Look Where You're Told: Instruction-Consistent Attention for GUI Grounding

Seonhoon Kim, Zhiyu Chen, Xin Li, Qun Liu
Amazon.com Inc, Seattle, USA
{seonhoon,zhiyuche,xinll,qunliu}@amazon.com

Abstract

Visual grounding in graphical user interface (GUI) requires accurate localization of UI elements from natural language instructions. Conventional coordinate generation approaches face inherent limitations, including sensitivity to resolution variations and lack of interpretability. Recently, coordinate-free attention-based methods have emerged as a promising alternative, but these methods primarily rely on spatial location signals from ground-truth bounding boxes to supervise attention, with limited mechanisms to explicitly verify that the learned attention distributions reflect genuine semantic correspondence between the instruction and the attended visual regions. We propose Attention Cycle-Consistency (ACC), a self-supervised regularization framework that enforces bidirectional alignment between visual attention and instruction semantics. ACC introduces two complementary constraints: semantic consistency, which ensures attended visual regions contain sufficient information to reconstruct the original instruction, and spatial consistency, which requires attention distributions to remain invariant when cycled through instruction reconstruction. We further incorporate entropy regularization to encourage spatially concentrated attention. ACC is applicable as a lightweight, model-agnostic regularizer for attention-based coordinate-free grounding methods, adding zero computational overhead at inference as all auxiliary components are discarded after training.

1 Introduction

Graphical user interfaces (GUIs) are the primary abstraction through which users operate modern software. Building agents that follow natural-language instructions and act directly on pixels offers a unified interface across heterogeneous platforms, without requiring structured representations such as DOM trees or accessibility graphs that can be incomplete or unavailable (Xie et al., 2024; Cheng et al., 2024). Recent evaluations suggest that overall agent success is frequently constrained by failures at the perception-to-action boundary, where models must identify the correct on-screen target (Xie et al., 2024). This bottleneck is naturally formalized as *GUI visual grounding*: given a screenshot and an instruction, predict the actionable UI element.

A prevalent modeling choice casts GUI grounding as coordinate generation, where models directly produce a click point or bounding box (Cheng et al., 2024; Gou et al., 2024; Wu et al., 2024). While coordinates provide a convenient execution interface, this formulation has structural limitations: performance degrades on high-resolution screenshots where targets are small and densely surrounded by distractors (Li et al., 2025), supervision is ambiguous since multiple points within an element are valid, and coordinate outputs offer limited diagnostic value when failures occur (Wu et al., 2025).

These weaknesses have motivated coordinate-free, attention-based formulations that predict distributions over visual patch tokens (Wu et al., 2025; Zhou et al., 2025). However, attention visualization alone does not guarantee *semantic grounding*. When attention is primarily trained with spatial targets, the resulting distributions may still reflect spurious correlations rather than genuine instruction-element correspondence. This concern aligns with findings that attention weights can be weakly coupled to model decisions (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). For GUI grounding, spurious attention is particularly damaging: interfaces vary widely across applications and platforms, so shortcut alignment to superficial cues is unlikely to transfer to unseen software or professional contexts.

We propose *Attention Cycle-Consistency* (ACC), a lightweight regularization framework that directly validates whether attention patterns encode the intended semantics. The core insight is that *se-*

Proceedings of the 4th Workshop on Advances in Language and Vision Research (ALVR), pages 155–163

mantically grounded attention should be recoverable: if attention truly captures instruction intent, attended regions should contain sufficient information to reconstruct the original instruction, and re-attending with the reconstruction should yield the same spatial distribution. ACC operationalizes this through two complementary losses. *Semantic consistency* requires attended visual content to reconstruct the instruction, enforcing that attention selects information-dense regions rather than spuriously correlated patches. *Spatial consistency* closes the loop: the reconstructed instruction must re-induce the original attention distribution, with deviations providing fine-grained self-supervision. We further incorporate entropy regularization to encourage spatially concentrated attention.

ACC is model-agnostic, requires no annotations beyond existing training data, and integrates with coordinate-free attention heads as a regularizer. Crucially, all auxiliary components are discarded after training, adding zero computational overhead at inference. We validate ACC on three benchmarks including ScreenSpot-Pro, which specifically tests grounding in high-resolution professional interfaces where existing methods show significant degradation.

