SafeGround: Know When to Trust GUI Grounding Models via Uncertainty Calibration

Qingni Wang [*1 2] **Yue Fan** [*2] **Xin Eric Wang** [1]

Abstract

Graphical User Interface (GUI) grounding aims to translate natural language instructions into executable screen coordinates, enabling automated GUI interaction. Nevertheless, incorrect grounding can result in costly, hard-to-reverse actions (e.g., erroneous payment approvals), raising concerns about model reliability. In this paper, we introduce SAFEGROUND, an uncertainty-aware framework for GUI grounding models that enables risk-aware predictions through calibrations before testing. SAFEGROUND leverages a distribution-aware uncertainty quantification method to capture the spatial dispersion of stochastic samples from outputs of any given model. Then, through the calibration process, SAFEGROUND derives a test-time decision threshold with statistically guaranteed false discovery rate (FDR) control. We apply SAFEGROUND on multiple GUI grounding models for the challenging ScreenSpot-Pro benchmark. Experimental results show that our uncertainty measure consistently outperforms existing baselines in distinguishing correct from incorrect predictions, while the calibrated threshold reliably enables rigorous risk control and potentials of substantial system-level accuracy improvements. Across multiple GUI grounding models, SAFEGROUND improves system-level accuracy by up to 5.38% percentage points over Gemini-only inference.

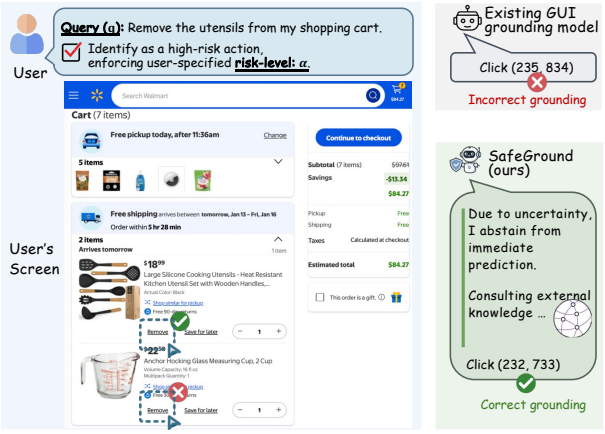


Figure 1. While existing models may commit costly errors on hard-to-undo actions (e.g., checkout), SAFEGROUND detects high uncertainty and defers the decision via cascading. This mechanism explicitly limits the risk of erroneous actions to a user-specified tolerance.

1. Introduction

Graphical User Interface (GUI) grounding is a critical component for autonomous GUI agents, enabling vision-language models (VLMs) to translate natural language instructions into executable screen coordinates (Nguyen et al., 2025; Cheng et al., 2024). Recent advances have substantially improved grounding accuracy across diverse GUI environments, making it increasingly feasible to deploy such agents in real-world applications (Fan et al., 2025; Hong et al., 2024). However, in practical GUI interactions, a single incorrect grounding can trigger costly and hard-to-reverse actions, including erroneous payment approvals or irreversible system configurations (Zhang et al., 2025). Despite these risks, existing GUI grounding models typically output only point predictions, offering no indication of when a prediction is unreliable or should be deferred (Gawlikowski et al., 2022; Hu et al., 2023) as shown in Figure 1.

The aforementioned limitation of existing GUI grounding models motivates the incorporation of uncertainty quantification (UQ) to enable safer decision-making. However, existing UQ techniques are poorly suited for GUI grounding and remain largely underexplored in this setting (Zhang et al., 2025). In particular, prior approaches suffer from several key limitations. (1) Uncertainty derived from model probabilities or logits (Hendrycks & Gimpel, 2017) assumes

[*]Equal contribution [1]University of California, Santa Barbara [2]University of California, Santa Cruz. Correspondence to: Xin Eric Wang <ericxwang@ucsb.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

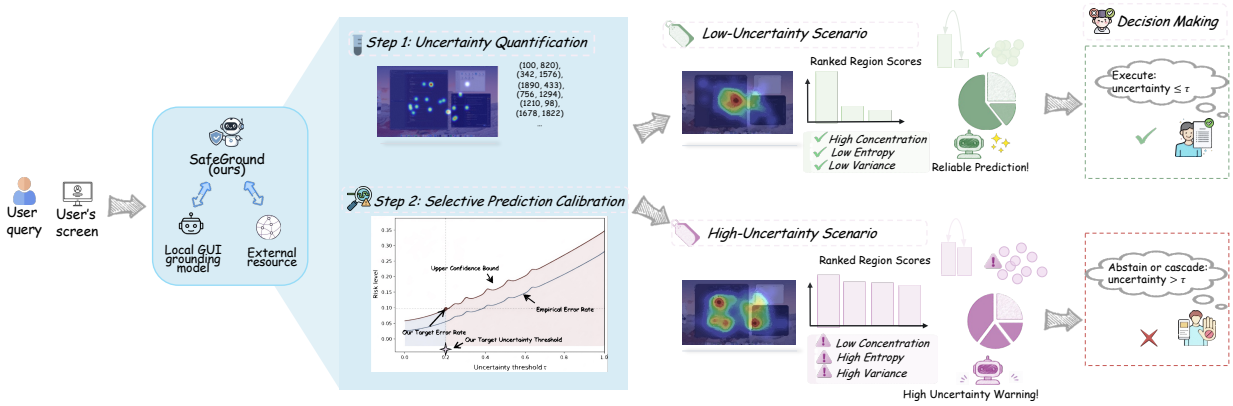


Figure 2. Overview of SAFEGROUND. Given a GUI input, the model performs multiple stochastic grounding samples to estimate predictive uncertainty. An uncertainty threshold τ is calibrated on a held-out set under a user-specified risk level (i.e, the maximum error rate). At test time, predictions with uncertainty $\leq \tau$ are executed directly, while high-uncertainty cases are abstained or cascaded. Low-uncertainty cases exhibit concentrated region scores, low entropy, and low variance, whereas high-uncertainty cases show dispersed predictions and trigger safety-aware deferral.

access to internal model states, making it infeasible for black-box vision-language models commonly used in GUI agents (Ye et al., 2024; Wang et al., 2025b). (2) Verbalized self-assessment (Kadavath et al., 2022) relies on strong instruction-following behavior and often fails when models do not explicitly reason about confidence. (3) Approaches that estimate uncertainty using ground-truth regions, such as Zhang et al. (2025), require annotation and cannot be applied at inference time. (4) Existing methods focus on producing uncertainty scores alone, without specifying how predictions should be acted upon at deployment time (e.g., whether to accept, defer, or abstain) despite this decision being critical in high-stakes GUI interactions (Geifman & El-Yaniv, 2017; Wang et al., 2025c). Collectively, these limitations expose a clear gap between existing UQ approaches and the practical requirements of GUI grounding, where uncertainty must be reliable under limited model access and without test-time supervision (Lin et al., 2023).

To address these challenges, we introduce SAFEGROUND, an uncertainty-aware framework that enables risk-aware predictions for existing state-of-the-art GUI grounding models, without requiring access to model internals. Concretely, as shown in Figure 2, SAFEGROUND first quantifies the predictive uncertainty of grounding outputs from the spatial distribution of multiple stochastic grounding samples from the same model. Then, given the model outputs with estimated uncertainty, we adopt a Learn Then Test (LTT) calibration paradigm to select a decision threshold that rigorously controls the false discovery rate (FDR) of accepted grounding predictions. This calibration procedure provides finite-sample guarantees: with high probability, the proportion of incorrect predictions among all accepted actions does not exceed a user-specified risk level α. At inference time, SAFEGROUND enables a principled selective prediction mechanism. Predictions deemed reliable under the calibrated threshold are executed directly, while high-uncertainty cases are abstained from or deferred to stronger models for further processing. Furthermore, with the selective prediction, we realized the cascading inference, where even when the primary model's base accuracy is limited, we can further leverage external resource to aid the prediction, achieving strong system-level accurarcy.

We evaluate SAFEGROUND on the challenging ScreenSpot-Pro benchmark across multiple state-of-the-art GUI grounding models. Experimental results demonstrate that our proposed uncertainty measure consistently outperforms existing baselines in distinguishing correct from incorrect predictions. Especially, SAFEGROUND achieves reliable FDR control in practice and significantly improves overall system accuracy through selective deferral, validating its effectiveness for high-stakes GUI interaction scenarios. Empirically, SAFEGROUND demonstrates clear system-level accuracy gains across different risk levels. For instance, on ScreenSpot-Pro, uncertainty-aware cascading with Holo1.5-7B achieves 58.66% accuracy at risk level 0.34, improving over Gemini-only inference by 5.38% points. The project website is available at `https://safeground-ericlab.github.io`. Our contributions can be summarized as follows:

- We propose SAFEGROUND, the first framework for uncertainty-aware selective GUI grounding with finite-sample risk guarantees via calibration.
- We introduce distribution-aware uncertainty quantification that leverage the spatial dispersion and concentration of stochastic grounding predictions.

- We demonstrate that SAFEGROUND with uncertainty-calibrated selective prediction enables reliable FDR control and improves system-level accuracy in cascading inference on the ScreenSpot-Pro benchmark.

2. Related Work

2.1. GUI Grounding

GUI grounding maps natural language instructions to actionable interface elements or click locations in graphical user interfaces (Nguyen et al., 2025; Fan et al., 2025). Most existing GUI grounding methods formulate the problem as a text-based coordinate prediction task, where models generate point locations conditioned on the input screenshot and instruction (Chen et al., 2023; Wang et al., 2024a; Qin et al., 2025a). Recently, motivated by how humans interact with digital interfaces, GUI-Actor introduces an attention-based formulation that aggregates spatial evidence into a single grounding decision (Wu et al., 2025b). These methods have achieved strong empirical accuracy across diverse GUI environments. However, most existing approaches produce deterministic point predictions and do not explicitly model predictive uncertainty, limiting their ability to assess decision reliability or defer actions under high uncertainty.

