SoGUI: Cognition-Conditioned Interaction for Social GUI Agents

Anonymous authors
Paper under double-blind review

Abstract

Progress in vision–language models has made GUI agents increasingly capable, yet most benchmarks provide the interaction goal and primarily evaluate where to act. Social interfaces introduce an earlier challenge: the agent must infer an interaction-relevant semantic state from observed content before a specified policy determines whether and how to interact. We introduce SoGUI, a cognition-conditioned interaction (CCI) formulation that enables separate analysis of semantic inference and policy-conditioned GUI grounding under a controlled policy. We further construct a 3,916-record Theory-Grounded Social-Semantic Cognition Benchmark covering affective valence, visible source-credibility cues, and gain/loss framing, with image-disjoint held-out-platform evaluation. Across three training seeds, SoGUI-SFT-CoG improves Strict Accuracy over the standard cognition-free NoCoG baseline by 6.56/12.37 points on ID/OOD and achieves 94.02%/87.21% E2E-Hit. A prompt-matched NoCoG+Rule control shows that explicit cognition supervision provides additional gains beyond policy-rule prompting alone, improving Strict Accuracy by 2.69/4.73 points on ID/OOD. Error decomposition further shows that once semantic inference is improved, GUI localization becomes the dominant residual bottleneck, particularly under held-out-platform evaluation. These results indicate that semantic cognition and GUI grounding are partially separable capabilities, motivating GUI agents that combine stronger semantic understanding with robust cross-platform grounding.

1 Introduction

Mobile GUI agents have recently made substantial progress in automating smartphone tasks, including cross-app navigation (Lu et al., 2025), multi-field form filling (Rawles et al., 2025), and general app control (Christianos et al., 2025). A central capability underlying these systems is GUI grounding: mapping an instruction or user intent to the corresponding text, icon, or interactive control on a screenshot (Nguyen et al., 2025). Large-scale grounding pre-training has substantially advanced this capability, as demonstrated by SeeClick (Cheng et al., 2024), while OS-ATLAS (Wu et al., 2025), UGround (Gou et al., 2025), and ScreenSpot-Pro (Li et al., 2025b) extend grounding across increasingly diverse interfaces. Figure 1 (a) illustrates this goal-conditioned setting: given "click the Settings icon," the agent identifies the specified target, answering *where to click*.

Recent studies have begun to push GUI agents beyond pure grounding through state-aware reasoning and reinforcement learning (Yang et al., 2026; Deng et al., 2026; Wu et al., 2026; Lian et al., 2026; Zhou et al., 2026; Lu et al., 2026c). State-aware reasoning (Wu et al., 2026), for example, determines whether an instructed action should be executed under the current UI state, extending the problem from *where to click* to *whether to execute a supplied action* (Figure 1 (b)). Despite this progress, existing settings still largely assume that the interaction goal is specified externally: the agent is told what to achieve and only needs to determine how—or whether —to execute it.

Beyond goal-conditioned interaction. A more challenging setting emerges when no per-content interaction action is explicitly provided. In social-media interfaces, an agent must first infer a task-relevant semantic state from the observed content, considering cues such as affective valence, source credibility, and message framing. Under a specified interaction policy, this semantic state determines whether an interaction is required and which behavior should be executed, such as like, forward, or no action. This raises a fundamental question: *Can GUI agents infer the semantic state required*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

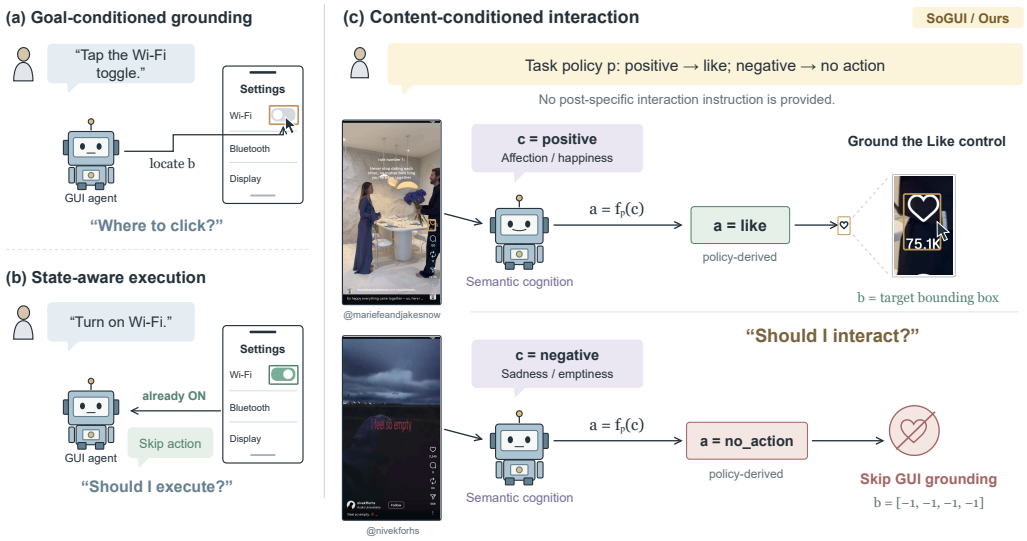

Figure 1: **Expanding GUI agent capability. (a) Goal-conditioned grounding** answers "Where to click?" by mapping an explicit instruction to a GUI target. **(b) State-aware execution** answers "Should I execute?" by comparing the current UI state with a supplied goal. **(c) Cognition-conditioned interaction (this work)** asks "Should I interact under the specified policy?" when no per-content action instruction is given. The agent infers the semantic state of the content, derives the corresponding interaction under a predefined policy, and grounds the relevant control only when interaction is required.

for interaction and subsequently ground the corresponding GUI control? We refer to this setting as **cognition-conditioned interaction (CCI)** (Figure 1 (c)). Unlike goal-conditioned grounding and state-aware execution, CCI does not directly provide a per-content action instruction; instead, the agent must first infer the semantic state consumed by the interaction policy. To isolate this capability, we define the interaction policy as a fixed and auditable mapping from semantic states to actions. This controlled formulation enables separate evaluation of semantic inference and GUI grounding, providing clearer attribution of interaction failures.

Existing GUI benchmarks provide limited visibility into this capability. Benchmarks such as ScreenSpot (Cheng et al., 2024) and AndroidWorld (Rawles et al., 2025) have substantially advanced our understanding of GUI agents by evaluating instruction-conditioned grounding and goal-directed task completion. However, they all assume that the desired objective is explicitly specified to the agent. Consequently, strong performance on these benchmarks primarily reflects the ability to execute known goals, while leaving open whether agents can infer the task-relevant semantics needed before GUI interaction can be executed.

To investigate this gap, we construct a diagnostic benchmark designed to disentangle semantic cognition from GUI grounding. Unlike many existing GUI evaluations that primarily report end-to-end success without explicitly separating these factors, our benchmark provides structured annotations for semantic state, policy-derived interaction behavior, and GUI localization. This design enables us to diagnose whether failures primarily arise from content understanding or from grounding the corresponding action. We operationalize cognition-conditioned interaction through three theory-grounded semantic factors: affective valence, source credibility, and message framing. These factors capture different forms of semantic evidence required for content-conditioned interaction in social-media scenarios (Tversky & Kahneman, 1981; Pornpitakpan, 2004; Yan et al., 2012). We further evaluate cross-platform generalization to examine whether cognition-conditioned interaction transfers beyond the training environments.

A capability gap beyond grounding. As illustrated in Figure 1, conventional GUI grounding and cognition-conditioned interaction require different but related capabilities. Models that perform strongly on grounding-oriented tasks do not necessarily excel at inferring the semantic states

required for policy-conditioned interaction. Conversely, general-purpose multimodal models can achieve competitive performance in this setting despite lacking specialized GUI grounding training. This observation suggests that these capabilities are partially separable: accurate localization of interface elements does not necessarily imply identifying the semantic states that trigger interaction under a given policy. To understand the sources of failure in cognition-conditioned interaction, we separate two potential bottlenecks: semantic cognition and GUI localization. Oracle-cognition analysis further indicates that improving semantic-state inference substantially benefits interaction performance, suggesting that semantic cognition represents a distinct challenge beyond grounding. Meanwhile, cross-model comparisons reveal that strong GUI localization does not systematically translate into reliable policy-conditioned interaction, motivating the need to evaluate semantic inference and grounding as separate capability dimensions.

Testing the diagnosis. To validate this capability decomposition, we conduct controlled supervision studies that examine whether explicit semantic-state supervision improves cognition-conditioned interaction. By comparing models with and without cognition supervision under matched training settings, we find that improving semantic-state inference consistently enhances interaction performance across training seeds. Further error decomposition reveals that, after semantic inference is strengthened, GUI localization becomes the dominant remaining challenge. These results support the view that semantic cognition and GUI grounding constitute complementary yet partially separable capabilities. Our contributions are threefold:

1. **Capability formulation.** We formulate **cognition-conditioned interaction (CCI)**, a setting in which GUI agents must infer interaction-relevant semantic states before grounding policy-conditioned actions. CCI extends existing GUI evaluations beyond settings where goals or actions are explicitly provided.
2. **Diagnostic benchmark.** We introduce a diagnostic benchmark with 3,916 interaction records annotated with semantic states, policy-derived actions, and GUI localization targets, enabling separate evaluation of semantic inference and GUI grounding.
3. **Capability analysis.** Through comprehensive evaluation and controlled studies, we show that semantic cognition and GUI grounding are partially separable capabilities. Improving semantic inference reduces cognition-related failures, while localization remains the dominant residual challenge.

2 Theory-Grounded Social-Semantic Cognition Benchmark

2.1 Benchmark Motivation and Design Principle

To study cognition-conditioned interaction, we construct the Theory-Grounded Social-Semantic Cognition Benchmark, which decomposes interaction into semantic-state inference, policy-conditioned action derivation, and GUI localization. Unlike conventional GUI settings in which the desired action is explicitly specified, each instance requires the agent to infer an interaction-relevant semantic state from the observed content before grounding the corresponding GUI control. Given a social-media screenshot, the agent first predicts the semantic state, applies a predefined interaction policy to obtain the corresponding action, and localizes the relevant control when an interaction is required. The benchmark treats the interaction policy as a controlled variable. Each semantic state is deterministically mapped to an action, making action prediction a policy-consistency check rather than an independent learned capability. This design enables separate analysis of semantic inference and GUI localization without introducing policy learning as an additional confounding factor. Detailed policy mappings, dataset statistics, and annotation procedures are provided in Appendix H.

2.2 Theory-Grounded Semantic Constructs

The semantic states in Social GUI are grounded in established theories of human communication and information processing. These theories provide principled semantic dimensions for constructing observable content-level signals rather than arbitrary annotation categories. We consider three representative constructs: affective valence, source credibility, and message framing. Affective valence captures the emotional orientation conveyed by content, motivated by emotional contagion

theory (Kramer et al., 2014), which studies how emotional expressions influence subsequent perceptions and behaviors. Source credibility captures the trustworthiness and authority associated with information sources, following classical credibility theory and its extensions in online environments (Hovland & Weiss, 1951; Westerman et al., 2014; Lin et al., 2016). Message framing captures how equivalent information presented through different perspectives can influence interpretation and decisions, motivated by framing theory (Tversky & Kahneman, 1981). These constructs define the semantic states that an agent must recognize, rather than prescribing human interaction preferences. The predefined policy serves only as an evaluation mechanism that converts semantic states into reproducible GUI actions, enabling controlled measurement of content-conditioned interaction capability.

2.3 Task Formulation and Annotation

We formulate Social GUI around three components: semantic cognition, policy-derived action, and GUI grounding. Each task record contains (1) a **cognition label** describing the task-relevant semantic state, (2) an **action label** deterministically derived from the stated policy, and (3) a **target box** when interaction is required. Action therefore serves as a policy-consistency check rather than an independent learned capability, enabling separate analysis of semantic and localization errors. The action space is `no_action`, `like`, or `forward`; actionable boxes are normalized to $[0, 1000]^4$, while no-action records use the sentinel `[-1,-1,-1,-1]`.

Cognition annotation. Affective-valence labels capture positive and negative textual or visual affect; source-credibility labels distinguish official organizations, individual accounts, and low-credibility accounts using visible identity, verification, and publisher cues; and framing labels distinguish gain- from loss-oriented presentation rather than isolated sentiment.

Quality control. For annotation-quality auditing, three annotators independently labeled a balanced sample of 500 examples per cognition dimension before adjudication. Appendix F reports the annotation criteria and agreement statistics. Disagreements were adjudicated, policy actions were derived automatically from cognition labels, and target boxes were manually verified.

2.4 Dataset Construction and Held-out-Platform Evaluation

Social GUI contains 3,916 instances from six platforms. Its image-disjoint ID and held-out-platform OOD protocol evaluates cross-platform generalization under platform-level visual and layout shifts. Training and ID evaluation use Weibo, TikTok (Douyin), and Instagram, whereas OOD evaluation uses X, RedNote (Xiaohongshu), and WeChat Channels (Weixin Channels). The held-out platforms differ in layout, interaction patterns, presentation style, and visual conventions. Screenshots never cross splits, and all task records derived from the same screenshot remain grouped to avoid screenshot-level leakage. Detailed split statistics and platform-level composition are provided in Appendix H.

2.5 Capability-Oriented Evaluation Metrics

We report stage-wise and end-to-end metrics. **Cognition Accuracy** tests the annotated semantic state; **Action Accuracy** is a policy-consistency diagnostic because action is derived from cognition. **Center Hit** tests whether the predicted point lies in the target, and **Mean IoU** measures box overlap. Across output formats, **E2E-Hit (Action+Hit)** is:

$$\text{E2E-Hit} = \frac{\sum_{i\in\mathcal{A}} \mathbb{1}[\hat{a}_i = a_i^\star \wedge \hat{p}_i \in b_i^\star] + \sum_{i\in\mathcal{N}} \mathbb{1}[\hat{a}_i = \text{no_action} \wedge \hat{p}_i = (-1,-1)]}{|\mathcal{A}| + |\mathcal{N}|}, \quad (1)$$

where $\mathcal{A}/\mathcal{N}$ are actionable/no-action examples. GT boxes only score the predicted point; point agents need not emit boxes, while box models use the predicted center. For box-output models, **Strict Accuracy** additionally requires the correct interaction and $\text{IoU} \geq 0.65$ on actionable examples, with the same no-action rule. Together the metrics separate semantics, grounding, and final success. Appendix I provides the complete parsing, denominator, native-point, box-scoring, and coordinate-handling rules.