In summary, this work makes the following contributions:

- We introduce **ACC**, a self-supervised attention regularizer that enforces semantic and spatial cycle-consistency for GUI grounding, providing a principled mechanism to validate instruction-element correspondence without additional annotations.
- We design ACC as a **plug-and-play regularizer** applicable to coordinate-free grounding architectures, with **zero inference overhead** since all auxiliary components are discarded after training.
- We conduct extensive experiments on three GUI grounding benchmarks, demonstrating that ACC **consistently improves** existing methods across diverse settings, with notable gains on challenging high-resolution professional interfaces.

2 Related Work

2.1 GUI Grounding

GUI grounding has recently emerged as a core bottleneck for computer-use agents operating directly on pixels (Xie et al., 2024). A dominant paradigm formulates grounding as coordinate regression, where models directly predict click points or bounding boxes in continuous screen space, such as SeeClick, UGround, OS-Atlas, and Aria-UI (Cheng et al., 2024; Gou et al., 2024; Wu et al., 2024; Yang et al., 2025). While practical for execution, coordinate-based methods suffer from resolution sensitivity, supervision ambiguity (multiple valid click points within a region), and limited interpretability when failures occur (Wu et al., 2025; Li et al., 2025).

More recently, coordinate-free approaches predict attention distributions over visual patch tokens instead of explicit coordinates. Methods such as GUI-Actor and GUI-AIMA align instruction semantics with visual tokens through attention-based matching, producing interpretable heatmaps and improved robustness across resolutions and layouts (Wu et al., 2025; Zhou et al., 2025). In particular, GUI-AIMA aligns intrinsic multimodal attention with patch-level grounding signals through a context anchor, moving beyond purely spatial supervision (Zhou et al., 2025). A complementary direction explores adaptive exploration policies that encourage semantic alignment during grounding (Liu et al., 2026). Additionally, intrinsic attention extraction from pretrained multimodal large language models has demonstrated that attention maps can serve as grounding signals without explicit coordinate regression (Xu et al., 2025). In concurrent work, attention-based GUI grounding has been enhanced through multimodal fusion with OCR-derived textual cues and icon-level caption semantics (Ma et al., 2026).

Despite these advances, existing coordinate-free methods primarily rely on spatial location signals from ground-truth bounding boxes when supervising attention, and they do not provide an explicit mechanism to verify that the learned attention reflects genuine semantic correspondence between instructions and visual regions. This becomes particularly challenging in high-resolution professional environments where existing grounding models show significant performance degradation (Xie et al., 2024; Li et al., 2025). Our work complements these directions by regularizing attention recoverability via cycle-consistency, providing a closed-loop verification signal that is orthogonal to richer alignment cues.

2.2 Attention Faithfulness and Reconstruction-Based Alignment

The interpretability and faithfulness of attention mechanisms have been widely debated. Jain and Wallace (Jain and Wallace, 2019) and Serrano and Smith (Serrano and Smith, 2019) show that attention weights can be weakly correlated with feature importance and can sometimes be altered with limited impact on predictions. Wiegreffe and Pinter (Wiegreffe and Pinter, 2019) further argue that attention faithfulness depends on evaluation design, while evidence from vision-language tasks demonstrates divergence between model attention and human reasoning (Das et al., 2017). These findings suggest that attention distributions alone do not guarantee semantic grounding.

Reconstruction-based grounding offers a principled mechanism for enforcing meaningful visual-text alignment. GroundeR (Rohrbach et al., 2016) learns phrase localization by requiring attended image regions to reconstruct textual phrases, effectively turning grounding into an information bottleneck. Similar reconstruction objectives have been used in grounded captioning and multimodal alignment (Ma et al., 2020; Wang et al., 2024). While effective, these approaches typically supervise localization indirectly and do not explicitly enforce stability or recoverability of attention distributions.

ACC extends this line of work by not only requiring attended regions to reconstruct the instruction (semantic consistency), but also enforcing that the reconstructed instruction re-induces the same spatial attention pattern (spatial consistency), directly validating attention recoverability.

2.3 Cycle-Consistency for Self-Supervised Alignment

Cycle-consistency has emerged as a general principle for learning meaningful correspondences without additional supervision. In image translation, CycleGAN enforces reconstruction after round-trip domain mapping to discourage degenerate solutions (Zhu et al., 2017). In vision-language tasks, cycle constraints have been applied to robust visual question answering (Shah et al., 2019), grounded captioning (Ma et al., 2020), and mutual consistency between captioning and grounding (Wang et al., 2024).