2.2. Uncertainty Estimation

Uncertainty estimation is widely used to support reliable decision making in AI systems by quantifying the confidence of model predictions (Liu et al., 2025). In large language models, uncertainty has also been derived from probabilistic measures, semantic entropy, or verbalized self-reports (Hou et al., 2025; Wang et al., 2024b; Xu et al., 2025; Kuhn et al., 2023b). In GUI grounding, uncertainty estimation remains largely underexplored. Existing GUI grounding approaches typically rely on probabilistic uncertainty or verbalized uncertainty, both of which have been shown to be systematically miscalibrated, exhibiting a mismatch between predicted confidence and actual grounding accuracy (Zhang et al., 2025). This misalignment motivates uncertainty estimation methods that rely solely on model outputs while providing more reliable signals for downstream decision-making, as considered in our work.

2.3. Learn then Test Calibration

Learn Then Test (LTT) is a post-hoc calibration paradigm that separates model learning from statistical risk control (Angelopoulos et al., 2022). Given a fixed predictive model, LTT frames decision making as a hypothesis testing problem over a low-dimensional decision space, and uses held-out calibration data to identify parameters that satisfy user-specified risk constraints with finite-sample guarantees. Split conformal prediction (SCP) (Angelopoulos & Bates, 2022) follows this principle by leveraging data splitting and concentration-based confidence bounds to perform valid risk estimation. Prior work builds on this paradigm to enable reliable decision making in large foundation models (Jung et al., 2025; Wang et al., 2025a;c; Wang et al.). Our approach also builds on the LTT paradigm and extends it to GUI grounding through uncertainty-based calibration of spatial action decisions for the first time.

3. Methodology

3.1. Problem Formulation and Notations

Let the GUI grounding model be a function $f : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^2$, which takes a UI screenshot $x \in \mathcal{X}$ and a user instruction $q \in \mathcal{T}$ as input. Given an input pair (x, q), the model predicts a coordinate $\hat{y} = (\hat{u}, \hat{v}) \in \mathbb{R}^2$ on the screen. Although the model produces a single point prediction, the ground truth for a target UI element is typically provided as a spatial region on the screen, denoted by $B^* \subset \mathbb{R}^2$. A predicted coordinate is considered correct if and only if it falls within the ground-truth region, which we conclude as an admission function $A : \mathbb{R}^2 \times \mathcal{P}(\mathbb{R}^2) \to \{0, 1\}$ with 1 indicating a correct prediction:

$$A(\hat{y}, B^*) = \begin{cases} 1, & \text{if } \hat{y} \in B^*, \\ 0, & \text{otherwise.} \end{cases}$$

In current coordinate-based GUI grounding models, predictions are deterministic and are not accompanied by explicit uncertainty or confidence estimates, which leaves the trustworthiness of model outputs largely uncharacterized, and may cause users to place unwarranted trust in incorrect predictions, without any indication of potential failure.

3.2. Method Overview

To address this issue, we propose SAFEGROUND, an uncertainty-aware GUI grounding framework that can be integrated with diverse state-of-the-art GUI grounding models without requiring access to internal model states, as illustrated in Figure 2. SAFEGROUND introduce a user-specified risk level $\alpha \in (0, 1)$ that quantifies the maximum tolerable proportion of incorrect predictions, serving as a high-level control signal for how conservatively the system should behave. The risk level α is then translated into an uncertainty threshold τ through a calibration procedure. Specifically, the GUI grounding model's predictive uncertainty, $U(\hat{y}^{(\mathrm{MLG})}) \in \mathbb{R}$ for a prediction $\hat{y}^{(\mathrm{MLG})}$, is estimated by SAFEGROUND through sampling multiple additional predictions from the GUI grounding model given the same input. The larger values of such uncertainty score indicate lower reliability. A prediction $\hat{y}$ is correct if $U(\hat{y}) \leq \tau$ and rejected otherwise, in which case it is deferred to a stronger model. The threshold τ is chosen such that, among all admitted

predictions, the fraction of incorrect ones, measured by the admission function $A(\hat{y}, B^*)$, is controlled below α.

3.3. Uncertainty Quantification

We first quantify model uncertainty by analyzing the distributional properties of the ranked region scores. Then, three complementary uncertainty measures are introduced, where they are designed to capture complementary failure modes of GUI grounding: local ambiguity among competing targets, global dispersion of belief across regions, and lack of dominant spatial concentration.

Sampling-Based Spatial Distribution Construction To move beyond deterministic point predictions and capture the output distribution of GUI grounding models, we employ a Monte Carlo (Gal & Ghahramani) sampling strategy followed by spatial aggregation, drawing inspiration from attention-based aggregation mechanisms in (Wu et al., 2025b). Specifically, for each input (x, q), we perform K stochastic forward passes of the grounding model, generating a set of coordinates $\mathcal{S} = \{\hat{y}^{(i)}\}_{i=1}^{K}$, where $\hat{y}^{(i)} \in \mathbb{R}^2$.

These sampled coordinates are then projected onto a discretized screen grid to estimate a normalized local density map P, which empirically characterizes the spatial distribution of the model's predictions using only sampled outputs from the model. Intuitively, high density in a localized area indicates model consistency and thus low uncertainty. To establish object-level representations, we aggregate connected high-density patches in P into disjoint regions $\mathcal{R} = \{R_m\}_{m=1}^{M}$ through density-based clustering. Each region R_m is scored by its *average probability density*, denoted as S_m, serving as a proxy for the likelihood that the region corresponds to the intended UI element. Regions are further ranked such that $S_{(1)} \geq S_{(2)} \geq \cdots \geq S_{(M)}$. More implementation details are provided in the Appendix B.3.

Uncertainty Measurement 1. Top-Candidate Ambiguity (TA). To measure the distinctiveness of a certain prediction from a GUI grounding model, we compute the margin between the two leading candidates. A vanishing margin indicates that the model is uncertain between multiple plausible targets (e.g., two identical exit buttons), therefore, we propose the uncertanty score measured by top-candidate ambiguity:

$$U_{TA} = \begin{cases} 1 - \frac{S_{(1)} - S_{(2)}}{S_{(1)} + \epsilon}, & \text{if } M \geq 2 \\ \max(0.1, 1 - S_{(1)}), & \text{otherwise} \end{cases} \quad (1)$$

where ϵ ensures numerical stability. High U_{TA} signifies localized confusion at the decision boundary.

Uncertainty Measurement 2. Informational Dispersion (IE). We assess global uncertainty using the entropy of the region score distribution. To ensure a valid probabilistic interpretation, we induce a categorical distribution over the M regions:

$$\hat{p}_i = \frac{S_{(i)}}{\sum_{j=1}^{M} S_{(j)}}, \quad (2)$$

and then we define the uncertainty score based on information dispersion as the normalized entropy:

$$U_{\text{IE}} = -\frac{1}{\log M} \sum_{i=1}^{M} \hat{p}_i \log(\hat{p}_i + \epsilon). \quad (3)$$

Such measurement captures the dispersion of probability mass across regions; a high U_{IE} indicates that the model's confidence is fragmented, failing to converge on a single consistent hypothesis.

Uncertainty Measurement 3. Concentration Deficit (CD). While entropy assesses global disorder, we explicitly quantify the lack of focus with another uncertainty score U_{CD} by examining the quadratic concentration of the distribution:

$$U_{CD} = 1 - \sum_{i=1}^{M} \hat{p}_i^2 \quad (4)$$

Unlike entropy, U_{CD} is more sensitive to the dominance of the top candidates. Higher values of U_{CD} indicate a highly fragmented distribution, suggesting that the model lacks a clear spatial focus and distributes confidence across multiple interface regions.

Combined Uncertainty Score. Each uncertainty score captures a distinct aspect of predictive dispersion, and no single measurement is universally dominant across all models and scenarios. To obtain a unified and deployment-friendly uncertainty signal, we aggregate these three scores into a single one via a fixed weighted combination:

$$U_{COM}(\hat{y}) = w_{CD} \cdot U_{CD} + w_{IE} \cdot U_{IE} + w_{TA} \cdot U_{TA}. \quad (5)$$

We adopt a single set of weights across all models to preserve a plug-and-play interface without model-specific tuning.

3.4. Uncertainty Calibration for Selective Prediction

Although the proposed uncertainty measures capture predictive uncertainty, they cannot fully distinguish between correct and incorrect predictions. To enable user-specified deployment, we further introduce a selective prediction mechanism by calibrating a statistically rigorous decision threshold τ on the uncertainty score, such that, among all accepted predictions, the proportion of incorrect predictions does not exceed a desired level α.

Following prior SCP-based frameworks, we hold out a calibration set of N data points: $\mathcal{D}_{cal} = \{(x_i, q_i, B_i^*)\}_{i=1}^N$. For each calibration input pair (x_i, q_i), we produce $\hat{y}_i^{(MLG)}$ and quantify its uncertainty score $u_i = U\left(\hat{y}_i^{(MLG)}\right)$. Given a candidate threshold τ, we obtain the number of accepted predictions $\sum_i^N \mathbf{1}\{u_i \le \tau\}$, and the number of incorrect predictions $\sum_i^N \mathbf{1}\{u_i \le \tau, A(\hat{y}_i^{(MLG)}, B_i^*) = 0\}$. We then compute the false discovery rate (FDR) on $\mathcal{D}_{cal}$ under threshold τ:

$$\mathrm{FDR}_{cal}(\tau) = \frac{\sum_i^N \mathbf{1}\{u_i \le \tau, A(\hat{y}_i^{(MLG)}, B_i^*) = 0\}}{\sum_i^N \mathbf{1}\{u_i \le \tau\}} \quad (6)$$

To provide finite-sample FDR guarantees for the accepted samples at test time, we first introduce an auxiliary lemma.