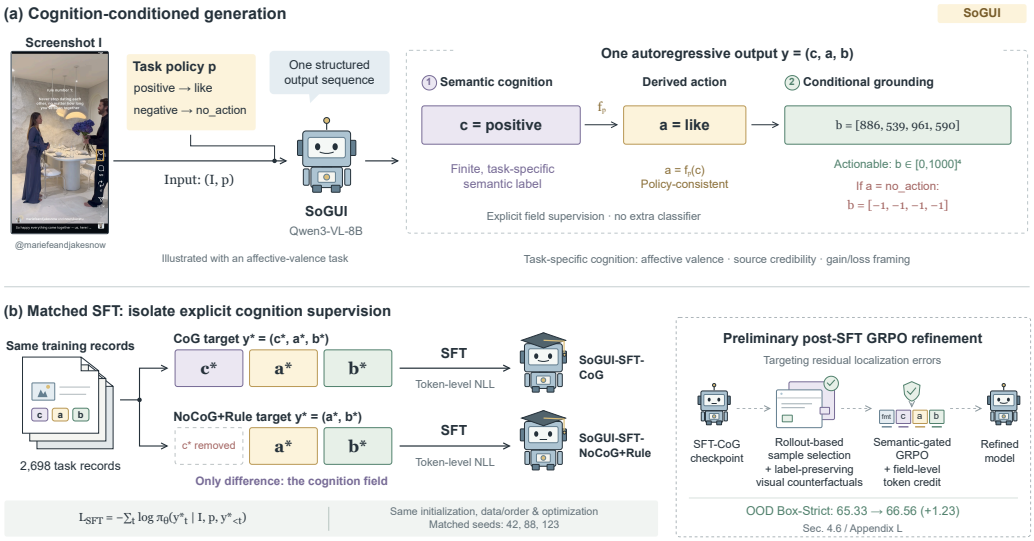

Figure 2: Overview of SoGUI. **(a)** The model predicts a structured interaction representation $y = (c, a, b)$, where c denotes semantic cognition, $a = f_p(c)$ denotes the policy-derived interaction action, and b denotes the GUI localization target. **(b)** The matched CoG/NoCoG+Rule pair isolates explicit cognition supervision; the preliminary GRPO follow-up then targets the residual localization bottleneck identified in Finding 4, rather than redefining the main method.

3 Cognition-Conditioned Interaction

3.1 Problem Formulation

Conventional GUI grounding maps a screenshot and explicit instruction to an action and target, $(I, p) \rightarrow (a, b)$. Social content introduces an additional capability requirement: the agent must infer the semantic state of the content before the interaction policy can determine whether an interaction is required and which action to take. We therefore introduce semantic cognition c and formulate $(I, p) \rightarrow (c, a, b)$. Under the controlled benchmark policy, $a = f_p(c)$ is deterministic: the interaction action is derived from the inferred cognition rather than learned as an independent capability. This separation ensures that a failure can arise from incorrect semantic inference or from localization after correct inference, while the emitted action simply checks policy consistency.

3.2 Explicit Cognition Supervision

Figure 2 shows our Qwen3-VL-8B-based model (Bai et al., 2025). It generates $y^\star_{\mathrm{CoG}} = (c^\star, f_p(c^\star), b^\star)$, incorporating the annotated cognition state directly into one autoregressive sequence rather than adding a classifier or separate reasoning stage:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{T} \log \pi_\theta(y^\star_t \mid I, p, y^\star_{<t}). \tag{2}$$

This directly supervises the decision preceding grounding and exposes a compact, predefined, measurable intermediate state, enabling reproducible separation of semantic and localization failures rather than relying on unconstrained traces.

3.3 Standard and Prompt-Matched Cognition Controls

The standard NoCoG baseline uses the cognition-free target $y^\star_{\mathrm{NoCoG}} = (a^\star, b^\star)$ and a direct semantic instruction that requests the appropriate action and location without enumerating the intermediate label-to-action rule. This is the conventional cognition-free condition used in the main comparison.

To isolate target-side supervision, the prompt-matched NoCoG+Rule control shares Qwen3-VL-8B initialization, records, policies, boxes, order, optimization, and the policy-explicit prompt with CoG. Only the target differs: $y^{\star}_{\mathrm{NoCoG+Rule}} = (a^{\star}, b^{\star})$ versus $y^{\star}_{\mathrm{CoG}} = (c^{\star}, a^{\star}, b^{\star})$, isolating cognition supervision from data or training scale. Both conditions use matched training seeds and their corresponding output schemas; inference-time cognition prompting is reported separately as a complementary diagnostic.

4 Experiments

4.1 Experimental Setup

Evaluation groups. We evaluate three groups of models to characterize the relationship between conventional GUI grounding and cognition-conditioned interaction. (1) *GUI-specialized agents* include Qwen2.5-VL-3B-UI-R1-E (Lu et al., 2026a), GUI-R1-7B (Luo et al., 2025), POINTS-GUI-G (Zhao et al., 2026), and ScaleCUA-7B (Liu et al., 2026b). These models are designed or adapted specifically for GUI perception and interaction. (2) *General-purpose MLLMs* include GPT-5.6 Luna, Gemini 3.7 Flash, Kimi K2.5, Doubao-Seed-2.1-turbo, and Qwen3-VL-8B. They serve as broadly capable multimodal baselines without specialized GUI grounding training. (3) *SoGUI models* consist of SoGUI-SFT-CoG, the standard SoGUI-SFT-NoCoG baseline, and the prompt-matched SoGUI-SFT-NoCoG+Rule control. All three models share the same initialization, training records, and optimization configuration. Standard NoCoG uses neither a cognition output nor an explicit cognition-to-action mapping. NoCoG+Rule adds the policy mapping to the prompt while retaining the cognition-free target, whereas CoG additionally supervises the cognition field. This controlled hierarchy allows us to distinguish the effect of prompt-side policy scaffolding from that of target-side cognition supervision.

Evaluation protocol. All models are evaluated on Social GUI under the same interaction space and evaluation protocol. We report E2E-Hit as the primary end-to-end metric, measuring whether the predicted action is policy-consistent and, when an interaction is required, grounded to the correct GUI target. To analyze different failure sources, we additionally report Cognition Accuracy and localization metrics. Mean IoU and Strict Accuracy are reported for models that output bounding boxes, while point-native GUI agents are evaluated using their predicted interaction points.

Grounding preservation evaluation. To examine whether improving cognition-conditioned interaction affects conventional GUI grounding ability, we additionally evaluate SoGUI-SFT-CoG and its initialization model on ScreenSpot-Mobile. This provides an external check that gains in CCI do not come at the expense of conventional GUI grounding. Training details, complete metric and coordinate conventions, and dataset statistics are provided in Appendix E, Appendix I, and Appendix H, respectively.

4.2 Finding 1: Grounding Does Not Imply Interaction

We first investigate whether conventional GUI grounding ability is sufficient for cognition-conditioned interaction. Figure 3 directly compares grounding performance on ScreenSpot-Mobile with end-to-end performance on Social GUI. Models that perform strongly on conventional grounding do not exhibit the same relative ordering on Social GUI, where semantic inference is required before interaction. A similar pattern is observed among similarly sized models. ScaleCUA-7B and Qwen3-VL-8B achieve identical 94.01% grounding accuracy on ScreenSpot-Mobile, while their Social GUI E2E-Hit differs substantially (56.77% vs. 75.22%). Thus, similar localization performance can correspond to markedly different cognition-conditioned interaction performance. Table 1 extends this observation across a broader set of models. GUI-specialized agents do not consistently outperform general-purpose MLLMs on Social GUI E2E-Hit despite their strong conventional grounding performance.

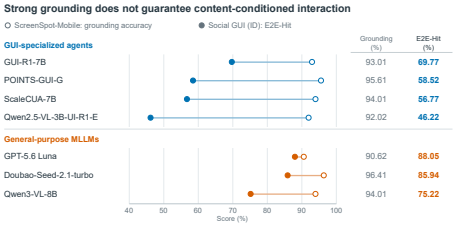

Figure 3: ScreenSpot-Mobile grounding versus Social GUI ID E2E-Hit.

Table 1: Social GUI main results on ID/OOD (%). E2E-Hit is Action+Hit. Mean IoU and Strict are undefined for point-native agents. NoCoG denotes the standard direct semantic-to-action baseline; Table 2 separates prompt-rule and cognition-target effects.

	ID ($N = 569$)						Held-out-platform OOD ($N = 649$)					
Model	Cognition Accuracy	Action Accuracy	Mean IoU	Center Hit	E2E-Hit	Strict Accuracy	Cognition Accuracy	Action Accuracy	Mean IoU	Center Hit	E2E-Hit	Strict Accuracy
GUI-specialized agents												
Qwen2.5-VL-3B-UI-R1-E	53.43	56.94	–	33.05	46.22	–	50.54	52.54	–	28.87	39.45	–
GUI-R1-7B	82.43	84.71	–	73.95	69.77	–	78.89	80.89	–	60.82	62.25	–
POINTS-GUI-G	64.32	66.96	–	78.71	58.52	–	61.63	70.42	–	42.01	43.76	–
ScaleCUA-7B	71.00	73.46	–	68.63	56.77	–	69.18	67.18	–	64.69	56.09	–
General-purpose MLLMs												
GPT-5.6 Luna	92.09	92.09	63.56	93.28	88.05	64.67	91.37	91.37	47.14	**85.05**	82.74	47.15
Gemini 3.7 Flash	91.21	89.98	55.47	70.31	72.93	67.49	90.45	89.68	36.38	54.12	64.87	50.54
Kimi K2.5	91.92	91.92	32.95	58.54	67.66	35.68	89.68	89.68	27.35	43.30	58.71	35.59
Doubao-Seed-2.1-turbo	89.81	89.81	74.08	**93.84**	85.94	78.21	89.06	89.06	55.73	84.54	80.74	50.85
Qwen3-VL-8B	88.22	88.22	50.94	76.19	75.22	52.20	82.43	82.43	39.19	63.92	68.72	45.15
SoGUI models												
SoGUI-SFT-NoCoG	–	88.05	71.11	82.35	85.41	77.50	–	77.81	40.51	63.14	69.18	48.38
SoGUI-SFT-CoG	**97.19**	**97.19**	**80.29**	92.72	**94.02**	**85.76**	**93.84**	**93.84**	**58.02**	84.28	**87.21**	**65.33**

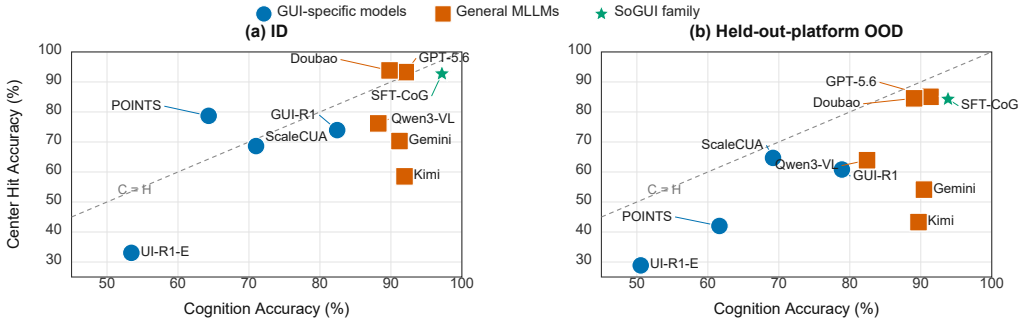

Figure 4: Cognition Accuracy versus Center Hit: two partially separable capabilities.

Finding 1. *Strong GUI grounding does not necessarily translate into strong cognition-conditioned interaction, indicating that semantic inference introduces an additional capability requirement beyond localization.*

4.3 Finding 2: Cognition and Grounding Are Separable Capabilities

Finding 1 shows that conventional grounding does not necessarily transfer to content-conditioned interaction. To locate this gap, Figure 4 plots Cognition Accuracy against Center Hit for every evaluated model. The two capabilities are only partially correlated: GUI-specialized agents tend toward stronger localization but weaker semantic cognition, whereas general-purpose MLLMs show the opposite tendency and variable grounding. The lack of a clear monotonic relationship shows that content-conditioned interaction requires both understanding *what* the content implies and grounding *where* the corresponding action should be performed.

A paired Oracle-cognition intervention supports the same diagnosis from a different direction. Supplying the ground-truth semantic label increases Cog-Hit by 17.22/11.09 points for GUI-R1-7B and 15.82/18.34 points for POINTS-GUI-G on ID/OOD. Appendix Table 19 reports the four paired comparisons and Appendix I specifies exactly what the Oracle condition changes. This diagnostic does not replace the protocol-fair E2E-Hit comparison.

Finding 2. *Semantic cognition and GUI grounding are partially separable capabilities. Reliable cognition-conditioned interaction requires both inferring the interaction-relevant semantic state and accurately grounding the corresponding GUI control.*

4.4 Finding 3: Explicit Cognition Supervision Improves Interaction

Given the partially separable roles of semantic cognition and GUI grounding, we next examine whether explicitly supervising the semantic state improves cognition-conditioned interaction.

Table 2: Controlled SFT ablation over three training seeds (mean ± sample standard deviation, %).

Training configuration	Split	Action	Mean IoU	Center Hit	Strict
SoGUI-SFT-CoG	ID	$\mathbf{97.31 \pm 0.20}$	$\mathbf{80.13 \pm 0.56}$	$\mathbf{92.62 \pm 0.16}$	$\mathbf{85.53 \pm 0.90}$
	OOD	$\mathbf{93.89 \pm 0.24}$	$\mathbf{57.46 \pm 0.53}$	$\mathbf{83.51 \pm 0.68}$	$\mathbf{64.71 \pm 0.53}$
− cognition target (NoCoG+Rule)	ID	96.25 ± 0.62	78.65 ± 0.21	92.06 ± 0.43	82.84 ± 0.37
	OOD	90.96 ± 1.91	53.86 ± 1.24	79.98 ± 1.32	59.99 ± 1.69
− rule scaffold (NoCoG)	ID	87.64 ± 2.06	71.64 ± 1.66	81.60 ± 2.33	78.97 ± 1.50
	OOD	78.22 ± 1.44	41.39 ± 0.81	61.94 ± 1.46	52.34 ± 3.43
Total Δ (CoG − NoCoG)	ID	**+9.67**	**+8.49**	**+11.02**	**+6.56**
	OOD	**+15.67**	**+16.07**	**+21.57**	**+12.37**

We compare three controlled SFT conditions. Standard NoCoG uses the cognition-free (a, b) target together with a direct semantic instruction. NoCoG+Rule retains the same target but explicitly provides the task-specific cognition-to-action rule in the prompt. CoG uses the same policy-explicit prompt as NoCoG+Rule while additionally supervising c in the output target. The CoG–NoCoG comparison captures the overall difference from the standard cognition-free setting, while the matched CoG–NoCoG+Rule comparison isolates the contribution of target-side cognition supervision. Table 2 separates the effects of policy scaffolding and explicit cognition supervision. Under the same policy-explicit prompt, adding the cognition target improves Strict Accuracy by 2.69 points on ID and 4.73 points on held-out-platform OOD. Removing the rule scaffold further reduces Strict Accuracy by 3.87 and 7.65 points, respectively, with consistent degradation across the shared metrics. Overall, CoG improves over standard NoCoG by 6.56/12.37 Strict points on ID/OOD, with larger gains under held-out-platform evaluation. Complete per-seed results, seed-42 paired-bootstrap intervals, and complementary inference-time prompting analyses are provided in Appendix Tables 4–8 and Appendix B.5.