These methods share the intuition that valid correspondences should be recoverable under a closed loop transformation. Unlike prior cycle-based approaches that operate at the level of images, captions, or answers, ACC applies cycle-consistency directly to spatial attention distributions in GUI grounding. By enforcing round-trip agreement between instruction semantics and attention maps, ACC regularizes the internal alignment mechanism itself rather than only its outputs, providing fine-grained supervision without requiring additional annotations.

3 Method

In this section, we describe our Attention Cycle-Consistency (ACC) framework. Given an existing GUI grounding model as the base, ACC introduces two additional components: (1) an instruction reconstruction decoder, and (2) cycle-consistency and attention regularization losses. These components regularize the base model's attention to be semantically grounded without modifying its architecture. We denote the input screenshot as I and the natural language instruction as $T = \{t_1, t_2, \cdots, t_L\}$ where t_i is the i-th token and L is the sequence length. The overall architecture of the proposed ACC is shown in Figure 1.

3.1 Preliminaries: Base GUI Grounding Model

ACC is designed as a model-agnostic regularizer that can be integrated into attention-based GUI grounding architectures. We first briefly describe how these methods produce attention distributions, which ACC leverages for cycle-consistency regularization.

Given a screenshot I and instruction T, attention-based GUI grounding models such as GUI-Actor (Wu et al., 2025) and GUI-AIMA (Zhou et al., 2025) produce an attention distribution $A \in \mathbb{R}^{H \times W}$ over the visual patch grid, where H and W denote the height and width of the patch grid, respectively. This attention is computed through a dedicated action head:

$$A = \text{softmax}\left(\frac{z^\top Z}{\sqrt{d}}\right) \tag{1}$$

where $z \in \mathbb{R}^d$ is the contextual anchor embedding that encodes the instruction semantics, and $Z \in \mathbb{R}^{d \times HW}$ are the projected visual patch features. ACC operates on this attention distribution A to enforce semantic grounding.

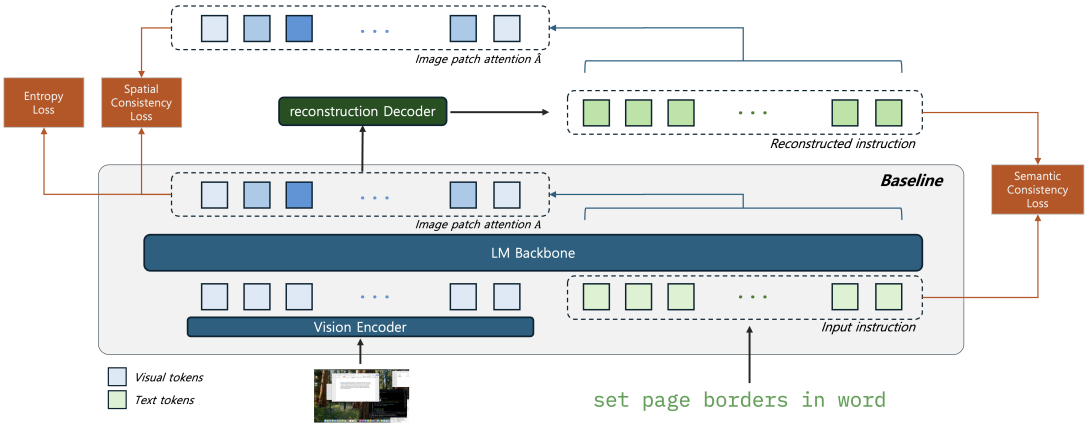

Figure 1: Overview of Attention Cycle-Consistency (ACC) framework. The baseline model (gray box) takes a screenshot and input instruction, processes them through a Vision Encoder and LM Backbone, and produces image patch attention A for GUI grounding. ACC introduces three regularization losses during training: (1) Semantic Consistency Loss ensures that attended visual patches contain sufficient information to reconstruct the original instruction via a reconstruction decoder; (2) Spatial Consistency Loss enforces that the backward attention $\hat{A}$, computed from the reconstructed instruction, aligns with the forward attention A; (3) Entropy Loss encourages the attention distribution to sharpen, promoting more concentrated focus on the target element. Blue and green boxes denote visual and text tokens, respectively, with darker shades indicating higher attention weights. The reconstruction decoder and associated losses are used only during training and discarded at inference, resulting in zero computational overhead.

