.0	0.0	0.0	0.8
Claude Computer Use	Closed	22.0	3.9	25.9	3.4	14.5	3.7	33.9	15.8	30.1	16.3	11.0	4.5	17.1
Qwen2.5-VL-7B	Open	45.5	1.4	32.8	6.3	22.3	6.2	50.7	7.3	52.5	15.1	36.4	10.1	26.8
Qwen3VL-8B-Instruct*	Open	80.0	26.9	69.2	16.1	46.7	16.7	73.6	34.5	77.4	32.1	66.4	25.8	51.4
Specialist GUI Agents (SFT / RL Baselines)														
UGround-V1-7B	SFT	51.9	2.8	47.5	9.7	15.8	1.2	57.6	14.5	60.5	13.2	38.3	7.9	31.1
UI-TARS-1.5-7B	SFT	50.0	14.5	56.6	13.3	37.6	12.5	66.0	22.7	76.3	34.0	55.6	16.9	40.8
UI-Venus-Ground-7B	RL	74.7	24.1	63.1	14.7	**60.4**	21.9	76.4	31.8	75.7	41.5	49.5	22.5	50.8
Test-Time Scaling (TTS) Baselines														
RegionFocus-7B	TTS	59.7	15.9	59.6	11.9	30.5	7.8	67.4	30.0	69.5	30.2	45.8	21.3	41.2
DiMo-GUI-7B	TTS	66.9	21.4	60.6	21.7	50.3	14.1	68.1	21.8	80.8	52.8	69.2	28.1	49.7
BAMI-7B	TTS+RL	**80.5**	26.9	66.7	21.0	53.8	28.1	78.5	34.6	83.1	62.3	71.3	31.6	56.2
Ours (FastFocus Scaling)														
UGround-V1-7B	Base	51.9	2.8	47.5	9.7	15.8	1.2	57.6	14.5	60.5	13.2	38.3	7.9	31.1
+ **FastFocus**	Ours	58.2	15.1	53.3	22.6	28.1	14.6	65.3	25.5	67.2	24.5	45.8	21.3	40.8
	Improvement Δ	*+6.3*	*+12.3*	*+5.8*	*+12.9*	*+12.3*	*+13.4*	*+7.7*	*+11.0*	*+6.7*	*+11.3*	*+7.5*	*+13.4*	*+9.7*
UI-TARS-1.5-7B	Base	50.0	14.5	56.6	13.3	37.6	12.5	66.0	22.7	76.3	34.0	55.6	16.9	40.8
+ **FastFocus**	Ours	65.3	30.3	65.7	28.5	51.4	27.1	75.0	34.5	**84.7**	52.8	66.4	33.7	53.1
	Improvement Δ	*+15.3*	*+15.8*	*+9.1*	*+15.2*	*+13.8*	*+14.6*	*+9.0*	*+11.8*	*+8.4*	*+18.8*	*+10.8*	*+16.8*	*+12.3*
UI-Venus-Ground-7B	Base	74.7	24.1	63.1	14.7	60.4	21.9	76.4	31.8	75.7	41.5	49.5	22.5	50.8
+ **FastFocus**	Ours	78.2	**37.8**	**72.2**	**33.6**	60.0	**39.6**	**81.3**	**45.5**	80.8	**69.8**	**74.8**	**44.9**	**61.8**
	Improvement Δ	*+3.5*	*+13.7*	*+9.1*	*+18.9*	*-0.4*	*+17.7*	*+4.9*	*+13.7*	*+5.1*	*+28.3*	*+25.3*	*+22.4*	*+11.0*

Table 1: Main Results on ScreenSpot-Pro Benchmark. We evaluate FastFocus as a plug-and-play module across various base models. Red indicates performance gains, while blue indicates marginal performance trade-offs. **Bold** indicates the best performance in each category. * indicates the model is reproduced.