Finding 3. *Explicit supervision of the semantic state provides additional gains beyond policy-rule prompting alone, supporting semantic cognition as a distinct and learnable component of cognition-conditioned interaction.*

4.5 Finding 4: Localization Becomes the Remaining Bottleneck

After improving semantic cognition, we analyze the residual failures that remain when the semantic state is correctly inferred. Table 3 shows that cognition errors account for only 2.81%/6.16% of ID/OOD examples, whereas 11.43%/28.51% retain the correct cognition but fail the final interaction. Since the benchmark policy deterministically maps cognition to action, these residual failures are primarily associated with GUI localization. The substantially higher OOD rate further indicates the challenge of grounding under cross-platform visual and layout variation. The class-conditional results provide a more fine-grained view of this residual error. On OOD `forward`, CoG reaches 92.86% Action Accuracy but only 59.18% E2E-Hit; for OOD `like`, Center Hit reaches 92.76% while Box-Strict is only 44.48%. Appendix Table 21 provides the full matched breakdown, and Appendix K reports representative prediction cases.

Table 3: Primary-model failure decomposition.

Failure category	**ID**	**OOD**
Incorrect cognition	2.81	6.16
Cognition-correct localization failure	11.43	28.51
Total unresolved failure	**14.24**	**34.67**

Together, these results show a shift in the dominant source of error: once semantic cognition is strengthened, GUI localization accounts for most of the remaining failures, particularly under held-out-platform evaluation.

Finding 4. *After semantic cognition is improved, GUI localization becomes the dominant residual bottleneck, especially under cross-platform interface shift.*

4.6 Localization-Targeted GRPO Refinement

Motivated by Finding 4, a lightweight post-SFT GRPO follow-up (Shao et al., 2024) uses rollout selection, stable replay, and label-preserving counterfactuals to target localization headroom without layout shortcuts. A semantic-gated continuous IoU–center reward activates only after correct format, cognition, and action; field masks route credit, and a one-epoch update is KL-regularized against SFT (Chen et al., 2026c). GRPO preserves ID Cognition/Action Accuracy at 97.19% while raising Strict by 0.71 points; on OOD, Mean IoU, Center Hit, and Strict improve by 0.76, 0.26, and 1.23 points with stable cognition. This reinforces Finding 4: targeted optimization can recover residual localization performance after semantic decisions stabilize. This diagnosis-driven design

gates geometry on semantic correctness and routes field-level advantages to box tokens, preventing accidental overlap from rewarding a wrong decision and concentrating updates on the capability isolated by Finding 4. Appendix L provides the objectives, configuration, full results, and qualitative progression in Figure 9.

5 Related Work

GUI grounding and multimodal GUI agents. Recent work on GUI agents has substantially improved visual grounding, cross-platform interaction, and benchmark coverage. SeeClick, UGround, OS-ATLAS, ShowUI, UIPro, and ScreenSpot-Pro study scalable grounding, generalization, and evaluation across diverse interfaces (Cheng et al., 2024; Gou et al., 2025; Wu et al., 2025; Lin et al., 2025; Li et al., 2025a;b). A growing line of reasoning-centric GUI agents further incorporates intermediate reasoning or structured analysis before action execution (Wang et al., 2026b; Chen et al., 2026b; Liu et al., 2026a; Chen et al., 2026a). However, these settings generally assume that the user goal or intended action is already specified. In contrast, SoGUI studies the preceding capability of inferring an interaction-relevant semantic state from observed social content before grounding the policy-conditioned GUI action.

Reinforcement learning for GUI agents. Reinforcement learning has also become an important approach for improving GUI interaction, with recent methods exploring continuous rewards, trajectory resampling, visual critics, action history, and multi-turn optimization (Lian et al., 2026; Wang et al., 2026a; Zhou et al., 2026; Lu et al., 2026c;b; Chen et al., 2026c). These approaches primarily optimize action execution or grounding quality within a specified task objective. Our use of RL is complementary: after semantic cognition is strengthened through supervised training, we apply localization-focused refinement only to the remaining grounding errors. Accordingly, RL is treated as a secondary optimization stage rather than the main mechanism for learning cognition-conditioned interaction.

Theory-grounded social cognition. Beyond interface perception, social and web agents must interpret semantic properties of observed content before acting (Xue et al., 2026; Qiu et al., 2025). We operationalize three interaction-relevant constructs grounded in established communication and information-processing theories: affective valence motivated by Emotional Contagion (Hatfield et al., 1993), source credibility (Hovland & Weiss, 1951), and gain/loss framing (Tversky & Kahneman, 1981). Related theories such as Spiral of Silence, Agenda Setting, and heuristic information processing provide broader context for socially conditioned behavior (Noelle-Neumann, 1974; McCombs & Shaw, 1972; Chaiken, 1980; McCombs et al., 1997). Importantly, these theories motivate the semantic constructs to be recognized rather than prescribe normative interaction preferences; SoGUI uses a fixed benchmark policy only to convert recognized states into reproducible GUI actions.

6 Conclusion

This work introduces cognition-conditioned interaction (CCI), a setting in which GUI agents must infer a task-relevant semantic state before grounding an interaction under a predefined policy. Through a theory-grounded benchmark and controlled SFT studies, we show that semantic cognition and GUI grounding are partially separable capabilities. Explicit cognition supervision provides additional gains beyond policy-rule prompting alone, while residual-error analysis shows that localization becomes the dominant bottleneck once semantic inference is improved.

By treating the interaction policy as a controlled variable, CCI enables separate analysis of semantic inference and GUI grounding within a unified interaction protocol. Cross-model comparisons, Oracle-cognition analysis, and multi-seed controlled experiments consistently support this capability decomposition. These findings motivate future GUI agents that combine stronger semantic understanding with robust cross-platform grounding, and eventually extend CCI to learned or user-specific interaction policies.

AI USE STATEMENT

In this work, we used generative AI tools to assist research ideation and execution, including feedback on experimental design, code development and debugging, and organizing and interpreting analyses. We also used these tools to identify related literature, polish the manuscript, and improve the presentation of tables and figures. This empirical study contains no formal proofs, so AI-assisted proof writing is not applicable. The authors checked references against original publications, reviewed reported results against experimental records, examined AI-assisted code and presentation, and take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

This work studies cognition-conditioned GUI decision making in a controlled offline setting. Cognition-to-action mappings are synthetic benchmark policies designed for objective evaluation and should not be interpreted as normative recommendations for real users or social platforms. Social theories motivate measurable cognition variables rather than prescribe GUI behavior. Evaluated agents operate on screenshots and do not perform live social-media interactions. Social screenshots can nevertheless contain account names, avatars, and user-generated content. A public release will therefore include only images whose redistribution is permitted by the applicable source terms; otherwise we will release derived annotations and collection scripts. Before redistribution, visible identifiers will be reviewed and de-identified where needed. These safeguards and the exact release scope will be documented in the dataset card.

REPRODUCIBILITY STATEMENT

The paper specifies the structured task schema, theory-to-task operationalization, annotation protocol, image-disjoint held-out-platform split, evaluation criteria, and matched SFT comparison. SoGUI-SFT-CoG and SoGUI-SFT-NoCoG use the same 2,698 records and optimization setup; the standard NoCoG prompt omits the explicit cognition-to-action rule. The additional SoGUI-SFT-NoCoG+Rule condition shares CoG's rule-explicit prompt and differs only by the cognition field, enabling the matched target-side comparison. The appendix provides the output schemas, matched-control protocol, parsing rules, dataset statistics, and annotation-quality details. Appendix B reports seeds 42/88/123, sample standard deviations, per-seed metrics, and seed-42 paired-bootstrap procedures, including resampling units and interval construction. The common E2E-Hit definition, sample denominators, and native point protocol are also documented. Appendix L documents the training-only rollout audit, counterfactual curriculum, semantic-gated reward, field-level credit, and the single-seed evaluation scope of the GRPO follow-up. An anonymized code repository for review is available at `https://anonymous.4open.science/r/SoGUI-review-code-7789`. Code, processed annotations, and releasable data will be published with the final version.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shelly Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5):752–766, 1980.

Liangyu Chen, Hanzhang Zhou, Chenglin Cai, Jianan Zhang, Panrong Tong, Xu Zhang, Quyu Kong, Chen Liu, Yuqi Liu, Wenxuan Wang, Yue Wang, Qin Jin, and Steven C. H. Hoi. UI-Ins: Enhancing GUI grounding with multi-perspective instruction as reasoning. In *International Conference on Learning Representations*, 2026a.

Ruihan Chen, Qiming Li, Xiaocheng Feng, Weihong Zhong, Xiaoliang Yang, Yuxuan Gu, Zekun Zhou, Yunfei Lu, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. MPR-GUI: Benchmarking and enhancing multilingual perception and reasoning in GUI agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 29798–29832. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1375.

Yuhan Chen, Yuxuan Liu, Long Zhang, Pengzhi Gao, Jian Luan, and Wei Liu. STEP: Success-rate-aware trajectory-efficient policy optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30681–30692. Association for Computational Linguistics, July 2026c. doi: 10.18653/v1/2026.findings-acl.1532.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, 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)*, pp. 9313–9332, 2024. doi: 10.18653/v1/2024.acl-long.505.

Filippos Christianos, Georgios Papoudakis, Thomas Coste, Jianye Hao, Jun Wang, and Kun Shao. Lightweight neural app control. In *International Conference on Learning Representations*, 2025.

Zehao Deng, Tianjie Ju, Zheng Wu, Zhuosheng Zhang, and Gongshen Liu. Training high-level schedulers with execution-feedback reinforcement learning for long-horizon GUI automation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27525–27535, June 2026.

Boyu Gou, Demi 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 *International Conference on Learning Representations*, 2025.

Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. Emotional contagion. *Current Directions in Psychological Science*, 2(3):96–100, 1993.

Carl I. Hovland and Walter Weiss. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, 15(4):635–650, 1951.

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

Hongxin Li, Jingran Su, Jingfan Chen, Zheng Ju, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. UIPro: Unleashing superior interaction capability for GUI agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1613–1623, 2025a.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8778–8786, 2025b. doi: 10.1145/3746027.3755688.

Shuquan Lian, Yuhang Wu, Jia Ma, Yifan Ding, Zihan Song, Bingqi Chen, Xiawu Zheng, Hui Li, and Rongrong Ji. UI-AGILE: Advancing GUI agents with effective reinforcement learning and precise inference-time grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8575–8584, June 2026.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, and Mike Zheng Shou. ShowUI: One vision-language-action model for GUI visual agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19498–19508, 2025.

Xialing Lin, Patric R. Spence, and Kenneth A. Lachlan. Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, 63:264–271, 2016.

Yuhang Liu, Pengxiang Li, Zishu Wei, Congkai Xie, Xueyu Hu, Xinchen Xu, Shengyu Zhang, Xiaotian Han, Hongxia Yang, and Fei Wu. InfiGUIAgent: A multimodal generalist GUI agent with native reasoning and reflection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1035–1051. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.eacl-long.45.

Zhaoyang Liu, Jingjing Xie, Zichen Ding, Zehao Li, Bowen Yang, Zhenyu Wu, Xuehui Wang, Qiushi Sun, Shi Liu, Weiyun Wang, Shenglong Ye, Qingyun Li, Zeyue Tian, Gen Luo, Xiangyu Yue, Biqing Qi, Kai Chen, Bowen Zhou, Yu Qiao, Qifeng Chen, and Wenhai Wang. Scale-CUA: Scaling open-source computer use agents with cross-platform data. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=yBFUqdJFZn.

Quanfeng Lu, Wenqi Shao, Zitao Liu, Lingxiao Du, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, and Ping Luo. GUIOdyssey: A comprehensive dataset for cross-app GUI navigation on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22404–22414, 2025.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Pengxiang Zhao, Guangyi Liu, Guanjing Xiong, and Hongsheng Li. UI-R1: Enhancing efficient action prediction of GUI agents by reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17608–17616, 2026a. doi: 10.1609/aaai.v40i21.38816.

Zhengxi Lu, Fei Tang, Guangyi Liu, Jin Ma, Kaitao Song, Xu Tan, Wenqi Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. UI-copilot: Advancing long-horizon GUI automation via tool-integrated policy optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19741–19762. Association for Computational Linguistics, July 2026b. doi: 10.18653/v1/2026.acl-long.904.

Zhengxi Lu, Jiabo Ye, Fei Tang, Yongliang Shen, Haiyang Xu, Ziwei Zheng, Weiming Lu, Ming Yan, Fei Huang, Jun Xiao, and Yueting Zhuang. Experience-driven multi-turn reinforcement learning for GUI agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9474–9496. Association for Computational Linguistics, July 2026c. doi: 10.18653/v1/2026.acl-long.428.

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. GUI-R1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.

Maxwell McCombs, Juan Pablo Llamas, Esteban Lopez-Escobar, and Federico Rey. Candidate images in spanish elections: Second-level agenda-setting effects. *Journalism & Mass Communication Quarterly*, 74(4):703–717, 1997.

Maxwell E. McCombs and Donald L. Shaw. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187, 1972.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Seunghyun Yoon, Lina Yao, Branislav Kveton, Jihyung Kil, Thien Huu Nguyen, Trung Bui, Tianyi Zhou, Ryan A. Rossi, and Franck Dernoncourt. GUI agents: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22522–22538. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1158.

Elisabeth Noelle-Neumann. The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2):43–51, 1974.

Chanthika Pornpitakpan. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34(2):243–281, 2004. doi: 10.1111/j.1559-1816.2004.tb02547.x.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3978–4005. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.222.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry,

Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A dynamic benchmarking environment for autonomous agents. In *International Conference on Learning Representations*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

Wenkai Wang, Xiyun Li, Hongcan Guo, Wenhao Yu, Tianqing Fang, Haitao Mi, Dong Yu, and Shengyu Zhang. Measure twice, click once: Co-evolving proposer and visual critic via reinforcement learning for GUI grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20964–20984. Association for Computational Linguistics, July 2026a. doi: 10.18653/v1/2026.acl-long.960.

Xuehui Wang, Zhenyu Wu, JingJing Xie, Zichen Ding, Bowen Yang, Zehao Li, Zhaoyang Liu, Qingyun Li, Xuan Dong, Zhe Chen, Weiyun Wang, Xiangyu Zhao, Jixuan Chen, Haodong Duan, Tianbao Xie, Chenyu Yang, Shiqian Su, Yue Yu, Yanting Zhang, Xiangyu Yue, Weijie Su, Xizhou Zhu, Wei Shen, Jifeng Dai, and Wenhai Wang. MMBench-GUI: A unified hierarchical evaluation framework for multi-platform GUI agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6239–6248, 2026b.

David Westerman, Patric R. Spence, and Brandon Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2):171–183, 2014.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. OS-ATLAS: Foundation action model for generalist gui agents. In *International Conference on Learning Representations*, 2025.

Zongru Wu, Rui Mao, Zhiyuan Tian, Pengzhou Cheng, Tianjie Ju, Zheng Wu, Lingzhong Dong, Haiyue Sheng, Zhuosheng Zhang, and Gongshen Liu. See, think, act: Teaching multimodal agents to effectively interact with GUI by identifying toggles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27536–27546, June 2026.