Lemma 3.1 (Clopper–Pearson interval (Clopper & Pearson, 1934))**.** *Let $X \sim \mathrm{Bin}(n, p)$ be the number of successes in n i.i.d. Bernoulli trials with success probability p. For any $\delta \in (0, 1)$, define the Clopper-Pearson confidence interval*

$$\Big[p_L(X),\, p_U(X)\Big] = \Big[\mathrm{Beta}^{-1}\big(\tfrac{\delta}{2};\, X,\, n - X + 1\big),\; \mathrm{Beta}^{-1}\big(1 - \tfrac{\delta}{2};\, X + 1,\, n - X\big)\Big], \quad (7)$$

where $\mathrm{Beta}^{-1}(q; a, b)$ denotes the q-quantile from a beta distribution with shape parameters a and b. Then the interval has (at least) nominal coverage:

$$\mathbb{P}(p \in [p_L(X), p_U(X)]) \ge 1 - \delta. \quad (8)$$

In our setting, $X = \sum_i^N \mathbf{1}\{u_i \le \tau, A(\hat{y}_i^{(MLG)}, B_i^*) = 0\}$ and $n = \sum_i^N \mathbf{1}\{u_i \le \tau\}$. Since we focus on controlling the upper tail of the system FDR $R(\tau)$ (thereby constraining test-time FDR), based on Lemma 3.1, we construct a high-probability upper confidence bound, $\hat{\mathrm{FDR}}_{1-\delta}^{upper}(\tau)$, for $R(\tau)$, using its empirical estimate from the calibration data:

$$\begin{aligned}\hat{\mathrm{FDR}}_{1-\delta}^{upper}(\tau) &= \mathrm{Beta}(1 - \delta;\, X + 1,\, n - X) \\ &= \sup\{R : \Pr(\mathrm{Bin}(n, R) \le X) \ge \delta\}\end{aligned}, \quad (9)$$

where $\hat{\mathrm{FDR}}_{1-\delta}^{upper}$ guarantees

$$\Pr\left(R(\tau) \le \hat{\mathrm{FDR}}_{1-\delta}^{upper}(\tau)\right) \ge 1 - \delta. \quad (10)$$

Essentially, $\hat{\mathrm{FDR}}_{1-\delta}^{upper}(\tau)$ can be interpreted as the largest plausible value that the system FDR could take, given that an extremely small $\mathrm{FDR}_{cal}(\tau)$ is observed on the calibration set at significance level δ. If the true system FDR were to exceed this bound, then observing $\mathrm{FDR}_{cal}(\tau)$ in a single realization would be statistically impossible at the level δ. A formal proof of Eq. (10) is provided in Appendix A.

To rigorously constrain test-time FDR, we calibrate τ such that $\hat{\mathrm{FDR}}_{1-\delta}^{upper}(\tau)$ does not exceed the risk level α:

$$\hat{\tau} = \sup\{\tau : \hat{\mathrm{FDR}}_{1-\delta}^{upper}(\tau) \le \alpha\} \quad (11)$$

The choice of $\hat{\tau}$ maximizes the acceptance of model predictions (or minimizes the abstention rate), while maintaining marginal FDR control. For a test sample $(x_{test}, q_{test}, B_{test}^*)$ with the model prediction $\hat{y}_{test}^{(MLG)}$ and estimated uncertainty score $u_{test} = U\left(\hat{y}_{test}^{(MLG)}\right)$, by applying the calibrated decision threshold $\hat{\tau}$, we establish the following guarantee

$$\Pr\left(\Pr\left(A\big(\hat{y}_{test}^{(MLG)}, B_{test}^*\big) = 0 \mid u_{test} \le \hat{\tau}\right) \le \alpha\right) \ge 1 - \delta. \quad (12)$$

Cascading Inference. At inference time, for each test input (x_{test}, q_{test}), we first estimate the model uncertainty u_{test}, and then perform selective prediction and escalating:

- If $u_{test} \le \hat{\tau}$, we define the sample as "safe" and accept the prediction of the primary model.
- If $u_{test} > \hat{\tau}$, we flag the sample as "risky" and escalate the input to a stronger model to enhance performance.

4. Experiment

4.1. Experimental Settings

Models and Dataset We conduct our experiments over 6 GUI-grounding models, including Holo1.5 (Company, 2025), GUI-Actor (Wu et al., 2025a), UI-TARS-1.5 (Qin et al., 2025b), GTA1 (Yang et al., 2025): Holo1.5-3B, Holo1.5-7B, GUI-Actor-2VL-7B, GUI-Actor-2.5VL-7B, UI-TARS-1.5-7B and GTA1-7B. To assess reliability under high-stakes scenarios, we conduct all experiments on the challenging ScreenSpot-Pro (Li et al., 2025) benchmark. We further evaluate on OSWorld-G (Xie et al., 2025) to test generalization to realistic desktop GUI environments, with results in Appendix. Additional dataset details are provided in the Appendix B.1.

Evaluation Metrics To comprehensively evaluate both the discriminative ability of UQ methods and the reliability and effectiveness of SafeGround, we adopt four complementary metrics: Area Under Receiver Operating Characteristic (AUROC), Area Under Accuracy-Rejection Curve (AUARC), FDR, and power (Lin et al., 2024; Wang et al., 2025c). AUROC measures the ability of uncertainty estimates to distinguish correct from incorrect predictions, while AUARC evaluates whether prediction accuracy improves as high-uncertainty samples are progressively rejected. FDR quantifies the proportion of incorrect predictions among the accepted samples. Power measures

the proportion of correct samples that are retained after uncertainty-based selection, relative to the total number of correct samples. More details about the metrics can be found in Appendix B.2.

Hyperparameters For uncertainty estimation, we sample each input 10 times with the decoding temperature set to 1.0 to compute the corresponding UQ score. The most likely generation $\hat{y}_i^{(\mathrm{MLG})}$ is obtained by uniformly sampling one output from the generated candidates. Specifically, when computing UQ scores, we partition the input into patches with a patch size of 14 to obtain region-level scores S_i for uncertainty estimation. We repeat the random calibration–test split 100 times and report the mean and standard deviation (mean±std) over all runs. All confidence bounds are constructed at a significance level of $\delta = 0.05$. For the combined uncertainty score U_{COM}, we use a fixed weighting scheme $(w_{CD}, w_{IE}, w_{TA}) = (0.6, 0.2, 0.2)$ across all models.

4.2. Evaluation of Uncertainty Estimation

Following prior work (Kuhn et al., 2023a; Band et al., 2022), we evaluate the quality of uncertainty estimates using AUROC and AUARC, which measure the discriminative ability of uncertainty scores and their effectiveness for selective prediction, respectively. We compare our distribution-aware uncertainty with the probabilistic confidence (PC) baseline, defined as one minus the average token probability (Pouget et al., 2016).

Table 2 reports AUROC results across six GUI grounding models. When PC is available, our method consistently achieves higher AUROC. , and on Holo1.5-7B from 0.6983 to 0.7526. For models where PC is not directly applicable (e.g., GUI-Actor variants), our method still attains strong AUROC values (up to 0.8155), demonstrating robust error discrimination under limited model access. Overall, these results suggest that modeling the spatial distribution of grounding predictions yields more informative uncertainty signals than token-level confidence alone.

We further evaluate uncertainty quality using AUARC, which captures accuracy gains as high uncertainty predictions are progressively rejected. As shown in Table 3, our method consistently outperforms baselines across models. For example, on Holo1.5-3B, AUARC improves from 0.6444 to 0.6576 compared to PC. These results indicate that our uncertainty estimates are particularly effective for guiding selective prediction decisions.

4.3. Selective Prediction with FDR Guarantees

While AUROC and AUARC evaluate the quality of uncertainty estimates, reliable deployment further requires translating these scores into principled decision rules with explicit risk guarantees. We therefore study selective prediction under false discovery rate (FDR) control.

Table 1. System-level accuracy (%) of uncertainty-calibrated cascading under different risk levels. "–" indicates infeasible risk levels. Parentheses show Δ over the corresponding model baseline (no cascading). All reported accuracies are computed on the test split, with a test ratio of 0.8.

	Risk Level				
Model	**0.34**	**0.38**	**0.42**	**0.46**	**0.50**
Gemini-only	53.28				
Holo1.5-7B	52.41				
(+SafeGround)	**58.66** *(+ 6.25)*	57.87 *(+ 5.46)*	55.73 *(+ 3.32)*	53.20 *(+ 0.79)*	52.41 *(+ 0.00)*
Holo1.5-3B	45.45				
(+SafeGround)	**53.44** *(+ 7.99)*	52.73 *(+ 7.28)*	52.02 *(+ 6.57)*	49.25 *(+ 3.80)*	47.35 *(+ 1.90)*
UI-TARS-1.5-7B	41.58				
(+SafeGround)	53.68 *(+12.10)*	**54.70** *(+13.12)*	53.04 *(+11.46)*	50.43 *(+ 8.85)*	47.91 *(+ 6.33)*
GUI-Actor-2.5VL-7B	45.69				
(+SafeGround)	**55.18** *(+ 9.49)*	54.86 *(+ 9.17)*	53.60 *(+ 7.91)*	51.38 *(+ 5.69)*	49.17 *(+ 3.48)*
GUI-Actor-2VL-7B	40.79				
(+SafeGround)	**55.18** *(+14.39)*	53.28 *(+12.49)*	53.99 *(+13.20)*	52.96 *(+12.17)*	50.67 *(+9.88)*
GTA1-7B	46.88				
(+SafeGround)	–	–	–	**53.12** *(+ 6.24)*	49.96 *(+ 3.08)*

Table 2. AUROC comparison of uncertainty quantification methods across different models. The best results for each model are highlighted in **bold**. PC is the Probabilistic Confidence baseline.

Model	**Uncertainty Score**	
	PC	U_{COM} **(Ours)**
Holo1.5-3B	0.7576	**0.8056**
Holo1.5-7B	0.6983	**0.7526**
GUI-Actor-2.5VL-7B	-	**0.7793**
UI-TARS-1.5-7B	0.7844	**0.8021**
GUI-Actor-2VL-7B	-	**0.8155**
GTA1-7B	0.6114	**0.6344**

FDR Control Guarantee For each uncertainty method and risk level, we calibrate a decision threshold on the calibration set using the Clopper–Pearson upper confidence bound (Clopper & Pearson, 1934), ensuring that the test-time FDR does not exceed the specified risk level with high probability. Figure 3 illustrates the empirical FDR on the test set across various user-specified risk levels (α). Notably, the evaluated risk levels start from a minimum attainable value. This arises because the intrinsic limitations of the base model and the imperfect discriminative power of uncertainty estimates may cause some incorrect predictions to receive relatively low uncertainty scores, making them inseparable from correct ones by thresholding. As a result, very stringent FDR requirements may be infeasible to satisfy, as no decision threshold can meet the risk constraint under such conditions (Wang et al., 2025a). Importantly, this does not undermine the safety guarantee, as the calibration stage explicitly determines whether a user-specified risk level is achievable prior to deployment, providing a principled fail-safe mechanism for high-stakes interactions.