3.3 Parallel Local Verification Module

Problem Formulation. The objective of this stage is to identify the precise target from the coarse candidates generated in the Global Proposal Module. Formally, taking the proposal set $\mathcal{P}$ and the native high-resolution image I_{native} as input, the module functions as a fine-grained discriminator to determine the optimal target b^* and refine its coordinates.

High-Fidelity Cropping. We first extract high-resolution visual details to recover the information lost during the downsampling in Stage I. For each candidate bounding box $b_k \in \mathcal{P}$, we crop the corresponding region from the native image I_{native} to obtain a set of high-fidelity patches $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$. These patches preserve the original pixel density required for reading small text and distinguishing fine icon details.

Recursive Pairwise Verification. To robustly identify the target among these patches, we adopt a Recursive Pairwise Verification strategy inspired by [Zhang *et al.*, 2025a]. We formulate the selection process as a multi-round tournament rather than independent scoring, which forces the model to discriminate the target from distractors based on relative visual evidence.

In each round t, we pair the current candidates $\{C_i, C_j\}$ and construct a comparison prompt T_{pair}. These pairs are stacked into a single tensor batch to be processed in parallel:

$$\mathbf{B}^{(t)} = \text{Stack}\left(\{(C_i, C_j) \mid i, j \in \text{Indices}^{(t)}\}\right) \quad (4)$$

The VLM predicts the winner for each pair in a single forward pass. The winners form the candidate set for the next round, and this process recursively continues until a single optimal patch C_{k^*} remains. This mechanism effectively suppresses hard negatives by leveraging the model's stronger capability in relative comparison versus absolute verification.

Precise Localization. Upon determining the winning candidate k^*, the model performs a final coordinate regression on the patch C_{k^*}. The predicted local coordinates are then projected back to the global coordinate system of I_{native} to yield the final prediction b^*.

4 Experiments

4.1 Experimental Setup

Datasets and baselines. Our primary evaluation utilizes ScreenSpot-Pro [Li *et al.*, 2025], a benchmark comprised of high-resolution screenshots (larger than 1920×1080) from professional desktop applications (e.g., Blender, VS Code), which poses significant challenges for current agents due to element density and layout complexity. To assess generalization across platforms, we also report results on the ScreenSpot V2 benchmark. Baselines. We compare our method against three classes of models: (1) Generalist VLMs including GPT-4o, Claude Computer Use, Qwen2.5-VL and Qwen3-VL [OpenAI, 2024; Bai *et al.*, 2025b; Bai *et al.*, 2025a]; (2) Specialist Agents such as SeeClick [Cheng *et al.*, 2024], OS-Atlas [Wu *et al.*, 2025c], UGround [Gou *et al.*, 2025] and UI-Venus [Gu *et al.*, 2025], which leverage Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) to align visual perception with actionable intent. These models typically undergo domain-specific training on large-scale

Model	Mob		Desk		Web		Avg.
	Tx	Ic	Tx	Ic	Tx	Ic	
Generalist VLMs							
Qwen2-VL-7B	39.3	61.3	52.0	45.0	33.0	21.8	42.9
CogAgent-18B	24.0	67.0	74.2	20.0	70.4	28.6	47.4
InternVL-2-4B	4.8	9.2	4.6	4.3	0.9	0.1	4.3
Specialist GUI Agents							
SeeClick	78.0	52.0	30.0	72.2	55.7	32.5	53.4
OS-Atlas-7B	68.7	92.1	88.7	60.7	89.7	75.9	81.2
UGround-7B	82.8	82.8	82.8	63.6	80.4	70.4	73.3
UI-TARS-1.5	80.6	94.1	88.7	76.4	88.0	84.2	86.4
Test-Time Scaling (TTS)							
BAMI-7B	94.1	80.6	88.7	76.4	88.0	84.7	86.5
DiMo-GUI-7B	94.8	85.3	94.3	82.1	93.2	80.3	89.2
Ours (FastFocus Scaling)							
UI-Venus-Ground-7B (Base)	99.0	90.0	97.0	90.7	96.2	88.7	94.1
+ FastFocus (Ours)	**98.9**	**91.3**	**96.4**	**92.1**	**96.5**	**90.3**	**94.6**
Improvement Δ	*-0.1*	*+1.3*	*-0.6*	*+1.4*	*+0.3*	*+1.6*	*+0.5*