Dizhan Xue, Jing Cui, Shengsheng Qian, Chuanrui Hu, and Changsheng Xu. Some: A realistic benchmark for llm-based social media agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1391–1399, 2026. doi: 10.1609/aaai.v40i2.37113.

Changmin Yan, James P. Dillard, and Fuyuan Shen. Emotion, motivation, and the persuasive effects of message framing. *Journal of Communication*, 62(4):682–700, 2012. doi: 10.1111/j.1460-2466.2012.01655.x.

Yan Yang, Dongxu Li, Yutong Dai, Yuhao Yang, Ziyang Luo, Zirui Zhao, Zhiyuan Hu, Junzhe Huang, Amrita Saha, Zeyuan Chen, Ran Xu, Liyuan Pan, Caiming Xiong, and Junnan Li. GTA1: GUI test-time scaling agent. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhongyin Zhao, Yuan Liu, Yikun Liu, Haicheng Wang, Le Tian, Xiao Zhou, Yangxiu You, Zilin Yu, Yang Yu, and Jie Zhou. POINTS-GUI-G: Gui-grounding journey. *arXiv preprint arXiv:2602.06391*, 2026.

Xurui Zhou, Gongwei Chen, Yuquan Xie, Zaijing Li, Kaiwen Zhou, Shuai Wang, Shuo Yang, Zhuotao Tian, and Rui Shao. HiconAgent: History context-aware policy optimization for GUI agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13028–13038, June 2026.

A Appendix Roadmap and Experimental Detail Index

The appendix is organized by evidentiary purpose. Appendix B reports repeatability across matched training seeds and conditional paired-bootstrap uncertainty, with full per-seed metrics in Table 7 and controlled prompt-scaffolding gaps in Table 8. Appendix C examines general grounding retention. Appendices E–I document training, annotation, dataset construction, prompts, and evaluation. Appendices J and K provide action-stratified diagnostics and traceable prediction cases, respectively. Appendix L concludes with the targeted post-SFT GRPO refinement and its full training and reproducibility specification.

B Statistical Robustness of Cognition and Prompt Ablations

We distinguish training variation across three matched seeds from evaluation variation measured by paired resampling of the reference checkpoint pair. The two analyses address different sources of uncertainty.

B.1 Per-seed Strict results and aggregate metrics

Table 4 lists the individual Strict results and paired differences. Tables 5 and 7 give the aggregate and full per-seed metrics. Table 8 summarizes the prompt-rule gain and the total gain over standard NoCoG under the same three training seeds. All metric values use a 0–100 scale, including Mean IoU multiplied by 100; differences use percentage points. Mean $\pm$ Std denotes the arithmetic mean and sample standard deviation over three seeds (denominator $3-1$), not a standard error or confidence interval. Aggregate and per-seed statistics are rounded separately; recomputing from two-decimal cells can differ by 0.01.

Strict in this matched comparison means Box-Strict. An actionable prediction must have the correct action, a box center inside the action-specific target, and $\mathrm{IoU} \geq 0.65$; a no-action prediction must contain `no_action` and the explicit box sentinel $[-1,-1,-1,-1]$. Cognition correctness is not an additional criterion. Cognition Accuracy is omitted here because neither NoCoG nor NoCoG+Rule is trained or asked to emit that field. Action and Strict use all 569 ID or 649 OOD examples; Mean IoU and Center Hit use the 357 ID or 388 OOD actionable examples. Center Hit alone does not require action correctness.

Table 4: Strict Accuracy for all three SFT configurations (upper panel) and paired differences in percentage points (lower panel). Each Δ is the left configuration minus the right; Mean $\pm$ Std uses three seeds.

Split	Configuration	Seed 42	Seed 88	Seed 123	Mean $\pm$ Std
ID	NoCoG (a,b)	77.50	78.91	80.49	78.97 ± 1.50
ID	NoCoG+Rule (a,b)	82.95	82.43	83.13	82.84 ± 0.37
ID	CoG SFT (c,a,b)	85.76	84.53	86.29	85.53 ± 0.90
OOD	NoCoG (a,b)	48.38	54.24	54.39	52.34 ± 3.43
OOD	NoCoG+Rule (a,b)	60.55	61.33	58.09	59.99 ± 1.69
OOD	CoG SFT (c,a,b)	65.33	64.41	64.41	64.71 ± 0.53
Split	**Comparison**	**Seed 42 Δ**	**Seed 88 Δ**	**Seed 123 Δ**	**Mean Δ**
ID	CoG $-$ NoCoG+Rule	$+2.81$	$+2.11$	$+3.16$	$+2.69$
ID	NoCoG+Rule $-$ NoCoG	$+5.45$	$+3.52$	$+2.64$	$+3.87$
ID	CoG $-$ NoCoG	$+8.26$	$+5.62$	$+5.80$	$+6.56$
OOD	CoG $-$ NoCoG+Rule	$+4.78$	$+3.08$	$+6.32$	$+4.73$
OOD	NoCoG+Rule $-$ NoCoG	$+12.17$	$+7.09$	$+3.70$	$+7.65$
OOD	CoG $-$ NoCoG	$+16.95$	$+10.17$	$+10.02$	$+12.37$

The cognition-specific CoG$-$NoCoG+Rule Strict difference is positive in every observed seed: 2.11–3.16 points on ID and 3.08–6.32 on OOD. The three-seed mean gain is 2.69/4.73 points on ID/OOD. The total CoG$-$NoCoG gap is also positive throughout and averages 6.56/12.37 points, while the intervening NoCoG+Rule$-$NoCoG comparison isolates the contribution of prompt scaffolding.

Table 5: All shared metrics across three training seeds: mean ± sample standard deviation. Bold marks the highest mean within each split and metric.

Split	Configuration	Action Acc. ↑	Mean IoU ↑	Center Hit ↑	Strict Acc. ↑
ID	NoCoG	87.64 ± 2.06	71.64 ± 1.66	81.60 ± 2.33	78.97 ± 1.50
ID	NoCoG+Rule	96.25 ± 0.62	78.65 ± 0.21	92.06 ± 0.43	82.84 ± 0.37
ID	CoG	$\mathbf{97.31 \pm 0.20}$	$\mathbf{80.13 \pm 0.56}$	$\mathbf{92.62 \pm 0.16}$	$\mathbf{85.53 \pm 0.90}$
OOD	NoCoG	78.22 ± 1.44	41.39 ± 0.81	61.94 ± 1.46	52.34 ± 3.43
OOD	NoCoG+Rule	90.96 ± 1.91	53.86 ± 1.24	79.98 ± 1.32	59.99 ± 1.69
OOD	CoG	$\mathbf{93.89 \pm 0.24}$	$\mathbf{57.46 \pm 0.53}$	$\mathbf{83.51 \pm 0.68}$	$\mathbf{64.71 \pm 0.53}$

B.2 Paired bootstrap for the reference checkpoint pair

Bootstrap procedure. For each split, we perform 10,000 paired bootstrap resamples over evaluation records. Each replicate draws N indices uniformly with replacement, where $N = 569$ for ID and $N = 649$ for OOD. The same index vector is applied to the CoG and NoCoG+Rule predictions from training seed 42. For replicate r, we compute

$$\Delta_{42}^{(r)} = \frac{100}{N} \sum_{j=1}^{N} \left(s_{42,i_j^{(r)}}^{\mathrm{CoG}} - s_{42,i_j^{(r)}}^{\mathrm{NoCoG+Rule}} \right), \tag{3}$$

where s is the binary Box-Strict success indicator. The 2.5th and 97.5th percentiles, with linear interpolation, define the 95% interval. The denominator includes all resampled actionable and no-action records.

Table 6: Strict differences and 95% paired-bootstrap intervals for the reference checkpoint pair. Accuracy values are percentages; differences and interval endpoints are percentage points.

Split	Metric	CoG	NoCoG+Rule	Δ	95% paired-bootstrap CI
ID	Box-Strict	85.76	82.95	+2.81	[+1.05, +4.57]
OOD	Box-Strict	65.33	60.55	+4.78	[+2.16, +7.55]

Scope of inference. The resampling unit is a task record. Records derived from the same screenshot may be dependent. These record-level intervals quantify Box-Strict uncertainty for the reference checkpoint pair; the three-seed mean and standard deviation summarize training variation separately.

B.3 Full per-seed metrics

B.4 Cognition-Aware Prompt Scaffolding

Standard NoCoG uses the cognition-free target $(a^\star, b^\star)$, training records, output schema, and optimization settings, together with a direct request for the appropriate action and location. NoCoG+Rule changes only the prompt by explicitly listing the semantic labels and their action mapping. Thus, NoCoG+Rule versus NoCoG quantifies the contribution of cognition-aware prompt scaffolding rather than an output-format or target-space change. The gain occurs in every split, seed, and reported metric and is largest on OOD.

Intervention boundary. Standard NoCoG is not a task-agnostic clicking prompt: it still identifies the relevant semantic dimension and requires the same structured `action`/`bbox` output. What it removes is the explicit inventory of semantic labels and the step-wise mapping from each label to `like`, `forward`, or `no_action`. The canonical templates are documented in Appendix I. Cognition Accuracy is consequently undefined for both NoCoG and NoCoG+Rule because neither emits a cognition field. As in the matched comparison, Action and Strict use all 569/649 ID/OOD records, whereas Mean IoU and Center Hit use the 357/388 actionable records.

Table 7: Per-seed metrics for all three controlled SFT configurations. All values are percentages and use the same Box-Strict criterion. NoCoG and NoCoG+Rule share target (a, b) and differ only in the explicit policy-rule scaffold.

Split	Configuration	Seed	Action Acc.	Mean IoU	Center Hit	Strict Acc.
ID	NoCoG	42	88.05	71.11	82.35	77.50
ID	NoCoG	88	85.41	70.32	78.99	78.91
ID	NoCoG	123	89.46	73.50	83.47	80.49
ID	NoCoG+Rule	42	96.31	78.44	92.16	82.95
ID	NoCoG+Rule	88	95.61	78.86	91.60	82.43
ID	NoCoG+Rule	123	96.84	78.64	92.44	83.13
ID	CoG	42	97.19	80.29	92.72	85.76
ID	CoG	88	97.19	79.51	92.72	84.53
ID	CoG	123	97.54	80.59	92.44	86.29
OOD	NoCoG	42	77.81	40.51	63.14	48.38
OOD	NoCoG	88	77.04	41.53	60.31	54.24
OOD	NoCoG	123	79.82	42.12	62.37	54.39
OOD	NoCoG+Rule	42	91.99	54.67	81.44	60.55
OOD	NoCoG+Rule	88	88.75	54.48	78.87	61.33
OOD	NoCoG+Rule	123	92.14	52.43	79.64	58.09
OOD	CoG	42	93.84	58.02	84.28	65.33
OOD	CoG	88	94.14	56.98	82.99	64.41
OOD	CoG	123	93.68	57.38	83.25	64.41

Table 8: Across-seed mean gains over standard NoCoG (percentage points). NoCoG+Rule−NoCoG isolates prompt-side rule scaffolding; CoG−NoCoG is the total gap across prompt- and target-side interventions.

Split	Comparison	ΔAction	ΔMean IoU	ΔCenter Hit	ΔStrict
ID	NoCoG+Rule − NoCoG	+8.61	+7.01	+10.46	+3.87
ID	CoG − NoCoG	**+9.67**	**+8.49**	**+11.02**	**+6.56**
OOD	NoCoG+Rule − NoCoG	+12.74	+12.47	+18.04	+7.65
OOD	CoG − NoCoG	**+15.67**	**+16.07**	**+21.57**	**+12.37**

Across-seed consistency. The NoCoG+Rule-minus-NoCoG Strict gap is positive for every paired seed: ID gains are $+5.45$, $+3.52$, and $+2.64$ points for seeds 42, 88, and 123; OOD gains are $+12.17$, $+7.09$, and $+3.70$ points. Table 8 summarizes the aggregate gaps. Their larger and uniformly positive OOD values support the interpretation that policy decomposition is particularly valuable when visual layout and platform conventions change. The prompt-matched CoG-versus-NoCoG+Rule comparison then holds the policy-explicit instruction fixed to isolate target-side cognition supervision.

B.5 Inference-time Cognition Prompting Analysis

In addition to training-time cognition supervision, we conduct a complementary inference-time analysis to examine whether models can utilize explicit semantic intermediate information without parameter updates.

For each evaluated model, we compare two input settings: a standard prompt without cognition information (NoCog) and an augmented prompt containing the semantic cognition state before the required GUI action (Cog). Model parameters remain unchanged in both settings. This analysis differs from the matched SFT experiment in Section 4, where cognition is learned through supervised training.

Table 9 reports the performance changes from NoCog to Cog prompting. The results show that explicit cognition information improves interaction performance for different models, although the magnitude of the improvement varies substantially. For example, Qwen3-VL-8B obtains larger gains than GPT-5.6 Luna, suggesting that models differ in their ability to exploit structured semantic intermediate representations.

These results show that supplied cognition can help frozen models translate content interpretation into GUI actions. They complement the matched SFT comparison, which measures the effect of learning the cognition field.

Table 9: Inference-time cognition prompting analysis on frozen models. Values represent percentage-point changes from the NoCog prompt to the Cog prompt. Model parameters are unchanged.

Model	Split	ΔAction	ΔIoU	ΔHit	ΔStrict
GPT-5.6 Luna	ID	+1.76	+1.35	+1.40	+4.21
	OOD	+0.92	+0.36	+1.03	+0.46
Doubao-Seed-2.1-turbo	ID	+0.88	+11.00	+2.24	+16.52
	OOD	+0.77	+6.94	+3.10	+8.48
Qwen3-VL-8B	ID	+14.23	+13.17	+10.08	+18.63
	OOD	+23.57	+7.40	+13.40	+18.80

C ScreenSpot-Mobile Capability Retention

Table 10 reports the complete ScreenSpot-Mobile comparison between SoGUI-SFT-CoG and its Qwen3-VL-8B initialization. Overall click accuracy changes by +0.80 percentage points. The changes are +1.42 on Android, +0.84 on iOS, +2.37 on icons, -1.92 on Shop, and -0.35 on text. Both checkpoints achieve 100% parse success and coordinate validity; strict JSON compliance is 98.80% for SoGUI-SFT-CoG and 100% for Qwen3-VL-8B. Figure 3 summarizes the broader ScreenSpot comparison.

Table 10: Detailed ScreenSpot-Mobile results on the same 501 examples. Click Acc. intervals are the reported 95% confidence intervals; Parse, Coord., and Strict JSON are output-validity rates.