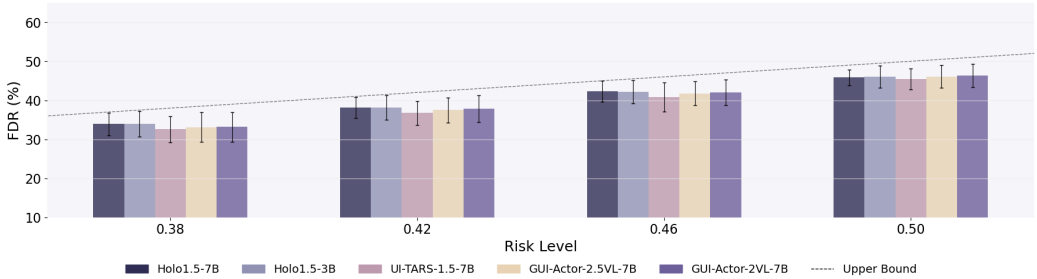


Figure 3. Test-time FDR (mean±std) on the ScreenSpot-Pro dataset under different risk levels.

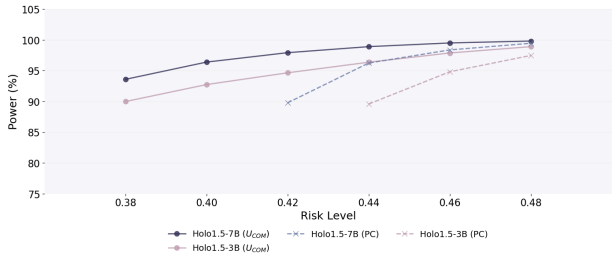


Figure 4. Test-time power (mean) of our U_{COM} and PC baseline on the ScreenSpot-Pro dataset under different risk levels.

Table 3. AUARC comparison of uncertainty quantification methods across different models. The best results for each model are highlighted in **bold**.

Model	Uncertainty Score		
	Random	*PC*	U_{COM} **(Ours)**
Holo1.5-3B	0.4706	0.6444	**0.6576**
Holo1.5-7B	0.5345	0.6686	**0.6705**
GUI-Actor-2.5VL-7B	0.4662	–	**0.7156**
GUI-Actor-2VL-7B	0.4130	–	**0.7166**
UI-TARS-1.5-7B	0.4231	0.6222	**0.6480**
GTA1-7B	0.4769	**0.5521**	0.5511

The results in Figure 3 show that for all tested models (e.g., Holo1.5, UI-TARS), the actual FDR is consistently bounded below the theoretical upper bound. This empirically verifies that SAFEGROUND provides rigorous safety guarantees, ensuring that, with high probability, the error rate among accepted predictions is controlled at the specified level.

Power Comparison In addition to FDR, we report power to further characterize the effectiveness of selective prediction. Higher power indicates that the uncertainty estimates more precisely identify truly risky cases, allowing the system to retain a larger set of reliable predictions without violating the target FDR. Figure 4 compares the power of our method U_{COM} versus the PC baseline under identical risk levels. Across the evaluated models, U_{COM} demonstrates superior robustness, particularly at strict risk levels (e.g., 0.38) where PC often fails to yield valid predictions. Notably, the minimum attainable risk level at which PC can satisfy the FDR constraint is consistently higher than that of U_{COM}, indicating a narrower feasible operating range for PC. U_{COM} consistently outperforms PC, retaining a significantly larger volume of correct responses. These results indicate that U_{COM} is systematically less conservative than PC: it accepts a larger fraction of correct predictions while still satisfying the same FDR constraint.

4.4. Cascading Inference

Finally, we study the system-level benefits of uncertainty-aware decision making in a cascaded inference setting. Given that powerful external models (e.g., Gemini) often incur latency and financial costs, our goal is to improve system accuracy by selectively invoking stronger models when the uncertainty of the base model exceeds a calibrated threshold. Specifically, we fix the calibration split ratio to 0.2 and use the remaining 80% of the data as the test set to evaluate the cascaded system. At test time, predictions with uncertainty scores below or equal to the threshold are handled by the primary local grounding model, while high-uncertainty cases are deferred to the stronger expert model, Gemini-3-pro (Team et al., 2023).

Table 1 reports the accuracy of uncertainty-aware Gemini cascading under different risk levels. Across a wide range of feasible risk levels, the proposed approach consistently improves system accuracy over both Gemini-only inference and the base models, demonstrating the effectiveness of uncertainty-aware cascading. At relatively small risk levels, uncertainty-aware cascading yields substantial accuracy gains. For instance, with Holo1.5-7B at risk level 0.34, the system achieves 58.66% accuracy, outperforming Gemini-only inference by 5.38%. As the risk level increases, the improvement gradually diminishes, since fewer high-uncertainty samples are deferred to Gemini, and the system behavior approaches that of the base model. The effect is more pronounced for models such as Holo1.5-3B and UI-TARS-1.5-7B, where uncertainty-aware cascading improves accuracy by more than 7% to 13% over the base models at relatively small risk levels. We also report the cascading rate in Figure 5, i.e., the fraction of test samples deferred to Gemini. As the risk level increases, the cascading rate con-

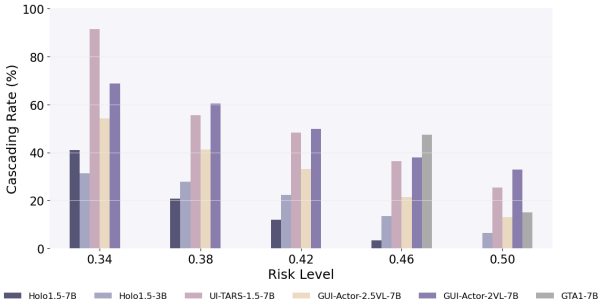


Figure 5. Cascading rate (fraction of test samples deferred to Gemini) across different risk levels.

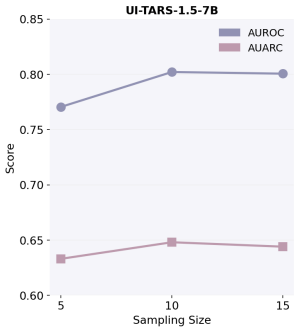


Figure 6. Effect of sampling size K on uncertainty estimation quality for UI-TARS-1.5-7B.

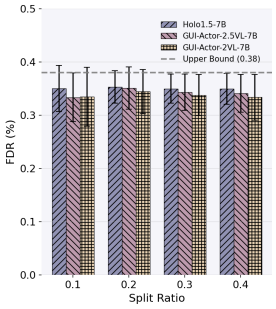


Figure 7. Test-time FDR results of various calibration test split ratios.

sistently decreases across all models, indicating that fewer uncertain cases are escalated to the expert model. This reflects the inherent trade-off between accuracy and expert invocation cost in uncertainty-aware cascading.

4.5. Sensitivity Analyses

Sampling Efficiency We investigate the trade-off between computational cost and estimation quality by varying the sample count K and measuring the resulting AUROC and AUARC. As shown in Figure 6, increasing the sample size from $K = 5$ to $K = 10$ yields a improvement for both metrics, indicating that the proposed uncertainty estimates are already effective with a small number of samples. In contrast, further increasing K from 10 to 15 leads to only marginal changes. Based on this trade-off between performance and computational cost, we set $K = 10$ as the default sampling size in all experiments.

Ablation of Uncertainty Components We analyze the contribution of individual uncertainty components, U_{TA} , U_{IE}, and U_{CD}, across different GUI grounding models. As shown in Table 4, the most informative uncertainty cue is model-dependent. On GTA1, U_{TA} is the strongest single signal, whereas for GUI-Actor-2VL and Holo1.5, U_{CD} is more effective, and U_{TA} alone is insufficient.

Table 4. Ablation study of uncertainty components on GTA1, GUI-Actor-2VL, and Holo1.5 models. Best results within each model block are highlighted in bold.

Model	Uncertainty	AUROC	AUARC
GTA1	U_{TA}	0.6228	0.5481
	U_{IE}	0.5916	0.5390
	U_{CD}	0.5917	0.5389
	U_{COM}	**0.6344**	**0.5511**
	w/o U_{TA}	0.5917	0.5389
GUI-Actor-2VL-7B	U_{TA}	0.4844	0.4335
	U_{IE}	0.7731	0.6435
	U_{CD}	0.7894	0.6505
	U_{COM}	**0.8155**	**0.7166**
	w/o U_{CD}	0.7987	0.6221
Holo1.5-7B	U_{TA}	0.6296	0.6284
	U_{IE}	0.7380	0.6670
	U_{CD}	**0.7529**	**0.6716**
	U_{COM}	0.7526	0.6705
	w/o U_{CD}	0.7303	0.6483

Across all models, no single component consistently dominates. U_{COM} achieves stable performance in all settings, and removing the dominant component for a given model leads to a clear drop in both AUROC and AUARC. This indicates that combining complementary cues yields a more robust, model-agnostic uncertainty estimate for selective prediction. Additional robustness analyses with respect to the uncertainty weighting are provided in the Appendix F.

Sensitivity to Calibration-Test Split Ratio We further study the sensitivity of our method to the calibration–test split ratio when using the combined uncertainty measure U_{COM}. Specifically, we vary the proportion of data allocated to the calibration set while keeping the target risk level fixed, and evaluate the resulting empirical FDR on the test set. As shown in Figure 7, across a wide range of split ratios, the empirical FDR achieved by all three models remains consistently below the target upper bound. These results suggest that our approach does not rely on a carefully tuned split ratio and can be applied robustly in practical settings.

5. Conclusion

We presented SAFEGROUND, an uncertainty-aware framework that enables reliable and risk controlled GUI grounding under limited model access. By modeling spatial uncertainty from stochastic grounding samples, SAFEGROUND captures distributional signals that go beyond point predictions and provide effective discrimination between correct and incorrect predictions. Based on uncertainty estimation, we further calibrate decision thresholds with finite-sample guarantees, supporting deployment-time decision making in high-stakes GUI interactions. Extensive experiments demonstrate that SAFEGROUND achieves accurate uncertainty discrimination, rigorous FDR control, and improved

system-level performance through selective prediction and cascading inference. We hope this work provides a principled way for deploying GUI agents with safety guarantees.