Table 2: Results on ScreenSpot V2. We verify our FastFocus module on the strong UI-Venus base. Red denotes improvement, while blue indicates trade-offs. FastFocus improves precision on hard Icon tasks while maintaining competitive performance on easier Text tasks. Mob, Desk, and Web denote the mobile, desktop, and web subsets, respectively; Tx and Ic denote text and icon grounding; Avg. denotes the overall average.

interaction datasets to achieve precise element localization and robust task execution; and (3) Test-Time Scaling Methods like RegionFocus [Luo *et al.*, 2025], DiMo-GUI [Wu *et al.*, 2025a], and BAMI [Zhang *et al.*, 2025a], which represent the current state-of-the-art in Test-Time visual refinement.

Implementation Details. We utilize UI-Venus-7B [Gu *et al.*, 2025], UGround-V1-7B, and UI-TARS-1.5-7B as the backbone models. For the Global Proposal Module, input images are resized to 448×448. We extract the intrinsic semantic attention by averaging the cross-attention weights from layers 12 to 28 of the vision encoder. To generate diverse region proposals, we employ an iterative greedy sampling strategy with the spatial suppression radius set to $0.15\times$ the image size, selecting the top $K = 4$ candidates. In the Parallel Local Verification Module, we extract high-resolution crops with a scale of $0.3\times$ the original image dimensions, enforcing a minimum shortest edge of 448px, to preserve sufficient local context. These candidates are processed in a single parallel batch to ensure efficient verification.

4.2 Main Results

As shown in Table 1, we present a comprehensive evaluation on the ScreenSpot-Pro benchmark. Unlike static baselines, we assess FastFocus as a model-agnostic perception module applied to three distinct base agents: UGround-V1, UI-TARS-1.5, and UI-Venus-Ground. The results demonstrate consistent performance gains across all backbones, indicating that FastFocus is not backbone-specific and transfers effectively across diverse agents. Notably, applying FastFocus to UI-Venus-Ground-7B achieves a new state-of-the-art average accuracy of 61.8%, surpassing the previous best Test-Time Scaling method, BAMI-7B (56.2%), by a significant margin. This confirms that our dual-stage strategy effectively complements existing SFT and RL-based optimization schemes without requiring model-specific tuning.

A granular analysis reveals a distinct performance divergence: while text grounding sees moderate gains, the improvement in Icon and Widget subsets is disproportionately significant (reaching up to +28.3% in Office/Icon for UI-Venus). We attribute this to the differing perception granularity required. Text relies on explicit features, whereas icons depend heavily on regional layout semantics—the understanding of an element's role within its surrounding functional block (e.g., a toolbar or navigation panel). Standard high-resolution processing often suffers from attention dispersion, failing to grasp this coarse-grained block structure as it fixates on high-frequency pixel details. Our framework resolves this via the Global Proposal Module, which utilizes the low-resolution prior to aggregate visual information into coherent semantic regions. Subsequently, the Recursive Pairwise Verification in Stage II discriminatively refines these regional proposals, ensuring the final localization aligns with the intended functional module.

To ensure that our optimization for high-density professional GUIs does not compromise performance in standard scenarios, we further evaluated FastFocus on the general ScreenSpot V2 benchmark (Table 2). On this relatively simpler dataset, where the UI-Venus baseline already approaches performance saturation (94.1%), our method maintains robust accuracy (94.6%). Notably, we observe a consistent trend where Icon grounding continues to improve (+1.3% to +1.6%), validating that our *Global Proposal Module* effectively generalizes across varying levels of layout complexity without introducing regression in simpler tasks.