Model	Slice	N	Hits	Click Acc.	95% CI	Parse	Coord.	Strict JSON
SoGUI-SFT-CoG	Overall	501	475	94.81	[92.50, 96.43]	100.00	100.00	98.80
	Android	211	206	97.63	[94.57, 98.98]	100.00	100.00	98.10
	iOS	238	224	94.12	[90.37, 96.46]	100.00	100.00	99.58
	Shop	52	45	86.54	[74.73, 93.32]	100.00	100.00	98.08
	Icon	211	187	88.63	[83.63, 92.24]	100.00	100.00	99.05
	Text	290	288	99.31	[97.52, 99.81]	100.00	100.00	98.62
Qwen3-VL-8B	Overall	501	471	94.01	[91.58, 95.77]	100.00	100.00	100.00
	Android	211	203	96.21	[92.70, 98.07]	100.00	100.00	100.00
	iOS	238	222	93.28	[89.36, 95.82]	100.00	100.00	100.00
	Shop	52	46	88.46	[77.03, 94.60]	100.00	100.00	100.00
	Icon	211	182	86.26	[80.96, 90.26]	100.00	100.00	100.00
	Text	290	289	99.66	[98.07, 99.94]	100.00	100.00	100.00

D EXPANDED LIMITATIONS

The cognition-to-action mappings are controlled benchmark policies rather than models of actual user behavior. Because each cognition label maps to one action, action selection is not an independent capability in the present benchmark. The benchmark covers three cognition dimensions and six platforms, and held-out platforms do not represent every form of visual or cultural distribution shift. Screenshots are kept disjoint across splits, as detailed in Appendix H. The matched ablation includes three seeds, but this is still a small sample of training variability, especially for standard NoCoG on OOD and ID CoG across the present three training seeds. The reported bootstrap intervals condition on seed 42 and do not form a joint interval over training and evaluation uncertainty. The example-level bootstrap does not adjust for dependence among records from the same screenshot. Human performance, free-form chain-of-thought controls, and semantically unrelated structured-prefix controls remain useful future comparisons.

E SFT TRAINING PROTOCOL

All three SFT conditions initialize from Qwen3-VL-8B-Instruct and use the same 2,698 Social GUI records. Images, task policies, actions, bounding boxes, record order, and optimization settings are held fixed. The CoG target is $(c^\star, a^\star, b^\star)$; both cognition-free targets are $(a^\star, b^\star)$. Each checkpoint is evaluated with the output schema used during its own training. All configurations are trained with seeds 42, 88, and 123; comparisons pair checkpoints with the same training seed. The broad baseline table, ScreenSpot analysis, error diagnosis, and paired bootstrap use the reference checkpoints; the robustness tables report all three seeds.

Standard NoCoG requests an action and location without enumerating the intermediate semantic label-to-action mapping. NoCoG+Rule uses the same target, records, seed set, and optimization settings but explicitly provides that rule in the prompt, which is also used by CoG. This design prevents target-schema or data-composition changes from confounding the prompt-scaffolding comparison and preserves a prompt-matched CoG–NoCoG+Rule pair.

Key training settings. We use LoRA with rank 8 and `lora_target=all`, while additionally training the visual projection/merger modules (`vision_proj` and `merger`). Training runs for three epochs with a learning rate of 5×10^{-5}, a per-device batch size of 2, and 8 gradient-accumulation steps. The image preprocessing budget is capped at 2,700,000 pixels. These are training settings; they should not be confused with the inference configuration in Appendix I.

F ANNOTATION PROTOCOL AND AGREEMENT

Table 11 reports pre-adjudication inter-annotator agreement on balanced samples of 500 examples per cognition dimension. Disagreement counts reflect the independent annotations; all disputed cases are adjudicated before the benchmark labels are finalized.

Table 11: Annotation agreement before adjudication ($n = 500$ per cognition dimension).

Dimension	Samples	Unanimous	Disagree.	Fleiss' κ
Affective valence	500	457 (91.4%)	43	**0.886**
Source credibility	500	421 (84.2%)	79	**0.833**
Message framing	500	407 (81.4%)	93	**0.764**

We further summarize the task-specific labeling criteria and adjudication procedure. For affective valence, annotators consider affective language, punctuation, emoji, and visible emotional cues. Source-credibility labels are assigned from visible cues, including identity, verification status, publisher type, and related interface signals. For message framing, annotators determine whether the content emphasizes gains or losses rather than relying on isolated lexical sentiment.

Once the cognition label is finalized, the corresponding action is generated deterministically from the controlled benchmark policy rather than independently annotated. Target boxes are verified separately against the intended GUI controls.

G Matched-Control Interpretation

The reference CoG model's Cognition Accuracy is 97.19% on ID and 93.84% on OOD, so cognition errors account for 2.81% and 6.16% of all examples. Because the emitted action exactly follows the controlled cognition-to-action mapping, Action Accuracy equals Cognition Accuracy. The gaps between Cognition Accuracy and Strict Accuracy are therefore 11.43 points on ID and 28.51 points on OOD. These are examples with correct semantics but unsuccessful end-to-end localization under the strict Center-Hit and IoU requirements.

Across the three matched seeds, mean Strict Accuracy increases from 82.84 ± 0.37 to 85.53 ± 0.90 on ID and from 59.99 ± 1.69 to 64.71 ± 0.53 on OOD. These gains support a contribution from the structured target beyond access to the Social GUI records. The common E2E-Hit results for the reference pair are 92.97%/94.02% for NoCoG+Rule/CoG on ID and 84.28%/87.21% on OOD. The paired-bootstrap intervals support a positive Strict difference under example resampling; E2E-Hit is reported separately as a point estimate.

H Benchmark Construction and Dataset Statistics

H.1 Dataset Composition and Split Protocol

The benchmark contains 3,916 task records constructed from screenshots of six social-media platforms. Each record pairs a screenshot with one task-specific cognition dimension, its cognition label, the action induced by the corresponding controlled policy, and a target bounding box when the policy requires an interaction. A single screenshot may serve as the interface context for multiple task records; consequently, the number of task records exceeds the number of unique screenshots. All records derived from the same screenshot are grouped together during data splitting to avoid cross-split leakage.

Table 12 summarizes the benchmark partitions. Training and ID evaluation use Weibo, TikTok, and Instagram, whereas the held-out-platform OOD partition contains only X, RedNote, and WeChat Channels. The OOD split therefore introduces platform-level visual and layout shifts rather than merely resampling screenshots from the training platforms.

Table 12: Dataset partitions. Percentages are relative to all 3,916 task records. Unique-screenshot counts are reported for the released evaluation partitions.

Split	Platforms	Records	Share	Unique screenshots
Training	Weibo, TikTok, Instagram	2,698	68.90%	–
ID evaluation	Weibo, TikTok, Instagram	569	14.53%	452
Held-out OOD	X, RedNote, WeChat Channels	649	16.57%	520
Total	Six platforms	3,916	100.00%	–

The 569 ID records consist of 230 affective-valence, 219 source-credibility, and 120 framing examples. The 649 OOD records consist of 253 affective-valence, 264 source-credibility, and 132 framing examples. Table 13 reports the complete evaluation distribution. Across the combined evaluation set, affective valence and source credibility each account for 39.66% of the records, while framing accounts for 20.69%. The action distribution includes both interactive and non-interactive cases: 38.83% of evaluation records require no action, 47.04% require liking, and 14.12% require forwarding.

Always-no-action reference. A constant predictor that emits `no_action` together with the native sentinel is correct exactly on the no-action records. Its Action Accuracy and E2E-Hit are therefore $212/569 = 37.26\%$ on ID and $261/649 = 40.22\%$ on OOD ($473/1218 = 38.83\%$ combined). When the box sentinel $[-1, -1, -1, -1]$ is emitted, the same values are obtained under the matched Box-Strict criterion. This analytical reference is derived directly from the label distribution and is not an additional trained model.

Table 13: Cognition-dimension and action distributions in the evaluation partitions. Each cell reports count (percentage within the corresponding split).

Split	Records	Affective valence	Credibility	Framing	No action	Like	Forward
ID	569	230 (40.42%)	219 (38.49%)	120 (21.09%)	212 (37.26%)	283 (49.74%)	74 (13.01%)
OOD	649	253 (38.98%)	264 (40.68%)	132 (20.34%)	261 (40.22%)	290 (44.68%)	98 (15.10%)
Combined	1,218	483 (39.66%)	483 (39.66%)	252 (20.69%)	473 (38.83%)	573 (47.04%)	172 (14.12%)

Table 14 further decomposes the OOD split. All three held-out platforms contain both actionable and no-action cases. The difference between task-record counts and unique-screenshot counts arises because a single screenshot may be associated with multiple task records; such records remain grouped and never cross split boundaries.

Table 14: Composition of the held-out-platform OOD evaluation set.

Platform	Records	Images	Affective valence	Credibility	Framing	Actionable
X	231	188	88	99	44	139
RedNote	254	211	105	105	44	157
WeChat Channels	164	121	60	60	44	92
Total	649	520	253	264	132	388

H.2 Controlled Cognition-to-Action Policies

Action labels are generated deterministically from cognition labels rather than independently annotated. This design enables separate analysis of semantic recognition and GUI localization while keeping the intended interaction objectively verifiable under each benchmark policy. Table 15 lists the complete mappings.

Because the action is deterministically derived from the cognition label, Action Accuracy should be interpreted as a policy-consistency diagnostic rather than an independent learned capability. Equality between Cognition Accuracy and Action Accuracy for a policy-consistent model therefore follows directly from the benchmark definition. We retain Action Accuracy to expose malformed or policy-inconsistent predictions.

Table 15: Controlled cognition-to-action mappings used by the benchmark.

Cognition dimension	Cognition label	Target action
Affective valence	Positive	Like
Affective valence	Negative	No action
Source credibility	Official organization	Forward
Source credibility	Positive individual	Like
Source credibility	Low-credibility account	No action
Message framing	Gain-oriented	Like
Message framing	Loss-oriented	No action

H.3 Annotation Workflow and Quality Control

Annotation follows a cognition-first workflow aligned with the task formulation.

Step 1: Screenshot inspection. Annotators inspect the complete input instance, including the screenshot, accompanying post text when available, visible account information, and task-relevant visual cues. Labels are assigned only from evidence observable in the provided instance, without relying on external knowledge about the author or event.

Step 2: Cognition labeling. The assigned cognition dimension determines the annotation criterion. Affective valence is judged from textual and visual emotional cues. Source-credibility labels are assigned from visible cues, including identity, verification status, publisher type, and related interface signals. Message framing is judged from whether the content emphasizes gains or losses. Annotators record the cognition label before considering the resulting GUI action.

Step 3: Deterministic action assignment. After the cognition label is fixed, the target action is derived according to Table 15. The action is therefore not a separate subjective annotation. This ordering reduces the risk of inferring cognition labels backward from preferred interaction behaviors.

Step 4: Target localization. For actionable records, annotators identify the corresponding like or forward control and record its bounding box, normalized to $[0, 1000]$ along both axes. For no-action records, the target box is set to the sentinel `[-1,-1,-1,-1]`.

Step 5: Validation and adjudication. Each record is checked for an allowed cognition label, consistency between cognition and policy-derived action, valid coordinate range, and agreement between the action and the annotated GUI control. Target boxes are separately verified.

For the annotation-quality audit, three annotators independently label a balanced sample of 500 examples for each cognition dimension before adjudication. Disagreements are resolved through discussion before the final labels are fixed. The resulting agreement statistics are reported in Table 11.

H.4 Screenshot-Level Splitting and Leakage Control

Splitting is performed at the screenshot level before dimension-specific task records are expanded. All records derived from the same screenshot therefore inherit the same split assignment, preventing repeated use of a screenshot across training and evaluation partitions. Training and ID evaluation use non-overlapping screenshot sets from the same three platforms, while OOD evaluation uses three held-out platforms that are absent from both training and ID evaluation.

I Prompts and Evaluation Protocol

This section specifies the prompts, structured output space, baseline and Oracle conditions, parsing rules, and coordinate transformations used in our evaluation. Figure 5 first illustrates the form of the benchmark using representative, visually complete examples.

I.1 Box-Output Prompt Templates

The following templates specify the box-output condition. In the policy-explicit CoG and NoCoG+Rule conditions, models receive the screenshot and original semantic task instruction, including its decision rule, without ground-truth cognition or localization. The standard NoCoG baseline below omits the explicit label-to-action steps while retaining the same task semantics and cognition-free output target. The literal `<image>` token below denotes the image supplied through the model's multimodal interface. Dataset records retain their stored Chinese or English wording. For readability, we give the complete canonical English rendering of each semantic template; the label set, decision mapping, coordinate range, and JSON keys are identical across languages. In the framing task, `[POST_TEXT]` is replaced by the text associated with that record. The templates expose the benchmark's compositional structure: valence and framing each map a binary semantic judgment to liking or abstention, whereas credibility adds a third state and a distinct Share target. Read alongside standard NoCoG, they show which decision steps are supplied by the instruction and which must be inferred from the post.

Affective valence.

> `<image>` Please analyze the emotional bias of the post in the image and take action accordingly. For posts with positive emotional content, perform a `like` action, find the position of the Like button, and output its normalized bounding-box coordinates (0–1000). For posts with negative emotional content, perform no action and output the bounding box as $[-1, -1, -1, -1]$. Output only a JSON

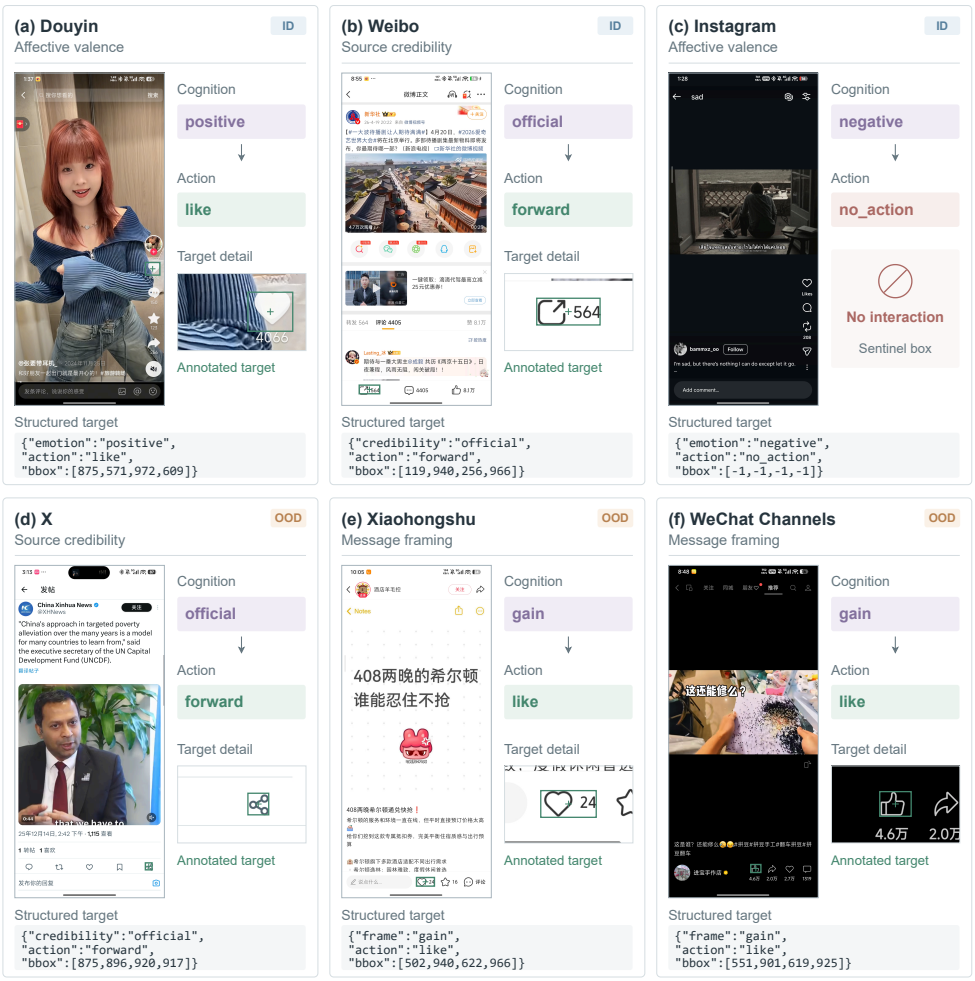

Figure 5: Representative SoGUI examples across all six platforms and three social-semantic tasks. The ID platforms are TikTok (shown as Douyin in the source screenshots), Weibo, and Instagram; the held-out OOD platforms are X, RedNote, and WeChat Channels. Affective valence, source credibility, and message framing are each shown twice. Each example pairs a cognition label with its induced action and a normalized target. Green boxes and cross markers denote the annotated GUI target; `no_action` uses the sentinel box $[-1, -1, -1, -1]$.

object in the following format; the actionable box is an example and must be replaced by the position in the given image:

```
{"emotion":"negative or positive",
  "action":"no_action or like",
"bbox":[-1,-1,-1,-1] or [875,571,972,609]}.
```

Source credibility.