Impact Statement

This paper introduces SAFEGROUND, a framework that significantly enhances the reliability and safety of autonomous GUI agents. By providing the first principled method for uncertainty quantification in GUI grounding with finite-sample statistical guarantees, our work addresses a critical bottleneck in the real-world deployment of visual agents, the risk of high-stakes, irreversible errors (e.g., erroneous financial transactions). Beyond improving individual model reliability, the proposed selective deferral mechanism demonstrates that local models, when combined with uncertainty-aware cascading to powerful external experts, can achieve superior system-level accuracy with substantially reduced computational costs. This research provides a foundational step toward trustworthy human-AI interaction in digital environments, ensuring that automated systems "know when they don't know" and make conservative decisions under ambiguous conditions.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022. URL https://arxiv.org/abs/2110.01052.

Band, N., Rudner, T. G. J., Feng, Q., Filos, A., Nado, Z., Dusenberry, M. W., Jerfel, G., Tran, D., and Gal, Y. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks, 2022. URL https://arxiv.org/abs/2211.12717.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. URL https://arxiv.org/abs/2306.15195.

Cheng, K., Sun, Q., Chu, Y., Xu, F., YanTao, L., Zhang, J., and Wu, Z. 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.

Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934.

Company, H. Holo1.5 - open foundation models for computer use agents, 2025. URL https://

huggingface.co/collections/Hcompany/holo15-68c1a5736e8583a309d23d9b.

Fan, Y., Zhao, H., Zhang, R., Shen, Y., Wang, X. E., and Wu, G. GUI-bee: Align GUI action grounding to novel environments via autonomous exploration. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, November 2025. Association for Computational Linguistics.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. A survey of uncertainty in deep neural networks, 2022. URL https://arxiv.org/abs/2107.03342.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., and Tang, J. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14281–14290, June 2024.

Hou, B., Zhang, Y., Andreas, J., and Chang, S. A probabilistic framework for llm hallucination detection via belief tree propagation, 2025. URL https://arxiv.org/abs/2406.06950.

Hu, M., Zhang, Z., Zhao, S., Huang, M., and Wu, B. Uncertainty in natural language processing: Sources, quantification, and applications, 2023. URL https://arxiv.org/abs/2306.04459.

Jung, J., Brahman, F., and Choi, Y. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=VD-AYtP0dve.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023b. URL https://arxiv.org/abs/2302.09664.

Li, K., Meng, Z., Lin, H., Luo, Z., Tian, Y., Ma, J., Huang, Z., and Chua, T.-S. Screenspot-pro: Gui grounding for professional high-resolution computer use, 2025. URL https://arxiv.org/abs/2504.07981.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. URL https://arxiv.org/abs/2305.19187.

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., and Wei, H. Uncertainty quantification and confidence calibration in large language models: A survey, 2025. URL https://arxiv.org/abs/2503.15850.

Nguyen, D., Chen, J., Wang, Y., Wu, G., Park, N., Hu, Z., Lyu, H., Wu, J., Aponte, R., Xia, Y., Li, X., Shi, J., Chen, H., Lai, V. D., Xie, Z., Kim, S., Zhang, R., Yu, T., Tanjim, M., Ahmed, N. K., Mathur, P., Yoon, S., Yao, L., Kveton, B., Kil, J., Nguyen, T. H., Bui, T., Zhou, T., Rossi, R. A., and Dernoncourt, F. Gui agents: A survey, 2025. URL https://arxiv.org/abs/2412.13501.

Pouget, A., Drugowitsch, J., and Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.

Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., Zhong, W., Li, K., Yang, J., Miao, Y., Lin, W., Liu, L., Jiang, X., Ma, Q.,

Li, J., Xiao, X., Cai, K., Li, C., Zheng, Y., Jin, C., Li, C., Zhou, X., Wang, M., Chen, H., Li, Z., Yang, H., Liu, H., Lin, F., Peng, T., Liu, X., and Shi, G. Ui-tars: Pioneering automated gui interaction with native agents, 2025a. URL https://arxiv.org/abs/2501.12326.

Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025b.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Wang, Q., Fan, Y., and Wang, X. E. Safer: Risk-constrained sample-then-filter in large language models, 2025a. URL https://arxiv.org/abs/2510.10193.

Wang, Q., Geng, T., Wang, Z., Wang, T., Fu, B., and Zheng, F. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=9WYMDgxDac.

Wang, Z., Wang, Q., Zhang, Y., Chen, T., Zhu, X., Shi, X., and Xu, K. SConU: Selective conformal uncertainty in large language models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July .

Wang, Z., Duan, J., Yuan, C., Chen, Q., Chen, T., Zhang, Y., Wang, R., Shi, X., and Xu, K. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond, 2024b. URL https://arxiv.org/abs/2402.14259.

Wang, Z., Duan, J., Wang, Q., Zhu, X., Chen, T., Shi, X., and Xu, K. Coin: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees, 2025c. URL https://arxiv.org/abs/2506.20178.

Wu, Q., Cheng, K., Yang, R., Zhang, C., Yang, J., Jiang, H., Mu, J., Peng, B., Qiao, B., Tan, R., Qin, S., Liden, L., Lin, Q., Zhang, H., Zhang, T., Zhang, J., Zhang, D., and Gao, J. Gui-actor: Coordinate-free visual grounding for gui agents, 2025a. URL https://arxiv.org/abs/2506.03143.

Wu, Q., Cheng, K., Yang, R., Zhang, C., Yang, J., Jiang, H., Mu, J., Peng, B., Qiao, B., Tan, R., et al. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*, 2025b.

Xie, T., Deng, J., Li, X., Yang, J., Wu, H., Chen, J., Hu, W., Wang, X., Xu, Y., Wang, Z., et al. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*, 2025.

Xu, Z., Song, T., and Lee, Y.-C. Confronting verbalized uncertainty: Understanding how llm's verbalized uncertainty influences users in ai-assisted decision-making. *Int. J. Hum.-Comput. Stud.*, 197(C), March 2025. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2025.103455. URL https://doi.org/10.1016/j.ijhcs.2025.103455.

Yang, Y., Li, D., Dai, Y., Yang, Y., Luo, Z., Zhao, Z., Hu, Z., Huang, J., Saha, A., Chen, Z., Xu, R., Pan, L., Savarese, S., Xiong, C., and Li, J. Gta1: Gui test-time scaling agent, 2025. URL https://arxiv.org/abs/2507.05791.

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., Shi, S., and Tu, Z. Benchmarking LLMs via uncertainty quantification. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=L0oSfTroNE.

Zhang, S., Fu, P., Zhang, R., Yang, J., Du, A., Xi, X., Wang, S., Huang, Y., Qin, B., Luo, Z., and Luan, J. Hyperclick: Advancing reliable GUI grounding via uncertainty calibration, 2025. URL https://openreview.net/forum?id=pXYwksqDyE.

Limitation

Our uncertainty estimation relies on the variability in the sampled predictions to characterize spatial ambiguity. For highly deterministic models with limited sampling diversity, the resulting spatial distributions may be less informative. Despite these limitations, SAFEGROUND provides a general and principled foundation for uncertainty-aware GUI grounding.

A. Proofs

In this section, we provide a compelete proof that the upper confidence bound $\hat{\text{FDR}}^{upper}_{1-\delta}(\tau)$ defined in Eq. (9) satisfies the statistical guarantee in Eq. (9). Recall $\hat{\text{FDR}}^{upper}_{1-\delta}(\tau) = \sup\{R : \Pr(\text{Bin}(n, R) \leq X) \geq \delta\}$, where $n = \sum_i^N \mathbf{1}\{u_i \leq \tau\}$ is the number of accepted calibration samples, and $X = \sum_i^N \mathbf{1}\{u_i \leq \tau, A(\hat{y}_i^{(MLG)}, B_i^*) = 0\}$ is the number of accepted incorrect calibration samples. In general, $\text{Bin}(n, R)$ denotes the random variable representing the number of successes in n Bernoulli trials when the system success probability is R. In our setting, it corresponds to the random variable counting the number of errors among n samples when the system FDR is R under a given threshold τ.

We define the cumulative distribution function (CDF) of the random variable $\hat{R}(\tau) = \frac{\text{Bin}(n;R(\tau))}{n}$, corresponding to the error rate over any n accepted samples when the system FDR is $R(\tau)$, as

$$\text{CDF}\big(r \mid R(\tau)\big) = \Pr\big(\hat{R}(\tau) \leq r \mid R(\tau)\big). \tag{13}$$

By the definition of $\hat{\text{FDR}}^{upper}_{1-\delta}(\tau)$, we have

$$\text{CDF}\left(\frac{X}{n} \mid \hat{\text{FDR}}^{upper}_{1-\delta}(\tau)\right) = \delta. \tag{14}$$

If $R(\tau) > \hat{\text{FDR}}^{upper}_{1-\delta}(\tau)$, we have $\text{CDF}\left(\frac{X}{n} \mid R(\tau)\right) \leq \delta$. Then, we have

$$\begin{aligned}\Pr\left(R(\tau) \leq \hat{\text{FDR}}^{upper}_{1-\delta}(\tau)\right) &= 1 - \Pr\left(R(\tau) > \hat{\text{FDR}}^{upper}_{1-\delta}(\tau)\right) \\ &\geq 1 - \Pr\left(\text{CDF}\left(\frac{X}{n} \mid R(\tau)\right) \leq \delta\right).\end{aligned} \tag{15}$$

We further the Inverse Cumulative Distribution Function (ICDF):

$$\text{CDF}^{-1}\big(p \mid R(\tau)\big) = \sup\left\{r : \text{CDF}\left(r \mid R(\tau)\right) \leq p\right\}. \tag{16}$$

If $\text{CDF}\left(\frac{X}{n} \mid R(\tau)\right) \leq \delta$, we have $\frac{X}{n} \leq \text{CDF}^{-1}\big(\delta \mid R(\tau)\big)$. We then obtain

$$\Pr\left(R(\tau) \leq \hat{\text{FDR}}^{upper}_{1-\delta}(\tau)\right) \geq 1 - \Pr\left(\frac{X}{n} \leq \text{CDF}^{-1}\big(\delta \mid R(\tau)\big)\right). \tag{17}$$

Since $\frac{X}{n}$ is exactly the empirical error rate observed over the n accepted samples in the calibration set, the probability that it is less than or equal to $\text{CDF}^{-1}\big(\delta \mid R(\tau)\big)$ does not exceed δ. Finally, we conclude

$$\Pr\left(R(\tau) \leq \hat{\text{FDR}}^{upper}_{1-\delta}(\tau)\right) \geq 1 - \delta. \tag{18}$$

In this way, we obtain an upper bound on the system FDR at threshold τ with at least $1 - \delta$ confidence. At test time, by the exchangeability condition, we provide marginal guarantees of FDR control.