4.3 Analysis of Resolution Priors

To quantitatively validate our qualitative observation regarding the conflict between resolution and attention in Section 1, we evaluated the Proposal Recall@4 of the UI-Venus-Ground-7B model across input image edge scales ranging from 224 to 4096. As illustrated in Figure 3, the results exhibit a distinct Inverted-U trajectory, peaking at 448. This empirical evidence strongly corroborates the inherent conflict between resolution and attention cohesion:

- Validation of Semantic Fragmentation: Performance degrades significantly at native high resolutions. This confirms that while high resolution is necessary for details, excessive visual token density leads to attention dispersion, preventing the model from effectively locking onto the object's center during the proposal phase.
- Optimal Structural Abstraction: The peak at 448 validates that moderate downsampling serves as an effective Structural Abstractor. It suppresses high-frequency noise while preserving the global layout, verifying our observation that lower resolutions yield more cohesive attention patterns for initial proposal generation.

4.4 Ablation Analysis

To validate the rationale behind our architectural decisions, we conducted a component-wise ablation study on the

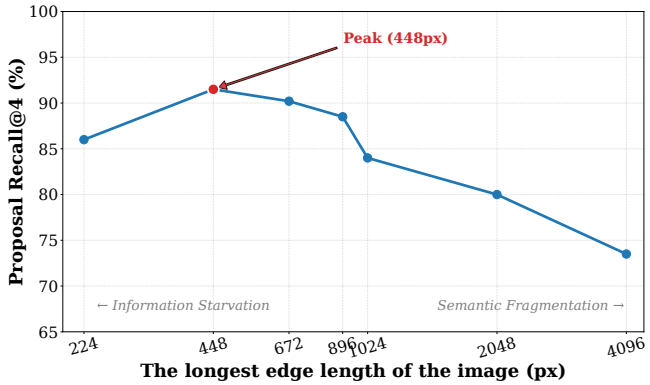

Figure 3: The Resolution-Attention Conflict. We evaluate the proposal performance of UI-Venus-Ground-7B across varying input scales. The x-axis represents the input resolution limit (max side length, downsampling only), and the y-axis (Recall@4) measures the proportion of ground-truth targets successfully covered by the model's intrinsic attention proposals. The inverted-U trajectory peaks at 448px, empirically validating that moderate downsampling acts as an optimal *Structural Abstractor*, whereas native high resolutions induce *Semantic Fragmentation*.

ScreenSpot-Pro benchmark, as detailed in Table 3.

Impact of Proposal Strategy (Rows a-c). We first investigate the optimal strategy for generating object proposals.

- **Intrinsic Prior vs. External Detectors:** Utilizing an external object detector (Grounding DINO, Row a) yields suboptimal performance (41.2%). This confirms the existence of a Semantic Gap: generic detectors are agnostic to user intent, generating geometrically valid but semantically irrelevant proposals. In contrast, our VLM-based intrinsic attention naturally aligns visual features with linguistic instructions.
- **Explicit Attention vs. Direct Regression:** While standard Direct Regression (Row b) and Native Resolution processing (Row c) achieve moderate baselines ($\sim$54%), they hit a performance ceiling. We attribute this to the **Resolution-Attention Conflict** discussed in Sec. 1: native resolution processing forces the model to handle excessive token density, leading to attention dispersion. Our Coarse-to-Fine approach (Row e, 61.8%) outperforms these single-stage baselines by decoupling structural localization from fine-grained identification.

Necessity of Coarse-to-Fine Synergy (Rows d-e). The most critical insight stems from the interaction between Stage I and Stage II.