3.2 Instruction Reconstruction Decoder

The core component of ACC is the instruction reconstruction decoder D, which reconstructs the original instruction from attended visual regions. This decoder validates whether the attention distribution genuinely captures semantic information from the instruction.

Attended Feature Extraction Given the attention distribution A and visual patch features $V = \{v_1, v_2, \cdots, v_{HW}\}$, we first compute the attended visual representation as a weighted sum:

$$\tilde{v} = \sum_{i=1}^{HW} a_i \cdot v_i \tag{2}$$

where a_i is the attention weight for the i-th patch. We then apply a region encoding function $f(\cdot)$ to obtain the final context feature:

$$v_{\text{ctx}} = f(\tilde{v}) \tag{3}$$

where $f(\cdot)$ is implemented as a multi-layer perceptron (MLP) that projects the attended representation into a suitable feature space for instruction reconstruction.

Decoder Architecture The instruction reconstruction decoder is a lightweight transformer decoder that generates the instruction sequence in an autoregressive manner. Starting from v_{ctx} as the initial context, the decoder reconstructs the instruction:

$$P(T|A, V) = \prod_{l=1}^{L} P(t_l|t_{<l}, v_{\text{ctx}}) \tag{4}$$

The decoder is designed to be lightweight to minimize additional parameters while maintaining sufficient capacity for instruction reconstruction.

3.3 Attention Cycle-Consistency Losses

ACC enforces bidirectional alignment through two complementary losses: semantic consistency loss and spatial consistency loss.

Semantic Consistency Loss The semantic consistency loss ensures that attended visual regions contain sufficient information to reconstruct the original instruction. We minimize the negative log-likelihood of the ground-truth instruction given the attended features:

$$\mathcal{L}_{\text{sem}} = -\frac{1}{L} \sum_{l=1}^{L} \log P(t_l|t_{<l}, v_{\text{ctx}}) \tag{5}$$

This loss directly measures whether the attention captures semantically meaningful regions. If the model attends to the correct regions, the decoder will successfully reconstruct the original instruction.

Spatial Consistency Loss The spatial consistency loss enforces that re-attending using the reconstructed instruction yields the same attention distribution. Given the reconstructed instruction $\hat{T} = D(A, V)$, we compute the backward attention by calculating the attention between the anchor embedding of $\hat{T}$ and the visual patch features:

$$\hat{A} = \text{softmax}\left(\frac{\hat{z}^\top Z}{\sqrt{d}}\right) \tag{6}$$

where $\hat{z}$ is the anchor embedding derived from the reconstructed instruction $\hat{T}$. The spatial consistency loss is defined as the symmetric KL divergence between the original and backward attention distributions:

$$\mathcal{L}_{\text{spa}} = \frac{1}{2}\left(\text{KL}(A\|\hat{A}) + \text{KL}(\hat{A}\|A)\right) \tag{7}$$

This loss ensures that the attention is recoverable through the reconstruction cycle.

3.4 Attention Entropy Regularization

While the cycle-consistency losses ensure that the attention is semantically correct, they do not explicitly encourage spatial precision. GUI screenshots often contain visually similar or repeated elements (e.g., multiple buttons with similar appearance), and the reconstruction decoder may successfully reconstruct the instruction from a diffuse attention distributed across these redundant regions. In such cases, the cycle is satisfied but the model has not localized the specific target element.

To address this, we introduce an entropy regularization loss that encourages the attention distribution to maintain a controlled level of concentration. We define the entropy of the attention distribution as:

$$H(A) = -\sum_{i=1}^{HW} a_i \log a_i \tag{8}$$

Rather than directly minimizing entropy, which could collapse the attention to a single patch and fail to cover the full spatial extent of the target element, we regularize the entropy toward a target value τ:

$$\mathcal{L}_{\text{ent}} = |H(A) - \tau| \tag{9}$$

This formulation prevents two failure modes: overly diffuse attention that spreads across irrelevant regions when $H(A) > \tau$, and overly concentrated attention that collapses to a single patch without covering the full spatial extent of the target element when $H(A) < \tau$.

Total Training Objective. The complete training objective combines the base grounding loss with the cycle-consistency and entropy regularization losses:

$$\mathcal{L} = \mathcal{L}_{\text{ground}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{spa}}\mathcal{L}_{\text{spa}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} \tag{10}$$

where $\mathcal{L}_{\text{ground}}$ is the original grounding loss, and λ_{sem}, λ_{spa}, λ_{ent} are balancing hyperparameters.