`<image>` Please evaluate the credibility of the information source based on the profile picture, account name, and post content, and act accordingly. For official and authoritative government or institutional accounts, perform a `forward` (Share) action and output the normalized bounding box (0–1000) of the Share button. For positive and constructive individual users, perform a `like` action and output the normalized box of the Like button. For anonymous marketing accounts or negative, low-credibility accounts, perform no action and output

$[-1,-1,-1,-1]$. Output only a JSON object in the following format; the box is an example and must be replaced by the position in the given image:

{"credibility":"official or positive_individual or low",
"action":"forward or like or no_action",
"bbox":[881,793,969,825] or [-1,-1,-1,-1]}.

Message framing.

<image> Determine whether the supplied post text uses a positive, gain-oriented frame or a negative, loss-oriented frame, and act at the corresponding location in the screenshot. Post text: [POST_TEXT]. For a gain-oriented post, perform a like action, locate the Like button, and output its normalized bounding box (0–1000). For a loss-oriented post, perform no action and output $[-1,-1,-1,-1]$. Output only a JSON object in the following format; the actionable box is an example and must be replaced by the position in the given image:

{"frame":"gain or loss",
"action":"like or no_action",
"bbox":[567,940,629,965] or [-1,-1,-1,-1]}.

Standard NoCoG prompt. For each cognition dimension, standard NoCoG replaces the corresponding policy-explicit template above with the following canonical instruction:

<image> Determine the appropriate action from the post's [SEMANTIC DIMENSION] and locate the corresponding interface control. Output only {"action":"like/forward/no_action", "bbox":[x1,y1,x2,y2]}.

Here [SEMANTIC DIMENSION] is instantiated as affective content, source credibility, or narrative framing; the framing records retain their associated [POST_TEXT]. The instruction intentionally does not enumerate the label set or its action mapping. The supervision target remains exactly $(a^\star, b^\star)$, so this intervention tests policy decomposition rather than a new output representation.

The resulting structured spaces are summarized in Table 16. Cognition labels are not free-form: the evaluator requires an exact match to one member of the corresponding set.

Table 16: Task-specific cognition and action spaces.

Task	JSON key	Cognition labels	Induced actions
Affective valence	emotion	positive, negative	like, no_action
Source credibility	credibility	official, positive_individual, low	forward, like, no_action
Message framing	frame	gain, loss	like, no_action

I.2 Baseline and Matched-Control Settings

We evaluate four GUI-specialized baselines: Qwen2.5-VL-3B-UI-R1-E, GUI-R1-7B, POINTS-GUI-G, and ScaleCUA-7B. The main comparison preserves native point outputs for these agents; the box-output templates above do not impose a bounding-box requirement on them. All output families share the same task semantics and cognition-to-action policy. Native coordinates are converted only to a common coordinate frame for evaluation. We do not synthesize boxes from points, repair malformed predictions, or inject missing cognition/action fields into model outputs.

The reported box-output inference setup uses greedy decoding, with sampling disabled, one beam, KV caching, and a limit of 128 newly generated tokens. It uses bfloat16 weights, SDPA attention, a maximum of 4,700,000 image pixels, and a minimum of 3,136 image pixels. These settings describe local inference, not the training image budget or closed-source API decoding parameters. Inference

Table 17: Recorded settings of the native-point normal E2E runs analyzed in Appendix J. All four use greedy decoding, bfloat16, and a minimum of 3,136 image pixels. The same settings apply to their ID and OOD runs.

Model	Max. new tokens	Max. image pixels	Attention
Qwen2.5-VL-3B-UI-R1-E	1,024	12,845,056	SDPA
GUI-R1-7B	1,024	2,097,152	SDPA
POINTS-GUI-G	256	4,700,000	Flash Attention 2
ScaleCUA-7B	256	2,109,744	SDPA

errors are retained as empty predictions and remain in the evaluation denominator. The native-point runs have model-specific image budgets and output limits; their recorded settings are given in Table 17, not inferred from the box-output configuration.

For protocol-fair E2E-Hit, every model receives the screenshot and original task instruction without Oracle cognition. On actionable examples, success requires both the correct action and a predicted click point inside the action-specific ground-truth box. On no-action examples, success requires `no_action` and the native point sentinel $(-1, -1)$. All 569 ID and 649 OOD examples remain in the denominator. Point-native GUI agents are scored directly; for a box-emitting model, the box center is used as its click point. Ground-truth boxes are visible only to the evaluator. E2E-Hit therefore does not require a predicted box, IoU, strict JSON compliance, or a cognition label. The no-action box sentinel has center $(-1, -1)$, so it provides the corresponding abstention point for box-output models. E2E-Hit is not a sum of Action Accuracy and Center Hit: it tests their conjunction on actionable records and correct abstention on no-action records. Because Center Hit uses only actionable records, its percentage can be lower than E2E-Hit.

Table 18: Output-family scope of the main evaluation. E2E-Hit uses the same action-and-point success rule across all rows.

Output family	Point used by E2E-Hit	Additional diagnostics
Native-point GUI agents	Native predicted point	Mean IoU and Strict undefined; no box is synthesized.
General-purpose box-output models	Predicted box center	Mean IoU and original schema-aware Strict retained.
SoGUI SFT variants	Predicted box center	Common Box-Strict, without a cognition-correctness requirement.

The cognition-free matched control is trained and evaluated with the same images, semantic decision rules, action space, target annotations, record order, optimization, and decoding setup. Its target and prompt request only `action` and `bbox`:

```
{"action":"like/forward/no_action", "bbox":[x1,y1,x2,y2]}.
```

Which actions are available still depends on the task-specific decision rule; only the explicit cognition field is removed from both training and evaluation. This NoCoG+Rule condition is therefore a training-level matched control, not a prompt ablation applied to the cognition-trained model.

Standard NoCoG retains this cognition-free target and all training factors but removes the explicit decision mapping from the prompt as specified above. Consequently, CoG versus NoCoG+Rule isolates target-side cognition supervision, whereas NoCoG+Rule versus NoCoG isolates instruction-side policy scaffolding.

I.3 Oracle Conditions

Oracle cognition. The normal and Oracle runs use the same model, screenshot, original instruction, decision rule, decoding parameters, and native action protocol. The Oracle condition changes only the information supplied to the model by appending:

> `[Oracle cognition]` The ground-truth cognition label is `[DIMENSION] = [LABEL]`. Treat this label as given; do not infer it again. Apply the decision rule in the original instruction using this label.

Here `[DIMENSION]` is one of `emotion`, `credibility`, or `frame`. We report this paired analysis for GUI-R1-7B and POINTS-GUI-G. Its common metric is Cognition-conditioned Hit (Cog-Hit): for an actionable target, the predicted point (or predicted box center) must lie inside the ground-truth box; for a non-actionable target, the model must return `no_action` or its native $(-1, -1)$ sentinel. All samples remain in the denominator. Unlike E2E-Hit, Cog-Hit does not require the action-and-hit conjunction and accepts either no-action indicator for non-actionable targets.

Table 19: Paired Oracle-cognition diagnostic. Cog-Hit is reported in percent; Δ is Oracle minus Normal in percentage points.

Model	Split	Normal	Oracle	Δ
GUI-R1-7B	ID	65.03	82.25	+17.22
GUI-R1-7B	OOD	65.64	76.73	+11.09
POINTS-GUI-G	ID	45.17	60.98	+15.82
POINTS-GUI-G	OOD	71.03	89.37	+18.34

Oracle grounding. To isolate localization from semantic decision making, we remove `no_action` examples and replace the original task by an explicit action command: `Click the Like button for the post.` for `like`, and `Click the Share button for the post.` for `forward`. The source label is named `forward`, but the annotated target is the Share control (including the rightmost Share icon on X), rather than X's separate Repost control. Cognition labels are retained only as metadata and are not exposed in the Oracle grounding prompt. This produces 357 ID and 388 OOD actionable examples. Center Hit Accuracy is the primary common metric because the evaluated native protocols emit points and therefore do not define predicted boxes for IoU.

I.4 Output Parsing and Scoring

For box-output schema diagnostics, we first remove optional Markdown fences, then take the substring from the first opening brace to the last closing brace and parse it with a strict JSON parser. A prediction is parse-valid only if the result is a JSON object containing both `action` and `bbox`, and the box is a four-element numeric list. The cognition key is additionally required for cognition-conditioned outputs but is not required for NoCoG or NoCoG+Rule outputs. Cognition and action labels are compared as exact strings. Extra prose outside the JSON object is ignored, but malformed JSON, unavailable geometry, nonnumeric boxes, and inference failures receive no geometry credit. We apply no semantic repair or coordinate post-correction. Schema-specific cognition requirements do not affect E2E-Hit, which scores the extractable action and point. Native-point predictions are evaluated through their native action and coordinate fields; missing or unparseable required predictions fail the corresponding metric without being removed from the denominator.

For actionable ground truth, Mean IoU and Center Hit are computed over the actionable subset. Center Hit is true when the center of the predicted box lies within the ground-truth box, including its boundary; for point-native outputs the predicted point is tested directly. Mean IoU is not defined for point-only outputs. The matched SFT Box-Strict criterion is computed over all examples. An actionable prediction is strictly correct only when the action matches, Center Hit is true, and $\mathrm{IoU} \geq 0.65$. A `no_action` example is strictly correct only when the action matches and the effective predicted box equals $[-1, -1, -1, -1]$. Cognition correctness is not required for either matched SFT condition. The other box-output baselines retain their original schema-aware Strict scores, including their schema-specific requirements; these are supplementary diagnostics rather than the protocol-fair metric for comparisons across all output families.

I.5 Coordinate Convention and Native-Protocol Conversion

Ground-truth boxes use normalized `xyxy` coordinates in $[0, 1000]^4$. For an original screenshot of width W and height H, pixel coordinates are normalized as

$$(x_1', y_1', x_2', y_2') = 1000 \left(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H} \right). \tag{4}$$

Valid actionable boxes satisfy $0 \le x_1' \le x_2' \le 1000$ and $0 \le y_1' \le y_2' \le 1000$. The sentinel $[-1, -1, -1, -1]$ is reserved exclusively for `no_action`.

The box-output prompt directly requests this normalized box, so no model-specific conversion is needed for those coordinates. Native-point end-to-end and Oracle evaluations preserve point outputs and convert their coordinate frame only for common scoring. For a unit-normalized point, $(x', y') = (1000x, 1000y)$; for a pixel point on an image of width W_p and height H_p, $(x', y') = (1000x/W_p, 1000y/H_p)$. Here W_p, H_p refer to the image frame in which the model reports its point, not automatically the original screenshot size. The native abstention sentinel $(-1, -1)$ is recognized before this conversion and is preserved. The existing Oracle coordinate conventions are summarized in Table 20.

Table 20: Coordinate handling in the Oracle evaluations.

Model	Native output	Conversion to $[0, 1000]$
POINTS-GUI-G	point in $[0, 1]^2$	multiply both coordinates by 1000
UI-R1-E	pixel click point	$(1000x/W_p,\ 1000y/H_p)$
GUI-R1-7B	pixel click point	$(1000x/W_p,\ 1000y/H_p)$

For a point-native prediction, Center Hit tests the point directly. For a box-native prediction, it tests the predicted box center. For actionable predictions, coordinates outside $[0, 1000]$, reversed boxes, and unparseable native outputs are counted as failures rather than clipped into the valid range.

J Action-Stratified Evaluation

Scope and denominators. We re-aggregate the existing per-example predictions underlying the main results. The SFT rows use the reference checkpoints (seed 42). Each example is assigned to its *ground-truth* action, rather than its predicted action. The ID split contains 283 `like`, 74 `forward`, and 212 `no_action` examples; the OOD split contains 290, 98, and 261, respectively. Action Accuracy, E2E-Hit, and Box-Strict use the full count of the indicated class. Mean IoU and Center Hit are undefined on the `no_action` class, denoted by a dash. For that class, E2E-Hit requires both the correct action and the explicit native abstention sentinel. An action label alone does not receive end-to-end credit.

The less frequent action remains difficult. For OOD `forward`, CoG improves E2E-Hit from 48.98% under standard NoCoG to 59.18% and Box-Strict from 38.78% to 57.14%. Nevertheless, its 92.86% Action Accuracy is substantially higher than its 59.18% E2E-Hit: selecting the right action does not guarantee a correct click. The `like` results show a different residual limitation. OOD CoG Center Hit is 92.76%, but Box-Strict is only 44.48%, exposing box-quality errors that a point-only metric does not measure. Improvements are not uniform across every cell: within the prompt-matched pair, ID `like` Action Accuracy decreases from 97.88% to 97.53%, while E2E-Hit remains 94.35% for both targets.

Connection to the main-table gains. Relative to standard NoCoG, the ID Box-Strict gain decomposes into 25 additional `like`, 10 additional `forward`, and 12 additional `no_action` successes. Thus $(25 + 10 + 12)/569$ yields the reported $+8.26$ percentage points on ID. On OOD, the corresponding net gains are 57, 18, and 35, yielding $(57 + 18 + 35)/649 = +16.95$ percentage points. These class totals recover the main-table differences.

Table 21: Controlled SFT results by ground-truth action (%). All rows use the common Box-Strict definition without a cognition-correctness gate. NoCoG is the standard baseline; NoCoG+Rule is the prompt-matched cognition-free control. Higher is better for every metric.