B. Details of Experimental Settings

B.1. Dataset

ScreenSpot-Pro. ScreenSpot-Pro (Li et al., 2025) consists of 1581 UI screenshots paired with natural language instructions that refer to target UI elements on the screen. Each target is annotated as a spatial region rather than a single point. Compared to earlier GUI grounding benchmarks, ScreenSpot-Pro features higher visual complexity, denser UI layouts, and more fine-grained distinctions between neighboring elements, making it particularly suitable for studying uncertainty-aware grounding.

OSWorld-G. OSWorld-G (Xie et al., 2025) contains 564 desktop GUI screenshots sampled from OSWorld agent rollouts, each paired with a natural language instruction and an annotated target bounding box. It covers diverse grounding capabilities, including text matching, element recognition, layout understanding, fine-grained manipulation, and refusal handling. With different UI element types and realistic desktop interfaces, OSWorld-G complements ScreenSpot-Pro by evaluating uncertainty-aware grounding in more practical computer-use scenarios.

B.2. Evaluation Metrics

We evaluate uncertainty estimation quality and selective prediction performance using four complementary metrics: AUROC, AUARC, FDR, and power. All metrics are defined with respect to the admission function $A(\hat{y}, B^*) \in \{0, 1\}$ introduced in Section 3.1, which indicates whether a grounding prediction is admissible.

Area Under Receiver Operating Characteristic (AUROC) Let $U(\hat{y})$ denote an uncertainty score, where larger values indicate higher uncertainty. AUROC measures how well $U(\hat{y})$ separates inadmissible predictions from admissible ones. Formally, AUROC is the area under the receiver operating characteristic curve obtained by thresholding $U(\hat{y})$ to predict whether $A(\hat{y}, B^*) = 0$. A higher AUROC indicates stronger discriminative ability of the uncertainty estimate.

Area Under Accuracy-Rejection Curve (AUARC) AUARC evaluates selective prediction behavior by measuring how accuracy changes as predictions with high uncertainty are rejected. Let $\mathcal{S}_\tau = \{i : U(\hat{y}_i) \leq \tau\}$ denote the set of accepted samples under threshold τ. The accuracy at τ is defined as

$$\mathrm{Acc}(\tau) = \frac{1}{|\mathcal{S}_\tau|} \sum_{i \in \mathcal{S}_\tau} A(\hat{y}_i, B_i^*).$$

In practice, τ is chosen to correspond to a target rejection rate, and AUARC is computed as the area under the curve of $\mathrm{Acc}(\tau)$ as a function of the rejection rate.

False Discovery Rate (FDR) Under a given uncertainty threshold τ, the false discovery rate is defined as

$$\mathrm{FDR}(\tau) = \frac{\sum_i \mathbb{I}\big(U(\hat{y}_i) \leq \tau\big)\, \mathbb{I}\big(A(\hat{y}_i, B_i^*) = 0\big)}{\sum_i \mathbb{I}\big(U(\hat{y}_i) \leq \tau\big)}.$$

FDR quantifies the proportion of inadmissible predictions among all accepted predictions and serves as the primary risk metric controlled by SafeGround.

Power Power measures the proportion of correct predictions retained by selective prediction under a risk constraint and is defined as:

$$\mathrm{Power}(\tau) = \frac{\sum_{i=1}^{N} \mathbb{I}\big(U(\hat{y}_i) \leq \tau\big)\mathbb{I}\big(A(\hat{y}_i, B_i^*) = 1\big)}{\sum_{i=1}^{N} \mathbb{I}\big(A(\hat{y}_i, B_i^*) = 1\big)}.$$

Higher power indicates that more correct predictions are retained while satisfying the specified FDR constraint.

B.3. Spatial Region Construction

Given an input image–instruction pair (x, q), we obtain a set of K sampled grounding predictions $\mathcal{S} = \{\hat{y}^{(i)} = (x^{(i)}, y^{(i)})\}_{i=1}^{K}$ via stochastic decoding. To lift these point-wise samples into a spatial distribution, we discretize the screen into a fixed $H \times W$ grid of patches and map each sampled coordinate to its corresponding patch. Let $C_{u,v}$ denote the number of samples falling into patch (u, v). We then normalize the resulting count map to obtain a spatial probability distribution

$$P_{u,v} = \frac{C_{u,v}}{\sum_{u',v'} C_{u',v'}}, \tag{19}$$

which serves as an empirical estimate of the model's predictive density over the output space.

Gemini

```
"""
You are a GUI agent that locates UI elements in screenshots.

CRITICAL RULES:
1. You MUST output ONLY valid JSON, nothing else
2. Do NOT include any explanation, markdown formatting, or natural language
3. Do NOT wrap the response in code blocks (```json)
4. Coordinates must be in PIXEL values (NOT normalized to 0-1000)
5. If you cannot find the element, output an empty list: []

Your response must be a valid JSON array.
"""
```

Figure 8. A system prompt example for Gemini-3-pro in ScreenSpot-Pro dataset.

Region Extraction To identify object-level grounding hypotheses, we first filter low-density patches using an instance-adaptive threshold. Specifically, let $P_{\max} = \max_{u,v} P_{u,v}$, and retain only patches satisfying $P_{u,v} > \beta P_{\max}$, where β is a fixed ratio (set to 0.3 in our experiments, following (Wu et al., 2025b)). We then group spatially adjacent retained patches (using 4-connected neighborhood) into connected components. This yields a set of disjoint regions $\mathcal{R} = \{R_m\}_{m=1}^{M}$, each corresponding to a plausible grounding target.

Region Scoring For each region R_m, we compute a region-level score

$$S_m = \frac{1}{|R_m|} \sum_{(u,v)\in R_m} P_{u,v}, \tag{20}$$

i.e., the average probability density within the region. This score reflects the relative support assigned to the region by the sampled predictions while remaining invariant to region size. The resulting region scores $\{S_m\}_{m=1}^{M}$ are subsequently normalized and used to compute the uncertainty metrics described in Section 3.3.

C. Threshold Calibration with Finite-Sample Guarantees

This section details the threshold calibration procedure used in SafeGround to obtain finite-sample guarantees on selective prediction risk, based on Clopper–Pearson confidence bounds, as summarized in Algorithm 1.

D. Prompt Template

To ensure a fair and reliable evaluation of large vision–language models on GUI grounding, we adopt a strictly constrained prompt template for Gemini in the ScreenSpot-Pro benchmark, as illustrated in Figure 8, 9.

E. Case Study

We present qualitative examples to illustrate how the proposed uncertainty score reflects the reliability of GUI grounding predictions in practice in Figure 10, 11, 12, 13, 14.

F. Additional Experimental Results

Sensitivity to Sampling Temperature. We further examine the sensitivity of the proposed uncertainty measures to the sampling temperature used during stochastic decoding. Table 5 and Table 6 report AUROC and AUARC results on Holo1.5-3B (Company, 2025) under different temperature settings. As the temperature increases, U_{IE} and U_{CD} become more informative, reflected by consistent gains in AUROC. In contrast, margin-based uncertainty exhibits relatively limited

Gemini

Task: Point to the UI element matching this instruction: {instruction}

Image size: {W} x {H} pixels.

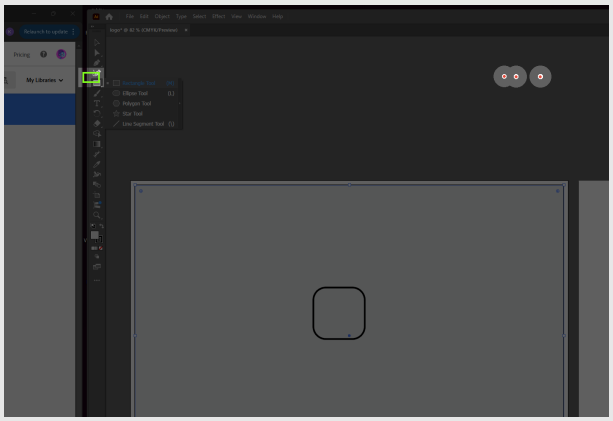
Output format (JSON only, no markdown):
[{"point": [y, x], "label": "description"}]

Where:
- point: [y, x] coordinates in PIXELS (NOT normalized to 0-1000)
- y is vertical (0 to {H})
- x is horizontal (0 to {W})

If no element found, output: []

Example: [{"point": [60, 230], "label": "submit button"}]

Figure 9. A user prompt example for Gemini-3-pro in ScreenSpot-Pro dataset.



Instruction: "Use Rectangle Tool. "

Uncertainty:
" U_{TA} ": 0.65
" U_{IE} ": 0.5
" U_{CD} ": 0.1
" U_{COM} ": 0.29

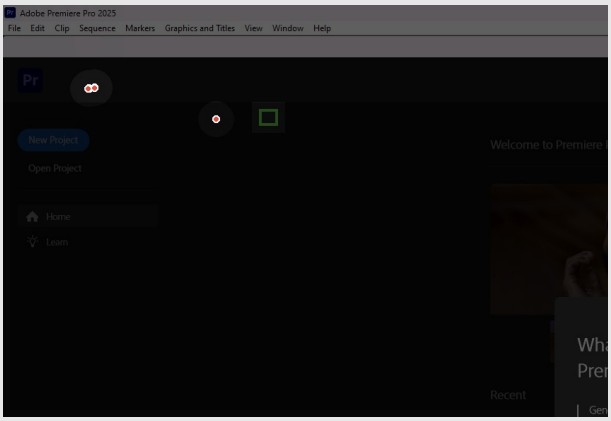
Hit: False

Figure 10. An example of GUI grounding task using our uncertainty score.

sensitivity to temperature changes. U_{COM} shows a dependence on the sampling temperature, reflecting its ability to adapt to changes in the diversity and dispersion of stochastic predictions, while remaining competitive across the evaluated temperature range.