3.5 Training Strategy

Two-Stage Training. We adopt a two-stage training strategy to stabilize the learning process. In the first stage, we train the base grounding model with only $\mathcal{L}_{\text{ground}}$ to obtain initial attention patterns. In the second stage, we introduce the ACC losses and jointly fine-tune the entire model. This prevents the reconstruction decoder from receiving random attention signals in the early training phase.

Gradient Stopping for Spatial Consistency. When computing $\mathcal{L}_{\text{spa}}$, we stop gradients through the instruction reconstruction process to prevent the model from trivially satisfying spatial consistency by making the decoder output constant tokens. This ensures that spatial consistency is achieved through improving attention quality rather than degrading reconstruction quality.

Inference. At inference time, the instruction reconstruction decoder is discarded, and only the base grounding model is used. This means ACC adds zero computational overhead during inference while providing regularization benefits during training.

4 Experiments

In this section, we evaluate the effectiveness of our proposed framework, ACC, on three representative GUI grounding benchmarks. We first describe the experimental setup including datasets, baselines, and implementation details, then present comparisons with strong coordinate-free baselines. Finally, we conduct ablation studies to analyze the contribution of each proposed regularization loss.

4.1 Experimental Setup

Baselines. We compare against two strong coordinate-free grounding frameworks:

- **GUI-Actor**: A coordinate-free grounding model that performs visual token alignment without explicit coordinate regression.

- **GUI-AIMA**: A context-anchored multimodal alignment model leveraging intrinsic multimodal attention mechanisms.

Importantly, our method does not modify the architecture of either baseline. Instead, we introduce additional training-time regularization losses. During inference, the architecture, parameter count, and computational cost remain identical to the original baselines.

Implementation Details. We train our models using the same data recipe as GUI-AIMA from publicly available GUI datasets. Hyperparameters follow the official baseline settings unless otherwise specified. For the instruction reconstruction decoder, we use GPT-2 medium (355M parameters), which consists of 24 transformer layers with 1024 hidden dimensions and 16 attention heads. We set $\lambda_{\text{sem}} = 0.3$, $\lambda_{\text{spa}} = 0.3$, and $\lambda_{\text{ent}} = 0.3$, with loss weights linearly warmed up from 0 to their target values between steps 500 and 1500. For the entropy regularization, we set the target entropy $\tau = 0.5$. For all experiments, we adopt the two-step inference with zoom-in strategy from GUI-AIMA (Zhou et al., 2025), where the model first predicts an approximate location on the full screenshot, then refines the prediction on a cropped and zoomed-in region centered at the initial prediction.

Evaluation Benchmarks. We evaluate on three well-established benchmarks:

- **ScreenSpot-v2**: A corrected version of ScreenSpot with fixed annotation errors and disambiguated instructions, containing samples across mobile, desktop, and web platforms.

- **ScreenSpot-Pro**: A benchmark for GUI grounding in professional high-resolution environments, featuring 1,581 expert-annotated tasks across 23 professional applications where target elements are significantly smaller, and interfaces are more visually complex than in general-use settings.

- **OSWorld-G**: A comprehensive benchmark comprising 564 finely annotated samples across diverse task types including text matching, element recognition, layout understanding, and fine-grained manipulation.

Table 1: Performance comparison on ScreenSpot-v2 and ScreenSpot-Pro benchmarks.

Model	ScreenSpot-v2			ScreenSpot-Pro		
	Text	Icon	Avg	Text	Icon	Avg
GUI-Actor 2B	94.15	75.99	86.24	56.60	25.17	44.59
GUI-Actor 2B + ACC	93.31	78.16	86.71	57.32	24.50	44.78
GUI-Actor 3B	95.40	80.69	88.99	60.70	27.48	48.01
GUI-Actor 3B + ACC	93.73	80.87	88.13	62.95	31.62	50.98
GUI-AIMA 2B	95.26	80.87	88.99	58.75	28.64	47.25
GUI-AIMA 2B + ACC	95.54	80.69	89.07	59.47	30.63	48.45
GUI-AIMA 3B	95.68	83.57	90.41	66.02	37.58	55.15
GUI-AIMA 3B + ACC	96.10	85.20	91.35	69.40	37.09	57.05

Table 2: Performance comparison on the OSWorld-G benchmark.