- **The Power of Verification:** Using the Global Proposal Module alone (Row d, w/o Verification) achieves 48.5% accuracy. This configuration corresponds to setting $K = 1$, where the model directly performs coordinate regression on the top-ranked proposal derived from the intrinsic attention map, bypassing the pairwise comparison entirely. While this is lower than the full model, it serves as a high-recall **Structural Abstractor**, effectively narrowing the search space from millions of pixels to a few salient regions. Integrating the Parallel Verification Module (Row e) yields a massive **+13.3%** performance jump (48.5% $\rightarrow$ 61.8%). This empirical evidence validates our core hypothesis: low-resolution priors are necessary to capture global layout semantics, but insufficient for precise localization. The Recursive Pairwise Verification in Stage II is indispensable for resolving the visual ambiguity among the candidate crops, providing the discriminative power needed to distinguish the target from subtle distractors.

Ablation Setting	Avg. Acc	Hypothesis Validation
1. Impact of Proposal Strategy		
(a) w/ External Detector (DINO)	41.2%	Intrinsic Prior > Generic Objectness
(b) w/ Direct Regression	53.5%	**Explicit Attn. > Coordinate Reg.**
(c) w/ Native Resolution	54.8%	Resolution-Attention Conflict
2. Impact of Verification Strategy		
(d) w/o Verification (Global Only)	48.5%	Necessity of Coarse-to-Fine
(e) **Full FastFocus Model**	**61.8%**	**Optimal Synergy**

Table 3: Ablation of Critical Design Choices. We systematically dissect the architectural contributions. Part 1 (Proposal Strategy) validates the *Resolution-Attention Conflict*, showing that our low-resolution intrinsic prior outperforms both external detectors (Row a) and native-resolution baselines (Row c). Part 2 (Verification Strategy) demonstrates the necessity of the coarse-to-fine synergy: while the Global Proposal Module (Row d) acts as a structural abstractor, the *Recursive Pairwise Verification* (Row e) is indispensable for resolving fine-grained ambiguities, yielding a +13.3% gain.

5 Conclusion

In this work, we critically examine the efficacy of native resolution processing in GUI grounding, uncovering the Resolution-Attention Conflict where excessive visual density leads to Semantic Fragmentation. We propose FastFocus, a biomimetic perception framework that resolves this dilemma by decoupling grounding into coarse-grained structural abstraction and fine-grained recursive verification.

Our empirical findings lead to three key conclusions: (1) Structure Precedes Detail: Low-resolution inputs act as superior Structural Abstractors, effectively aggregating semantic layout information where native high-resolution baselines often succumb to attention dispersion. (2) Discrimination over Regression: The success of our Recursive Pairwise Verification demonstrates that discriminative tournament selection offers significantly higher robustness against visual ambiguity compared to standard coordinate regression, particularly for abstract icons. (3) Consistency Across Evaluated Backbones: Validated across diverse SFT and RL baselines, FastFocus proves that a cognitive-inspired "System 2" perception module can consistently improve grounding across the evaluated backbones without backbone retraining.

By establishing new state-of-the-art results on the ScreenSpot-Pro benchmark, FastFocus demonstrates that precise grounding arises not from brute-force resolution scaling, but from the intelligent organization of visual attention. We hope this work inspires future research to move toward such hierarchical perception strategies for more robust autonomous agents.

Acknowledgments

Xiaocheng Feng and Bing Qin are the co-corresponding authors of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62522603 and 62276078, the Key R&D Program of Heilongjiang under Grant 2022ZX01A32, and the Fundamental Research Funds for the Central Universities under Grant XNJKKGYDJ2024013.

References

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[Bai *et al.*, 2025a] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Li Ying Meng, Xuancheng Ren, Xin yi Ren, Sibo Song, Yu-Chen Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Botao Zheng, Humen Zhong, Jingren Zhou, Fanxi Zhou, Jingren Zhou, Yuanzhi Zhu, and Keming Zhu. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025.

[Bai *et al.*, 2025b] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[Bar, 2004] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[Carrasco, 2011] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[Chen *et al.*, 2024] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[Cheng *et al.*, 2024] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332, 2024.

[Gou *et al.*, 2025] Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Gu *et al.*, 2025] Zhangxuan Gu, Zhengwen Zeng, Zhenyu Xu, Xingran Zhou, Shuheng Shen, Yunfei Liu, Beitong Zhou, Changhua Meng, Tianyu Xia, Weizhi Chen, et al. Ui-venus technical report: Building high-performance ui agents with rft. *arXiv preprint arXiv:2508.10833*, 2025.