Sensitivity to Uncertainty Weighting. We examine the sensitivity of the proposed framework to the weighting scheme used in the combined uncertainty score U_{COM}. Starting from the default setting $(w_{CD}, w_{IE}, w_{TA}) = (0.6, 0.2, 0.2)$, we evaluate several alternative weighting configurations that moderately vary the relative contributions of the three uncertainty components, while keeping the weights normalized.

Specifically, we consider the following weighting configurations for the combined uncertainty score U_{COM}:

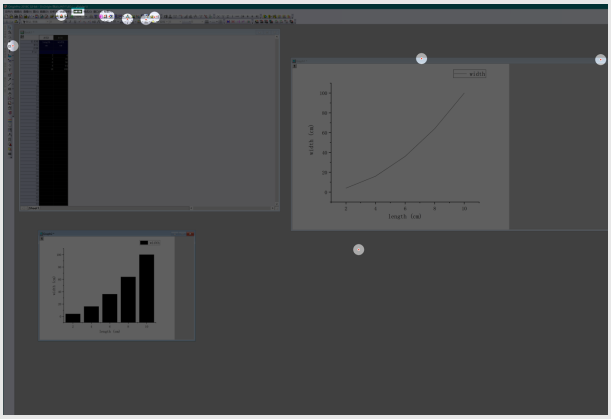


Instruction: "Create captions based on the transcription."

Uncertainty:
"U_{TA}": 0.55
" U_{IE} ": 0.5,
" U_{CD} ": 0.1,
" U_{COM} ": 0.27,

Hit: False

Figure 11. An example of GUI grounding task using our uncertainty score.



Instruction: "Click the tool menu."

Uncertainty:
" U_{TA} ": 1.0
" U_{IE} ": 0.99
" U_{CD} ": 0.89
" U_{COM} ": 0.94

Hit: False

Figure 12. An example of GUI grounding task using our uncertainty score.

- **v1:** $(w_{CD}, w_{IE}, w_{TA}) = (0.34, 0.33, 0.33)$;
- **v2:** $(w_{CD}, w_{IE}, w_{TA}) = (0.2, 0.2, 0.6)$;
- **v3:** $(w_{CD}, w_{IE}, w_{TA}) = (0.2, 0.6, 0.2)$;
- **v4:** $(w_{CD}, w_{IE}, w_{TA}) = (0.5, 0.25, 0.25)$;
- **v5:** $(w_{CD}, w_{IE}, w_{TA}) = (0.25, 0.25, 0.5)$;
- **v6:** $(w_{CD}, w_{IE}, w_{TA}) = (0.25, 0.5, 0.25)$;
- **original:** $(w_{CD}, w_{IE}, w_{TA}) = (0.6, 0.2, 0.2)$.

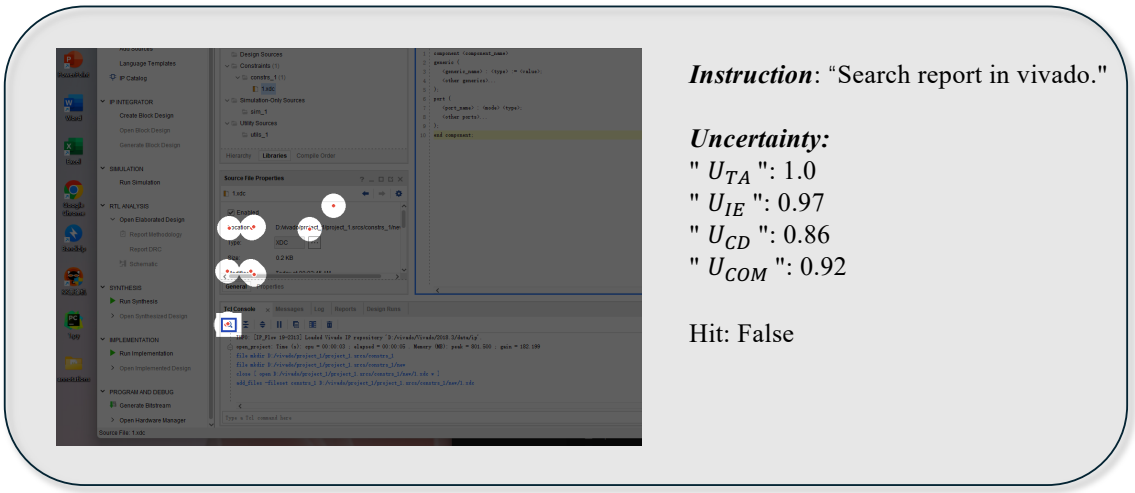


Figure 13. An example of GUI grounding task using our uncertainty score.

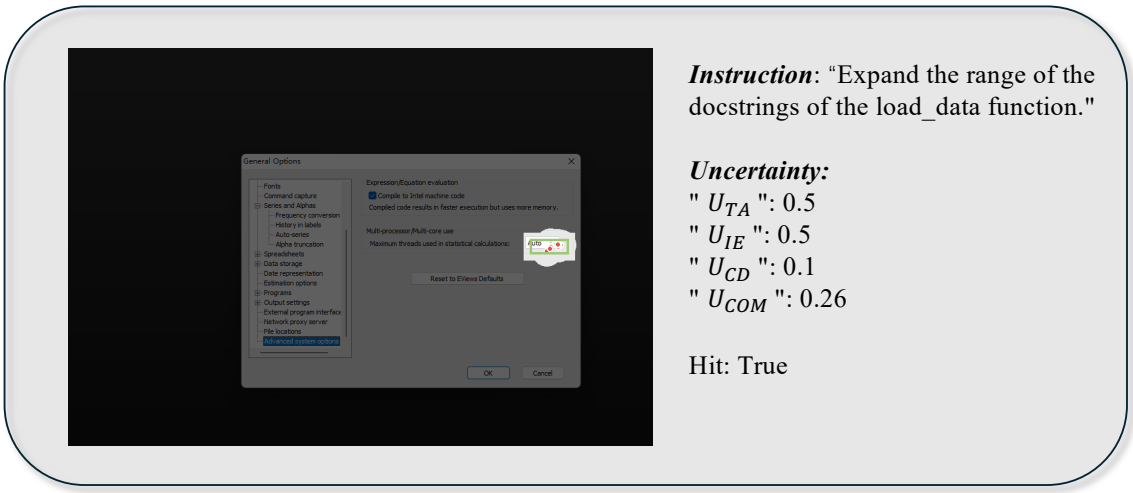


Figure 14. An example of GUI grounding task using our uncertainty score.

As shown in Figure 15, 16, 18, 19, 17, 20, across all evaluated models, both AUROC and AUARC exhibit only minor fluctuations under different weighting schemes. These results indicate that the proposed uncertainty aggregation is robust to moderate changes in the weighting scheme, supporting the use of a fixed, model-agnostic combination in practice.

Experiment results on OSWorld-G. We further evaluate SafeGround on OSWorld-G (Xie et al., 2025) to assess whether the proposed uncertainty scores generalize beyond ScreenSpot-Pro. Overall, the results show that our spatial uncertainty remains effective in this more realistic computer-use setting. For Holo1.5-7B, the combined score achieves the best AUROC of 0.7398 at patch size 21, substantially outperforming the PC baseline (0.6039). For UI-TARS-1.5-7B, the combined score also obtains the best AUROC at patch size 21, reaching 0.8350, compared with 0.7537 for PC. These results indicate that the spatial uncertainty representation provides stronger failure-discrimination ability than scalar confidence baselines.

The AUARC results show a similar trend. For Holo1.5-7B, the best combined AUARC is 0.7932 at patch size 56, while

Table 5. AUROC of different uncertainty measures on Holo1.5-3B under varying sampling temperatures.

Method	Temp=0.3	Temp=0.5	Temp=0.7	Temp=1.0
U_{TA}	0.6258	0.6270	0.6297	0.6404
U_{IE}	0.6621	0.6900	0.7329	0.7753
U_{CD}	0.6689	0.7078	0.7590	0.8060
U_{COM}	0.6819	0.7218	0.7578	0.8056

Table 6. AUARC of different uncertainty measures on Holo1.5-3B under varying sampling temperatures.

Method	Temp=0.3	Temp=0.5	Temp=0.7	Temp=1.0
U_{TA}	0.5373	0.5247	0.5186	0.5709
U_{IE}	0.5182	0.5165	0.5015	0.6534
U_{CD}	0.5219	0.5205	0.4977	0.6578
U_{COM}	0.5250	0.5308	0.4960	0.6576

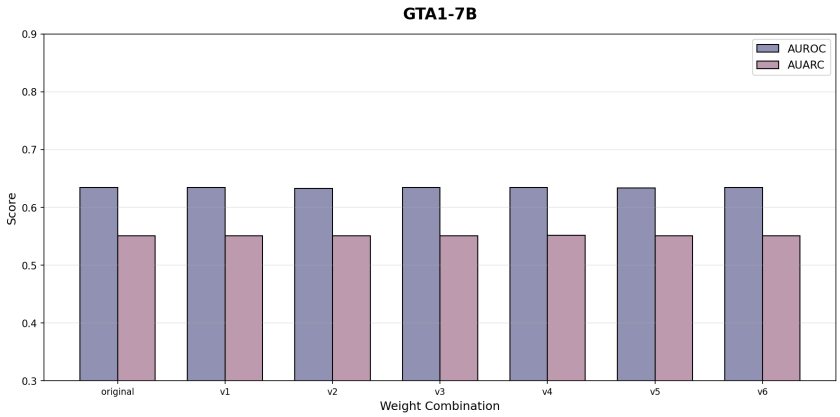


Figure 15. Sensitivity analysis of AUROC and AUARC to uncertainty weighting for GTA1-7B.