Model	OSWorld-G				
	Text Match	Elem. Recog.	Layout Und.	Fine-grain Manip.	All
GUI-Actor 2B	62.07	55.15	60.08	40.79	52.48
GUI-Actor 2B + ACC	63.98	58.18	62.85	40.13	54.43
GUI-Actor 3B	68.20	63.94	66.40	42.11	58.87
GUI-Actor 3B + ACC	66.67	66.36	70.36	42.11	60.64
GUI-AIMA 2B	63.22	59.39	63.24	41.45	55.50
GUI-AIMA 2B + ACC	64.75	60.91	65.61	38.82	55.85
GUI-AIMA 3B	69.35	65.76	70.36	46.71	60.99
GUI-AIMA 3B + ACC	69.73	65.76	67.98	47.37	61.52

We use Element Accuracy as the evaluation metric, which measures the proportion of predictions where the predicted point falls within the ground-truth bounding box.

4.2 Experimental Results

Tables 1 and 2 present the performance of ACC applied to both GUI-Actor and GUI-AIMA across two model scales (2B and 3B).

ScreenSpot-v2 and ScreenSpot-Pro. On ScreenSpot-v2, ACC yields consistent improvements for GUI-AIMA at both scales, with GUI-AIMA 3B + ACC achieving the highest overall accuracy of 91.35%. For GUI-Actor, ACC improves Icon accuracy but shows a slight trade-off on Text accuracy, resulting in mixed average performance. The improvements become more pronounced on ScreenSpot-Pro, which features high-resolution professional screenshots where target elements are significantly smaller relative to the full screen and embedded within visually complex interfaces. Notably, GUI-Actor 3B + ACC achieves a 2.97% absolute gain over the baseline (50.98% vs. 48.01%), and GUI-AIMA 3B + ACC improves by 1.90% (57.05% vs. 55.15%). These larger gains on the more challenging benchmark suggest that the cycle-consistency and entropy regularization are particularly effective when the model must localize small target elements

Table 3: Attention analysis on ScreenSpot-Pro. *In-box Mass*: fraction of attention inside the GT box (↑). *Global Entropy*: entropy of the full attention distribution (↓). *Density Ratio*: average attention per patch inside vs. outside the GT box (↑).

	In-box Mass (↑)			Global Entropy (↓)			Density Ratio (↑)		
Model	Text	Icon	Avg	Text	Icon	Avg	Text	Icon	Avg
Baseline	0.171	0.058	0.128	4.153	4.333	4.222	396.8	330.2	371.3
+ ACC	0.228	0.083	**0.173**	3.403	3.546	**3.457**	823.0	585.1	**732.1**

Table 4: Ablation study on ScreenSpot-Pro. Starting from the full ACC framework applied to GUI-Actor 3B, we progressively remove loss components.

Model	**ScreenSpot-Pro**		
	Text	Icon	Avg
GUI-Actor 3B + ACC	62.95	31.62	50.98
$-\mathcal{L}_{\text{ent}}$	63.56	29.64	50.60
$-\mathcal{L}_{\text{ent}} - \mathcal{L}_{\text{spa}}$	62.74	29.80	50.16

within cluttered, high-resolution interfaces.

OSWorld-G. On OSWorld-G, ACC improves the overall accuracy across all four configurations. GUI-Actor 3B + ACC achieves the largest gain of 1.77% (60.64% vs. 58.87%), with notable improvements in Element Recognition (+2.42%) and Layout Understanding (+3.96%). GUI-AIMA 3B + ACC shows a modest overall gain of 0.53%, with improvements concentrated in Text Match and Fine-grained Manipulation. The consistent overall gains across diverse task types suggest that the improved attention quality from ACC benefits a range of grounding capabilities, from recognizing specific elements to understanding spatial layouts.

4.3 Ablation Study

To understand the contribution of each proposed loss, we conduct an ablation study by progressively removing components from the full ACC framework applied to GUI-Actor 3B, evaluated on ScreenSpot-Pro. We compare three configurations: (1) full ACC with all three losses, (2) without entropy loss, and (3) without both entropy and spatial consistency losses. Results are shown in Table 4. Removing the entropy loss leads to a notable drop in Icon accuracy by -1.98%, resulting in a lower average despite a slight increase in Text accuracy. This suggests that entropy regularization is particularly important for localizing small, non-textual elements where attention concentration is critical. Further removing the spatial consistency loss degrades both Text and overall accuracy, confirming that the forward-backward cycle provides complementary regularization beyond what the reconstruction objective alone achieves. The progressive degradation in average accuracy across all three configurations validates that each loss component contributes meaningfully to the overall framework.