Split	GT action	Target	N	Action	Mean IoU	Center Hit	E2E-Hit	Box-Strict
ID	`like`	NoCoG	283	93.99	78.40	90.46	90.46	78.09
ID	`like`	NoCoG+Rule	283	97.88	81.71	94.35	94.35	83.04
ID	`like`	CoG	283	97.53	83.48	94.35	94.35	86.93
ID	`forward`	NoCoG	74	58.11	43.24	51.35	51.35	37.84
ID	`forward`	NoCoG+Rule	74	95.95	65.94	83.78	83.78	50.00
ID	`forward`	CoG	74	98.65	68.12	86.49	86.49	51.35
ID	`no_action`	NoCoG	212	90.57	–	–	90.57	90.57
ID	`no_action`	NoCoG+Rule	212	94.34	–	–	94.34	94.34
ID	`no_action`	CoG	212	96.23	–	–	96.23	96.23
OOD	`like`	NoCoG	290	82.07	41.56	67.93	67.93	24.83
OOD	`like`	NoCoG+Rule	290	96.21	57.88	92.07	92.07	38.97
OOD	`like`	CoG	290	96.21	60.12	92.76	92.76	44.48
OOD	`forward`	NoCoG	98	55.10	37.43	48.98	48.98	38.78
OOD	`forward`	NoCoG+Rule	98	88.78	45.18	50.00	50.00	50.00
OOD	`forward`	CoG	98	92.86	51.80	59.18	59.18	57.14
OOD	`no_action`	NoCoG	261	81.61	–	–	78.16	78.16
OOD	`no_action`	NoCoG+Rule	261	88.51	–	–	88.51	88.51
OOD	`no_action`	CoG	261	91.57	–	–	91.57	91.57

J.1 Cross-Model End-to-End Breakdown

We first summarize the prompt-matched CoG gain by action in Table 22; Table 23 then gives the common action-and-point metric for all eight models with verified per-example predictions. Native points are retained for GUI-specialized agents, and predicted box centers are used for box-emitting models.

Table 22: CoG minus prompt-matched NoCoG+Rule E2E-Hit by ground-truth action (percentage points), derived from Table 23.

Split	`like`	`forward`	`no_action`	All
ID	0.00	+2.71	+1.89	+1.05
OOD	+0.69	+9.18	+3.06	+2.93

Action prediction is not sufficient for abstention. For example, ScaleCUA's no-action Action Accuracy is 56.13% on ID and 56.32% on OOD, whereas its no-action E2E-Hit is 46.70% and 43.68%. The difference reflects the requirement for an explicit valid abstention point in E2E-Hit.

Audit and reproducibility. Re-aggregating saved per-example responses reproduces the main-table Action Accuracy, Center Hit, and E2E-Hit for all 16 model/split runs, as well as the box metrics for the four box-output models. We aggregate integer success counts before rounding and parse native coordinates using the same evaluation-time conversion as the main results. For the illustrated cases, Appendix K gives the dataset row and saved prediction line used for each panel.

Table 23: E2E-Hit by ground-truth action (%). The All column is the sample-count-weighted result over the three classes, not their unweighted mean. No-action success requires the native sentinel and correct action.

Model	Split	`like`	`forward`	`no_action`	All
Qwen2.5-VL-3B-UI-R1-E	ID	39.58	5.41	69.34	46.22
GUI-R1-7B	ID	79.51	51.35	63.21	69.77
POINTS-GUI-G	ID	68.20	75.68	39.62	58.52
ScaleCUA-7B	ID	66.78	47.30	46.70	56.77
Qwen3-VL-8B	ID	75.97	77.03	73.58	75.22
SoGUI-SFT-NoCoG	ID	90.46	51.35	90.57	85.41
SoGUI-SFT-NoCoG+Rule	ID	94.35	83.78	94.34	92.97
SoGUI-SFT-CoG	ID	94.35	86.49	96.23	94.02
Qwen2.5-VL-3B-UI-R1-E	OOD	33.10	1.02	60.92	39.45
GUI-R1-7B	OOD	69.31	34.69	64.75	62.25
POINTS-GUI-G	OOD	31.03	40.82	59.00	43.76
ScaleCUA-7B	OOD	71.03	44.90	43.68	56.09
Qwen3-VL-8B	OOD	66.55	55.10	76.25	68.72
SoGUI-SFT-NoCoG	OOD	67.93	48.98	78.16	69.18
SoGUI-SFT-NoCoG+Rule	OOD	92.07	50.00	88.51	84.28
SoGUI-SFT-CoG	OOD	92.76	59.18	91.57	87.21

The OOD gain is concentrated in `forward`, but this remains the hardest CoG action slice: its E2E-Hit is 59.18% on 98 examples, compared with 92.76% for OOD `like`. Within the prompt-matched pair, OOD `forward` Action Accuracy rises from 88.78% to 92.86%, Center Hit from 50.00% to 59.18%, and Box-Strict from 50.00% to 57.14%. The improvement thus spans the action decision, target hit, and box-quality checks, while leaving substantial room on this less frequent action.

Failure mass behind the aggregate. The class counts give a more concrete view of that limitation. On OOD, CoG succeeds on 269 of 290 `like`, 58 of 98 `forward`, and 239 of 261 `no_action` examples, yielding $566/649 = 87.21\%$ overall E2E-Hit. The 83 remaining failures comprise 21 `like`, 40 `forward`, and 22 `no_action` examples. Although `forward` contributes only 15.1% of OOD examples, it accounts for 48.2% of these failures. This concentration explains why the strong aggregate result coexists with a substantial tail-action deficit and pinpoints where further interaction and localization work has the most headroom.

Different localization bottlenecks by action. The two actionable classes also differ in how point success translates to box quality. For OOD `like`, CoG reaches 92.76% E2E-Hit but only 44.48% Box-Strict, a 48.28-point gap. For OOD `forward`, the corresponding values are 59.18% and 57.14%, a 2.04-point gap. Thus, the large residual loss for OOD `like` appears mainly when the full box is tested, whereas OOD `forward` already loses many successes at the action-plus-point stage. The pattern also changes across splits: on ID `forward`, the E2E-Hit–Box-Strict gap is 35.14 points (86.49% versus 51.35%). For ID `like`, E2E-Hit remains 94.35% in the prompt-matched pair while Box-Strict rises from 83.04% to 86.93%, showing a box-quality gain even without an additional successful click. These contrasts separate action-and-point failures from errors in predicted box extent.

K Traceable Qualitative Prediction Cases

These three cases illustrate distinct failure modes using original screenshots, ground-truth annotations, and saved model predictions. Colored boxes and points are visualization overlays; coordinates are normalized to $[0, 1000]$. The SFT outputs come from the reference checkpoints. In these panels, *No-Cog SFT* denotes the prompt-matched NoCoG+Rule checkpoint. The dataset row and prediction line under each panel identify its saved record.

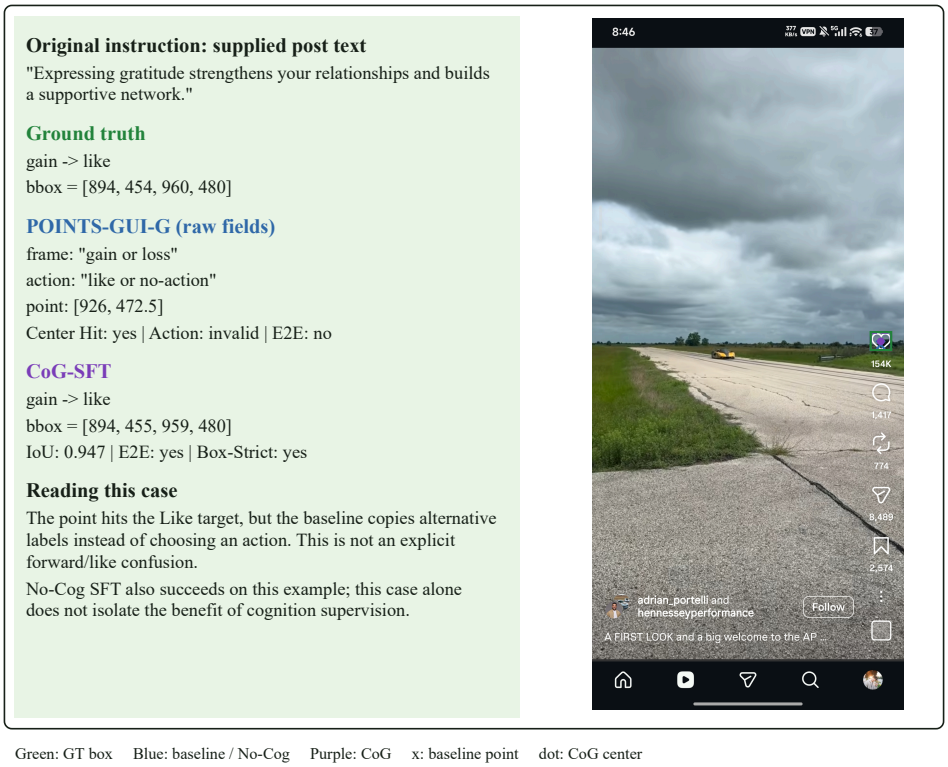

Figure 6: **A: a correct point does not resolve an action label.** On ID framing example `c_0004.jpg`, POINTS-GUI-G predicts a point inside the Like target but returns the alternatives `like or no-action` instead of a valid action. Center Hit therefore passes while Action Accuracy and E2E-Hit fail. CoG-SFT emits `gain`, `like`, and a valid box with IoU 0.947. NoCoG-SFT also succeeds on this example.

Recorded prediction and scoring. The saved POINTS-GUI-G response contains the following native fields (line breaks added for display):

```
{"frame":"gain or loss",
  "action":"like or no-action",
  "point":[926,472.5]}
```

The point lies in the ground-truth Like box $[894, 454, 960, 480]$, but `like or no-action` does not select a single executable action. The evaluator therefore records Center Hit without Action Accuracy or E2E-Hit. By contrast, the saved CoG-SFT output is:

```
{"frame":"gain","action":"like",
  "bbox":[894,455,959,480]}
```

This resolves the action and passes both E2E-Hit and Box-Strict.

Provenance. ID dataset row 123 (zero-based index 122), `test_id/images/c_0004.jpg`; POINTS-GUI-G normal ID shard 0, prediction line 62; matched CoG and NoCoG ID prediction line 123. All line numbers are one-based.

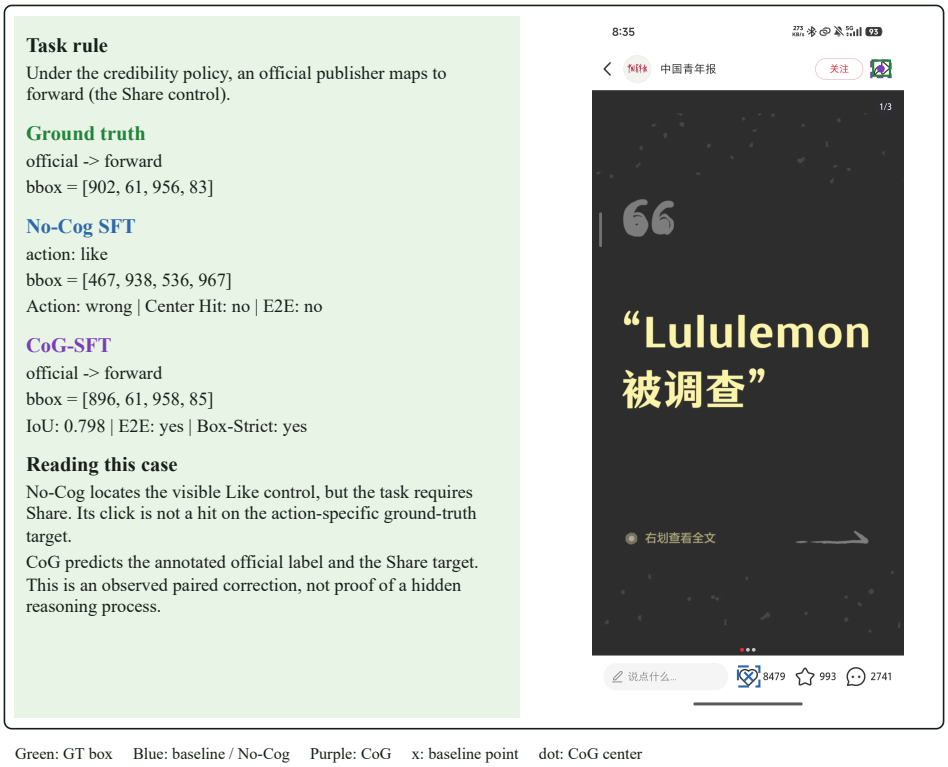

Figure 7: **B: locating a visible control is not the same as following the task policy.** On OOD credibility example `e_0061.jpg`, the annotated official source calls for `forward`, whose target is the Share control at the top right. NoCoG-SFT predicts `like` and locates the bottom Like control. This is an explicit action error, and its point does *not* hit the action-specific ground-truth target. CoG-SFT predicts `official`, selects `forward`, and hits Share with IoU 0.798, passing both E2E-Hit and Box-Strict.

Provenance. OOD dataset row 403 (zero-based index 402), `test_ood/images/e_0061.jpg`; matched CoG and NoCoG OOD prediction line 403.

Exact recorded predictions. The raw JSON objects are shown below (line breaks added for display only).

```
NoCoG:
{"action":"like", "bbox":[467,938,536,967]}

CoG:
{"credibility":"official", "action":"forward",
  "bbox":[896,61,958,85]}
```

The ground-truth box is $[902, 61, 956, 83]$.



Figure 8: **C: correct decision, imprecise box.** On ID credibility example `b_0032.jpg`, CoG-SFT predicts `positive_individual` and `like`; the blue *No-Cog SFT* trace (the prompt-matched NoCoG+Rule control) also selects `like`. Both predicted box centers fall inside the ground-truth Like target and pass E2E-Hit. Their box IoUs are 0.327 (CoG-SFT) and 0.507 (No-Cog SFT), both below the 0.65 Box-Strict threshold. The case shows why successful action and clicking do not imply precise box localization.

Provenance. ID dataset row 110 (zero-based index 109), `test_id/images/b_0032.jpg`; prompt-matched NoCoG+Rule and CoG ID prediction line 110. The GT box is $[760, 937, 823, 965]$; NoCoG+Rule predicts $[759, 937, 875, 967]$ and CoG predicts $[727, 938, 904, 967]$.

L GRPO Refinement for Residual Localization Errors

L.1 Motivation and Scope

After cognition-supervised SFT, semantic decision errors are substantially reduced, while localization remains the dominant residual error. We therefore investigate whether a lightweight post-SFT GRPO stage can improve localization without materially changing the learned cognition-to-action policy. The reported result evaluates the complete refinement pipeline, including training-set rollout diagnosis, correction-focused sample selection, label-preserving visual counterfactuals, a semantic-gated continuous geometry reward, field-level token credit, and one epoch of GRPO. This integrated study directly tests whether the diagnosis in Finding 4 can guide an effective targeted optimization stage, while the reward and credit formulations below make the role of each training signal explicit.

L.2 DETERMINISTIC TRAINING-SET ROLLOUT AUDIT

All selection decisions are computed from the original training split. Neither ID nor OOD evaluation examples are included in the selection procedure.

L.2.1 ROLLOUT PROTOCOL

Table 24: Training-set rollout protocol used for GRPO curriculum construction.