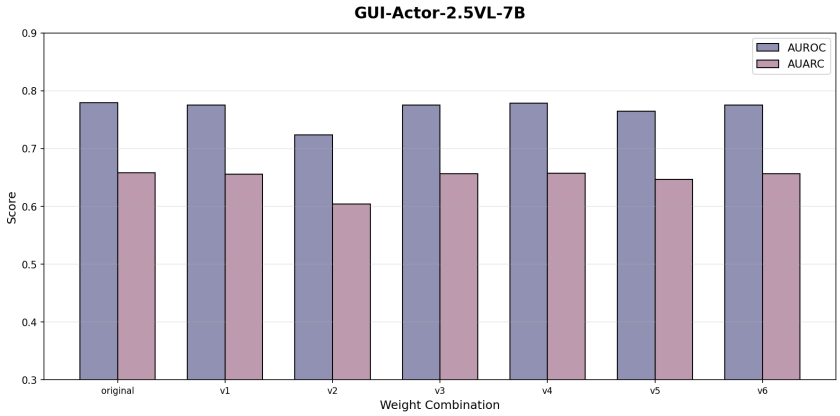


Figure 16. Sensitivity analysis of AUROC and AUARC to uncertainty weighting for GUI-Actor-2.5VL-7B.

for UI-TARS-1.5-7B, the best combined AUARC is 0.7227 at patch size 21. Although the optimal patch size differs across metrics and models, the combined uncertainty score remains consistently competitive across a wide range of patch resolutions, suggesting that SafeGround is not overly sensitive to a single grid configuration. This supports the robustness of the proposed uncertainty formulation under different GUI grounding environments.

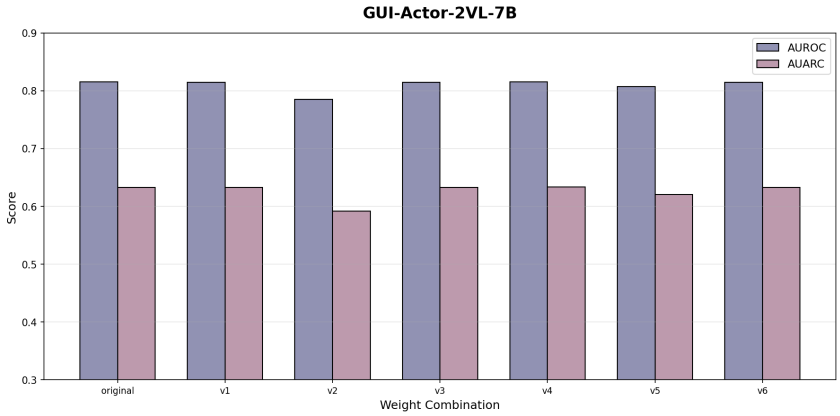


Figure 17. Sensitivity analysis of AUROC and AUARC to uncertainty weighting for GUI-Actor-2VL-7B.

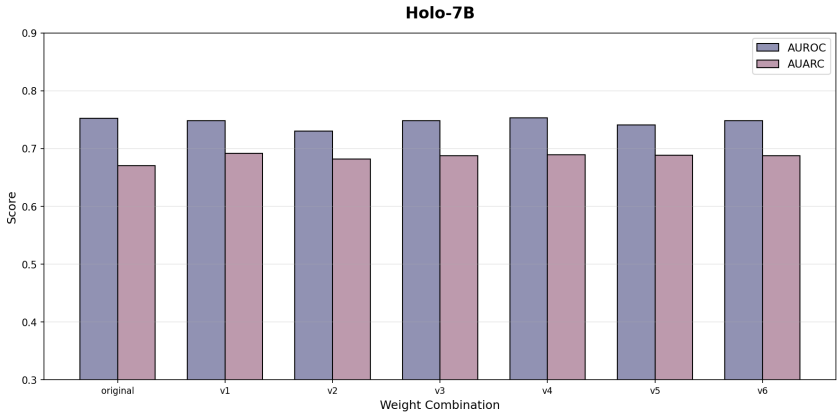


Figure 18. Sensitivity analysis of AUROC and AUARC to uncertainty weighting for Holo1.5-7B.

Figure 19. Sensitivity analysis of AUROC and AUARC to uncertainty weighting for Holo1.5-3B.

Algorithm 1 SafeGround: Clopper–Pearson Threshold Calibration with Sampling-Based Spatial Uncertainty

Input: GUI grounding model f; calibration set $\mathcal{D}_{cal} = \{(x_i, q_i, B_i^*)\}_{i=1}^{N}$; sample count K; patch grid size $H \times W$; region threshold ratio β; admission function $A(\hat{y}, B^*)$; risk level α; significance level δ; weights (w_{CD}, w_{IE}, w_{TA})
Output: calibrated uncertainty threshold $\hat{\tau}$
for $i = 1$ **to** N **do**
  (Primary prediction) Obtain $\hat{y}_i^{(MLG)} \leftarrow f(x_i, q_i)$
  (Sampling) Draw K stochastic predictions $\mathcal{S}_i = \{\hat{y}_i^{(k)}\}_{k=1}^{K}$ via stochastic decoding
  (Discretized density map) Initialize count map $C \in \mathbb{N}^{H \times W} \leftarrow 0$
  for $k = 1$ **to** K **do**
    Map $\hat{y}_i^{(k)}$ to patch index (u, v) and set $C_{u,v} \leftarrow C_{u,v} + 1$
  end for
  Normalize to density $P_{u,v} \leftarrow \frac{C_{u,v}}{\sum_{u',v'} C_{u',v'}}$
  (Region extraction) $P_{\max} \leftarrow \max_{u,v} P_{u,v}$; mask $M_{u,v} \leftarrow \mathbb{I}\{P_{u,v} > \beta P_{\max}\}$
  Group 4-connected active patches in M into M_i connected components via BFS, yielding regions $\mathcal{R}_i = \{R_{i,m}\}_{m=1}^{M_i}$
  (Region scoring) For each region $R_{i,m}$, compute $S_{i,m} \leftarrow \frac{1}{|R_{i,m}|} \sum_{(u,v) \in R_{i,m}} P_{u,v}$
  Sort scores in descending order: $S_{i,(1)} \geq \cdots \geq S_{i,(M_i)}$
  Induce categorical distribution $\hat{p}_{i,j} \leftarrow \frac{S_{i,(j)}}{\sum_{\ell=1}^{M_i} S_{i,(\ell)}}$
  (Uncertainty components)
  $U_{TA,i} \leftarrow \begin{cases} 1 - \frac{S_{i,(1)} - S_{i,(2)}}{S_{i,(1)} + \epsilon}, & M_i \geq 2 \\ \max(0.1,\, 1 - S_{i,(1)}), & M_i = 1 \end{cases}$
  $U_{IE,i} \leftarrow -\frac{1}{\log M_i} \sum_{j=1}^{M_i} \hat{p}_{i,j} \log(\hat{p}_{i,j} + \epsilon)$
  $U_{CD,i} \leftarrow 1 - \sum_{j=1}^{M_i} \hat{p}_{i,j}^2$
  (Combined uncertainty) $u_i \leftarrow w_{CD} U_{CD,i} + w_{IE} U_{IE,i} + w_{TA} U_{TA,i}$
  (Error indicator) $err_i \leftarrow \mathbb{I}\{A(\hat{y}_i^{(MLG)}, B_i^*) = 0\}$
end for
Sort uncertainties ascending: $u_{(1)} \leq \cdots \leq u_{(N)}$ with aligned $err_{(1)}, \ldots, err_{(N)}$
Initialize the selected threshold $\hat{\tau} \leftarrow \texttt{NULL}$
for $t = 1$ **to** N **do**
  Set candidate threshold $\tau \leftarrow u_{(t)}$
  $n \leftarrow \sum_{j=1}^{N} \mathbb{I}\{u_{(j)} \leq \tau\}$ (number of accepted samples)
  $X \leftarrow \sum_{j=1}^{N} \mathbb{I}\{u_{(j)} \leq \tau \wedge err_{(j)} = 1\}$ (number of errors among accepted)
  Compute Clopper–Pearson upper bound: $\mathrm{UCB} \leftarrow \mathrm{BetaInv}(1 - \delta;\, X + 1,\, n - X)$
  if $\mathrm{UCB} \leq \alpha$ **then**
    Update $\hat{\tau} \leftarrow \tau$
  end if
end for
if $\hat{\tau} = \texttt{NULL}$ **then**
  Return "The target risk level α is unattainable under calibration."
else
  Return $\hat{\tau}$
end if

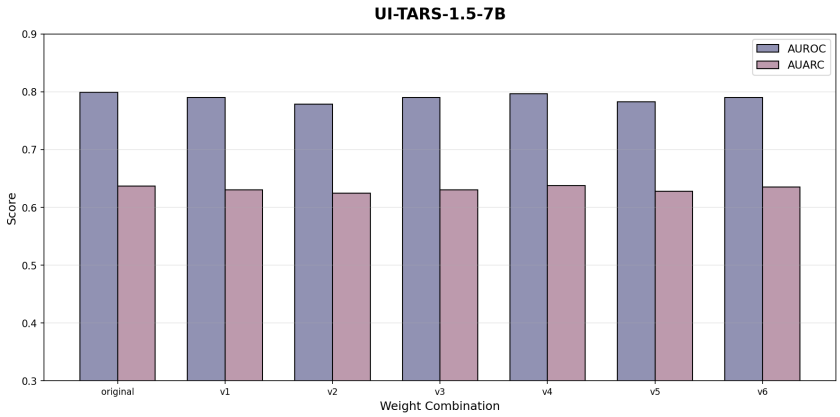


Figure 20. Sensitivity analysis of AUROC and AUARC to uncertainty weighting for UI-TARS-1.5-7B.

Table 7. OSWorld-G results under different patch sizes. We report AUROC and AUARC for the combined uncertainty score U_{COM} and the probabilistic confidence baseline (PC). Best results within each model and metric are highlighted in bold.

Model	**Patch Size**	**AUROC ↑**		**AUARC ↑**	
		U_{COM}	PC	U_{COM}	PC
Holo1.5-7B	7	0.6882	0.6039	0.7550	0.7178
	14	0.7191		0.7766	
	21	**0.7398**		0.7709	
	28	0.6885		0.7377	
	42	0.6918		0.7760	
	56	0.7246		**0.7932**	
UI-TARS-1.5-7B	7	0.8194	0.7537	0.6953	0.6546
	14	0.8283		0.6795	
	21	**0.8350**		**0.7227**	
	28	0.8329		0.6926	
	42	0.8214		0.7000	
	56	0.8138		0.6965	