4.4 Attention Analysis

To understand how ACC regularization reshapes the model's attention, we compare the attention distributions of the baseline and ACC-trained models using three complementary metrics computed over the ScreenSpot-Pro evaluation set. *In-box Attention Mass* measures the fraction of total attention that falls within the ground-truth bounding box, reflecting whether the model attends to the correct region. *Global Entropy* measures the entropy of the full attention distribution, where lower values indicate sharper, more concentrated attention. *In-box Density Ratio* is defined as the average attention per patch inside the ground-truth box divided by the average per patch outside, capturing how selectively the model focuses on the target relative to the background. Results are reported in Table 3.

Across all three metrics, the ACC model shows consistent improvement over the baseline. In-box attention mass increases from 0.128 to 0.173 on average, indicating that a larger share of the model's attention is directed toward the target element. Global entropy decreases from 4.222 to 3.457, confirming that ACC produces sharper, more concentrated attention distributions. The in-box density ratio nearly doubles from 371 to 732, meaning that each patch inside the ground-truth box receives roughly twice the average attention compared to the baseline, relative to background patches.

These results demonstrate that ACC regularization produces attention that is simultaneously more accurate (higher in-box mass), more focused (lower entropy), and more discriminative (higher density

ratio), confirming that the cycle-consistency and entropy losses address complementary aspects of attention quality.

5 Conclusion

In this paper, we presented Attention Cycle-Consistency (ACC), a self-supervised regularization framework for GUI visual grounding that enforces bidirectional alignment between visual attention and instruction semantics. ACC introduces two complementary cycle-consistency constraints: semantic consistency, which ensures attended regions contain sufficient information to reconstruct the original instruction, and spatial consistency, which requires attention distributions to remain invariant through the reconstruction cycle. We further incorporate entropy regularization to encourage spatially concentrated attention on target elements. Our approach is model-agnostic and integrates seamlessly with existing coordinate-free grounding methods without modifying their architecture. Experiments on three benchmarks demonstrate that ACC consistently improves baseline methods across diverse GUI environments. Since ACC components are discarded at inference time, our method achieves these gains with zero computational overhead during deployment. Extending ACC to coordinate-generation architectures is a promising direction for future work: attention mechanisms also serve as intermediate representations in these models, and investigating whether cycle-consistency regularization yields comparable benefits in that setting remains an open empirical question that we leave to subsequent study.

References

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. 2024. 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.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2024. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8778–8786.

Yuhang Liu, Zeyu Liu, Shuanghe Zhu, Pengxiang Li, Congkai Xie, Jiasheng Wang, Xueyu Hu, Xiaotian Han, Jianbo Yuan, Xinyao Wang, and 1 others. 2026. Infigui-g1: Advancing gui grounding with adaptive exploration policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32267–32275.

Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2020. Learning to generate grounded visual captions without localization supervision. In *European conference on computer vision*, pages 353–370. Springer.

Longhui Ma, Di Zhao, Siwei Wang, Zhao Lv, and Miao Wang. 2026. Trifuse: Enhancing attention-based gui grounding via multimodal fusion. *arXiv preprint arXiv:2602.06351*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2931–2951.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

Ning Wang, Jiajun Deng, and Mingbo Jia. 2024. Cycle-consistency learning for captioning and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5535–5543.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 11–20.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, and 1 others. 2025. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and 1 others. 2024. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

Hai-Ming Xu, Qi Chen, Lei Wang, and Lingqiao Liu. 2025. Attention-driven gui grounding: Leveraging pretrained multimodal large language models without fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8851–8859.

Yuhao Yang, Yue Wang, Dongxu Li, Ziyang Luo, Bei Chen, Chao Huang, and Junnan Li. 2025. Aria-ui: Visual grounding for gui instructions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22418–22433.

Shijie Zhou, Viet Dac Lai, Hao Tan, Jihyung Kil, Wanrong Zhu, Changyou Chen, and Ruiyi Zhang. 2025. Gui-aima: Aligning intrinsic multimodal attention with a context anchor for gui grounding. *arXiv preprint arXiv:2511.00810*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.