Item	Value
Original training examples	2,698
Greedy completions per example	1
Stochastic completions per example, K	16
Total greedy completions	2,698
Total stochastic completions	43,168
Sampling temperature	1.2
Top-p	0.95
Maximum new tokens	64
Retry budget for truncated outputs	256
Initially/finally truncated completions	0 / 0
Empty completions	0
Maximum image pixels	4,700,000
Random seed	42

A completion is *ready* when its output is parseable, its cognition and action fields are correct, and an actionable prediction contains a valid normalized box. Let $\delta_{\mathrm{IoU}} = 0.05$ and $\tau_{\mathrm{anchor}} = 0.90$. These fixed heuristics are used only for training-set diagnosis and are not selected on ID or OOD evaluation results. For an actionable example, a sampled candidate s dominates the greedy prediction g when either

$$\begin{aligned} &\mathrm{IoU}(s) - \mathrm{IoU}(g) \geq 0.05 \quad \text{and} \quad H_{\mathrm{center}}(s) \geq H_{\mathrm{center}}(g), \\ \text{or} \quad &H_{\mathrm{center}}(s) > H_{\mathrm{center}}(g) \quad \text{and} \quad \mathrm{IoU}(s) \geq \mathrm{IoU}(g). \end{aligned} \tag{5}$$

Table 25 summarizes the mutually exclusive audit outcomes used to construct the subsequent curriculum.

Table 25: Mutually exclusive training-set rollout-audit buckets.

Bucket	Deterministic rule	Count
Geometry-improvable	At least one ready stochastic completion dominates the greedy prediction.	324
Geometry-contrast	At least two ready candidates, minimum IoU below 0.90, and either IoU range at least 0.05 or different center-hit outcomes.	465
Semantic repair	Greedy output is not ready, while at least one stochastic output is ready because a semantic or format decision can be repaired.	65
Semantic/format stability	Greedy output is ready, but at least one stochastic output is not ready.	158
Stable replay candidate	Greedy output is ready and is either a correct no-action prediction or has center hit with IoU at least 0.90.	1,624
Low contrast	Valid candidates exist but none of the preceding contrast or anchor rules applies.	38
No eligible sampled candidate	No stochastic completion is ready.	24
Total		**2,698**

The complete audit identifies 1,012 informative examples: 324 geometry-improvable, 465 geometry-contrast, 65 semantic-repair, and 158 semantic/format-stability examples. An initial task/action-stratified 20% replay construction adds 253 stable examples, producing an auditable pool of 1,265 examples and 1,153 unique screenshots. This pool is an intermediate diagnostic artifact, not the final GRPO curriculum.

L.2.2 CORRECTION-FOCUSED CURRICULUM

For the reported refinement experiment, the broad geometry-contrast bucket is reserved rather than trained in full. The final curriculum contains all direct-correction and semantic/format-stability examples, plus a smaller task/action-stratified stable replay set. Stable replay is selected deterministically with seed 42 by round-robin sampling over task/action strata and then screenshots.

Table 26: Composition of the correction-focused GRPO curriculum before visual counterfactual augmentation.

Curriculum component	Rows	Unique images	Actionable	No action
Geometry-improvable	324	324	324	0
Semantic repair	65	65	16	49
Direct correction subtotal	**389**	**389**	**340**	**49**
Semantic/format stability	158	158	23	135
Stable replay	61	48	35	26
Final selected curriculum	**608**	**594**	**398**	**210**

The 61 replay rows constitute approximately 10% of the final 608-row curriculum. Its task/action composition is 185 Emotion/like, 75 Emotion/no-action, 102 Credibility/forward, 103 Credibility/like, 126 Credibility/no-action, 8 Frame/like, and 9 Frame/no-action rows.

L.3 LABEL-PRESERVING VISUAL COUNTERFACTUALS

We augment the correction-focused curriculum using transformations that alter GUI appearance or geometry without changing the post content, cognition label, or cognition-to-action policy. For every counterfactual example,

$$c^*_{\mathrm{cf}} = c^*, \qquad a^*_{\mathrm{cf}} = a^*, \qquad b^* \mapsto b^*_{\mathrm{cf}}, \tag{6}$$

where the target box is transformed consistently with the visual operation. The transformations therefore instantiate semantic invariance under GUI variation. All replacement buttons are cropped from real training screenshots and matched by action type; no generated icons or evaluation images are used. Table 27 summarizes the four transformations.

Table 27: Label-preserving visual counterfactual operations.

Operation	Construction	Rows
Position	Exchange action-bar cells and translate the target box by the same offset.	300
Target scale	Scale the local target region by 0.88 or 1.12 and apply the same affine transform to the box.	55
Real style replacement	Replace the target with a same-action button crop from another training screenshot; retain the recipient target region as the box.	48
Resolution and whitespace	Isotropically scale the screenshot by 0.90, 0.93, or 0.96, place it on a fixed canvas, and transform the box accordingly.	96
Counterfactual total		**499**

The reported run uses 608 selected original examples and 499 counterfactual examples, for 1,107 training rows: 655 `like`, 194 `forward`, and 258 `no_action`. For no-action examples, the cognition and action labels are retained and the target remains exactly $[-1, -1, -1, -1]$. Construction validates image existence, normalized box bounds, and no-action sentinels.

L.4 SEMANTIC-GATED CONTINUOUS GEOMETRY REWARD

Let B_p and B_g denote the predicted and ground-truth boxes, and let $c_p = (x_p, y_p)$ be the predicted box center. The distance from c_p to the ground-truth region is

$$\begin{aligned} d_x &= \max(x_{g1} - x_p, 0, x_p - x_{g2}), \\ d_y &= \max(y_{g1} - y_p, 0, y_p - y_{g2}), \\ d_{\text{out}}(c_p, B_g) &= \sqrt{d_x^2 + d_y^2}. \end{aligned} \tag{7}$$

After normalization by the ground-truth diagonal d_g, the smooth center quality term is

$$S_{\text{center}}(B_p, B_g) = \exp\left[-\left(\frac{d_{\text{out}}(c_p, B_g)}{d_g}\right)^2\right]. \tag{8}$$

A center inside the target receives $S_{\text{center}} = 1$, while an outside center is penalized smoothly according to its distance. For actionable examples, the geometry reward is

$$R_{\text{geo}} = \mathbb{I}_{\text{format}}\mathbb{I}_{\text{cognition}}\mathbb{I}_{\text{action}}\ \text{IoU}(B_p, B_g)S_{\text{center}}(B_p, B_g). \tag{9}$$

For no-action examples, the geometry term equals one only when the predicted action is `no_action` and the predicted box is exactly $[-1, -1, -1, -1]$, subject to the same format and cognition prerequisites. The training reward does not use the evaluation-time IoU ≥ 0.65 threshold; that threshold is used only by Box-Strict, whereas GRPO receives continuous geometry feedback over valid IoU.

L.5 FIELD-LEVEL TOKEN CREDIT

The structured response contains format, cognition, action, and bounding-box fields. Instead of broadcasting one sequence-level advantage to every generated token, we compute a group-relative signal per field and route it through mutually exclusive token masks:

$$A_t = M_{\text{fmt}}(t)A_{\text{fmt}} + M_{\text{cog}}(t)A_{\text{cog}} + M_{\text{act}}(t)A_{\text{act}} + M_{\text{box}}(t)A_{\text{geo}}. \tag{10}$$

Incorrect candidates remain eligible for their own field comparison, allowing negative group-relative advantages. Geometry is eligible only when format, cognition, and action prerequisites are all correct. Tokens not assigned to a parsed JSON value are assigned to the format field.

For field f in rollout group g, rewards are centered over eligible rollouts:

$$A_{\text{raw}}(g, k, f) = E(g, k, f)\left(R(g, k, f) - \text{mean}_E[R(g, :, f)]\right). \tag{11}$$

All active field advantages in the optimization batch share one RMS scale,

$$\rho = \sqrt{\text{mean}(A_{\text{raw}}^2)}, \qquad A(g, k, f) = \frac{A_{\text{raw}}(g, k, f)}{\max(\rho, 10^{-6})}. \tag{12}$$

This shared normalization preserves relative differences among fields without separate manually tuned additive reward weights, yielding a structured optimization signal aligned with the semantic roles in SoGUI's output schema.

L.6 GRPO TRAINING CONFIGURATION

Table 28: Configuration of the post-SFT GRPO refinement.

Hyperparameter	Value
Initialization	SoGUI-SFT-CoG
Training rows	1,107
Epochs / optimization steps	1 / 1,107
GPUs / precision	2× NVIDIA A100 40 GB / BF16
Distributed optimizer	DeepSpeed ZeRO-2
Attention implementation	Eager
Rollouts per prompt	16
Sampling temperature / top-p	1.2 / 0.95
Maximum prompt / completion length	7,680 / 64 tokens
Maximum image pixels	4,700,000
Per-device batch size / gradient accumulation	1 / 8
Learning rate	5×10^{-7}
Scheduler / warmup ratio	Cosine / 0.03
GRPO KL coefficient, β	0.03
Reward scaling	Shared batch RMS; library sequence-level scaling disabled
LoRA rank / alpha / dropout	8 / 16 / 0.0
LoRA target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
Additional trainable modules	`vision_proj`, `merger`
Random/data seed	42

L.7 REFINEMENT RESULTS

The same evaluator and predefined Box-Strict criterion are used for SFT and GRPO. An actionable prediction must have the correct action, its predicted center inside the ground-truth target, and $\mathrm{IoU} \geq 0.65$; the no-action rule remains unchanged.

Table 29: Post-SFT GRPO comparison. Values and differences are percentages and percentage points, respectively.

Method	Split	Cog. Acc.	Action Acc.	Mean IoU	Center Hit	Strict Acc.
SoGUI-SFT-CoG	ID	97.19	97.19	80.29	**92.72**	85.76
SoGUI-SFT-CoG + GRPO	ID	**97.19**	**97.19**	**80.34**	92.44	**86.47**
Δ	ID	0.00	0.00	+0.05	-0.28	**+0.71**
SoGUI-SFT-CoG	OOD	93.84	93.84	58.02	84.28	65.33
SoGUI-SFT-CoG + GRPO	OOD	**94.14**	**94.14**	**58.78**	**84.54**	**66.56**
Δ	OOD	+0.30	+0.30	**+0.76**	+0.26	**+1.23**

GRPO preserves ID cognition and action accuracy while improving ID Strict Accuracy by 0.71 points. On held-out-platform OOD data, Mean IoU, Center Hit, and Strict Accuracy improve by 0.76, 0.26, and 1.23 points, respectively. These results are consistent with the intended role of the refinement pipeline in reducing residual localization errors, and the larger OOD Strict gain further supports its value under cross-platform visual variation.

Figure 9 traces a complete progression on a held-out X framing example. The base model predicts the wrong cognition label (`loss`) and consequently abstains. Cognition-supervised SFT recovers the correct `gain` label and `like` action, but its predicted box reaches an IoU of 0.6499, narrowly below the Box-Strict threshold of 0.65. GRPO preserves the corrected semantic decision and refines the box to an IoU of 0.8047, converting the near-threshold miss into a strict success. This example directly illustrates the complementary progression diagnosed by our benchmark: semantic supervision enables the intended interaction, after which targeted reinforcement learning improves residual geometry.

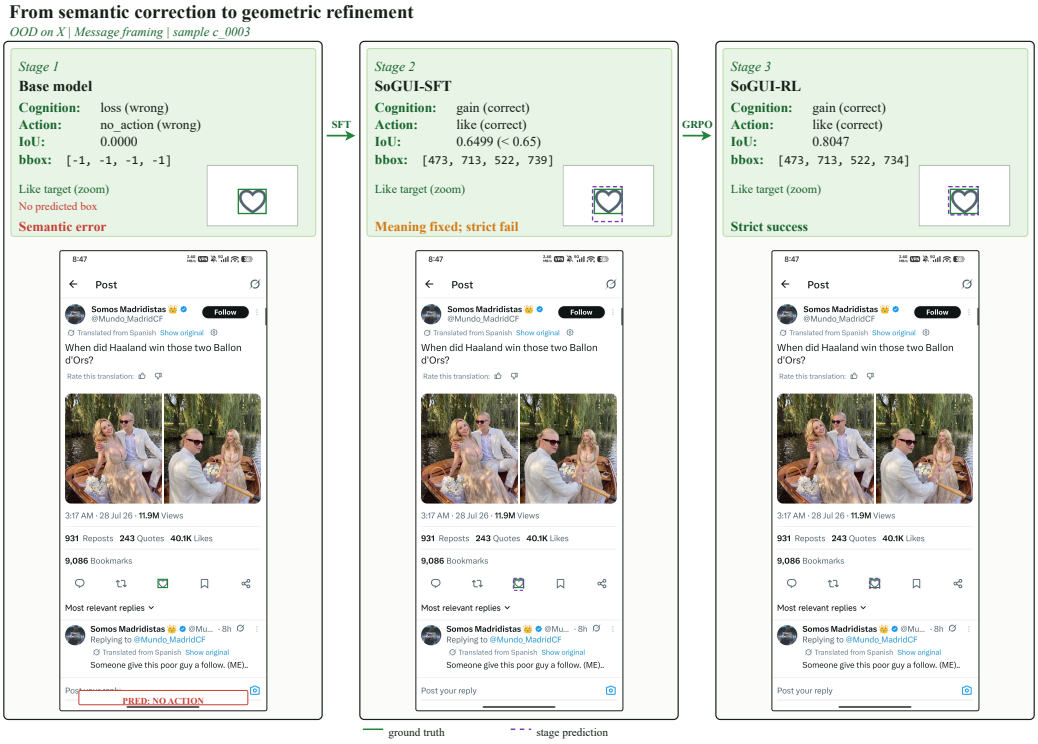

Figure 9: **From semantic correction to geometric refinement.** Base → SoGUI-SFT-CoG → SoGUI-SFT-CoG+GRPO on a held-out OOD X example. SFT first recovers the correct cognition and interaction; localization-targeted GRPO then improves the predicted box from IoU 0.6499 to 0.8047. All displayed outputs are taken from the corresponding prediction logs without modification.

L.8 Evaluation protocol and interpretation

Rollout-selection thresholds are fixed on training data, while held-out ID/OOD scores use the predefined Box-Strict criterion. The GRPO reward is continuous and does not reuse the evaluation-time IoU ≥ 0.65 threshold. Under this protocol, improvements on both splits—especially the $+1.23$-point OOD Strict gain—show that localization-directed post-SFT refinement transfers beyond the layouts used for curriculum construction.

The metric pattern further localizes the source of the gain. On ID, cognition and action remain fixed at 97.19%, Center Hit stays above 92%, and Strict improves; on OOD, Mean IoU, Center Hit, and Strict all increase while cognition is preserved. The update therefore improves box quality without relaxing the semantic gate or evaluation rule. The qualitative progression in Figure 9 provides the corresponding example-level view: supervised cognition first enables the intended interaction, after which GRPO supplies the residual geometric correction.

Together, the paired metrics and traced example show how the benchmark's cognition-versus-localization diagnosis can guide a concrete post-SFT update.