[Hochstein and Ahissar, 2002] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002.

[Hong *et al.*, 2024] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.

[Larson and Loschky, 2009] Adam M Larson and Lester C Loschky. The contributions of central versus peripheral vision to scene gist recognition. *Journal of vision*, 9(10):6–6, 2009.

[Li *et al.*, 2024] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773, 2024.

[Li *et al.*, 2025] Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Liu *et al.*, 2025] Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*, 2025.

[Lu *et al.*, 2025] Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Guanjing Xiong, and Hongsheng Li. Ui-r1: Enhancing action prediction of gui agents by reinforcement learning. *arXiv preprint arXiv:2503.21620*, 1(2):3, 2025.

[Luo *et al.*, 2025] Tiange Luo, Lajanugen Logeswaran, Justin Johnson, and Honglak Lee. Visual test-time scaling for gui agent grounding. *arXiv preprint arXiv:2505.00684*, 2025.

[Oliva and Torralba, 2006] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[OpenAI, 2024] OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

[Qian *et al.*, 2024] Yijun Qian, Yujie Lu, Alexander Hauptmann, and Oriana Riva. Visual grounding for user interfaces. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 97–107, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[Qin *et al.*, 2025] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

[Schyns and Oliva, 1994] Philippe G Schyns and Aude Oliva. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological science*, 5(4):195–200, 1994.

[Sun *et al.*, 2024] Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, et al. Os-genesis: Automating gui agent trajectory construction via reverse task synthesis. *arXiv preprint arXiv:2412.19723*, 2024.

[Toyama *et al.*, 2021] Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. Androidenv: A reinforcement learning platform for android. *arXiv preprint arXiv:2105.13231*, 2021.

[Vasu *et al.*, 2025] Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, et al. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19769–19780, 2025.

[Wang *et al.*, 2024a] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[Wang *et al.*, 2024b] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wu *et al.*, 2025a] Hang Wu, Hongkai Chen, Yujun Cai, Chang Liu, Qingwen Ye, Ming-Hsuan Yang, and Yiwei Wang. Dimo-gui: Advancing test-time scaling in gui grounding via modality-aware visual reasoning. *arXiv preprint arXiv:2507.00008*, 2025.

[Wu *et al.*, 2025b] Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, et al. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*, 2025.

[Wu *et al.*, 2025c] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: Foundation action model for generalist gui agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Xu *et al.*, 2025a] Hai-Ming Xu, Qi Chen, Lei Wang, and Lingqiao Liu. Attention-driven gui grounding: Leveraging pretrained multimodal large language models without fine-tuning. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025.

[Xu *et al.*, 2025b] Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous GUI interaction, 2025.

[Yan *et al.*, 2023] An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, et al. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*, 2023.

[Zhang *et al.*, 2024] Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, et al. Large language model-brained gui agents: A survey. *arXiv preprint arXiv:2411.18279*, 2024.

[Zhang *et al.*, 2025a] Borui Zhang, Bo Zhang, Bo Wang, Wenzhao Zheng, Yuhao Cheng, Liang Tang, Yiqiang Yan, Jie Zhou, and Jiwen Lu. Bami: Training-free bias mitigation in gui grounding. *Technical Report*, 2025.

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

[Zhang *et al.*, 2025c] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zheng *et al.*, 2024] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v (ision) is a generalist web agent, if grounded. In *International Conference on Machine Learning*, pages 61349–61385. PMLR, 2024.

[Zhou *et al.*, 2025] Yuqi Zhou, Sunhao Dai, Shuai Wang, Kaiwen Zhou, Qinglin Jia, and Jun Xu. Gui-g1: Understanding r1-zero-like training for visual grounding in gui agents. *arXiv preprint arXiv:2505.15810*, 2025.