MPR-GUI: Benchmarking and Enhancing Multilingual Perception and Reasoning in GUI Agents

Anonymous ACL submission

Abstract

With the advancement of computational resources, Large Vision–Language Models (LVLMs) exhibit impressive **P**erception and **R**easoning (**P&R**) performance on Graphical User Interface (GUI) tasks. However, although they demonstrate strong P&R capabilities in English GUI scenarios, their performance in multilingual settings has received little attention, which limits their global applications. Moreover, existing studies on GUI tasks lack fine-grained analyses, including widget functions and elements' spatial relationships, which are fundamental for more targeted improvements. To tackle these issues, we propose **MPR-GUI-Bench**, a **M**ultilingual fine-grained **P**erception and **R**easoning **GUI Bench**mark to evaluate GUI agents' P&R capabilities. Evaluation results demonstrate that LVLMs exhibit significantly worse P&R performance in non-English languages than in English. To address these gaps, we propose **GUI-XLI**, a **GUI C**ross-**L**ingual **I**ntervention method that applies interventions to the hidden states at P&R capability-related layers to mitigate the gaps between English and other languages, building on previous research showing that the hidden states of different language inputs exhibit significant differences in the latent space. Experimental results indicate that our method improves GUI agents' multilingual P&R capability by 6.5% on average.

1 Introduction

With the rapid development of large language-visual models (LVLMs) (Nguyen et al., 2024a), they exibit strong **P**erception and **R**easoning (**P&R**) capabilities across various Graphical User Interface (GUI) benchmarks. As present in Table 1, existing GUI benchmarks remain limited in two aspects: (1) the evaluations are limited in English environments (Rawles et al., 2024; Wang et al., 2024; Chen et al., 2025), which conflict with the global need for multilingual support; (2) current benchmarks lack systematic evaluation of GUI agents' fine-grained P&R capabilities (Cheng et al., 2024; Lu et al., 2024; Chen et al., 2024a) due to overlooking the inherent feature of GUI scenarios (i.e., sparse visual elements and concise layouts). To address these limitations, we propose **M**ultilingual fine-grained **p**erception and **R**easoning **GUI Bench**mark (**MPR-GUI-Bench**), the first benchmark to systematically evaluate the multilingual fine-grained P&R capabilities of GUI agents, featuring identical evaluation settings for each language. As exemplified in Figure 2, to construct **MPR-GUI-Bench**, we propose a semi-automatic pipeline leveraging both human resource and GPT-4o to automatically generate VQA samples and expand them to other languages, which significantly reduces manual effort while ensuring data quality. As presented in Figure 1, the evaluation results of seven baselines on **MPR-GUI-Bench** reveal a significant gap in fine-grained P&R capabilities between English and non-English languages, with an average FPR-ACC accuracy of 75.3% and 67.7%, respectively.

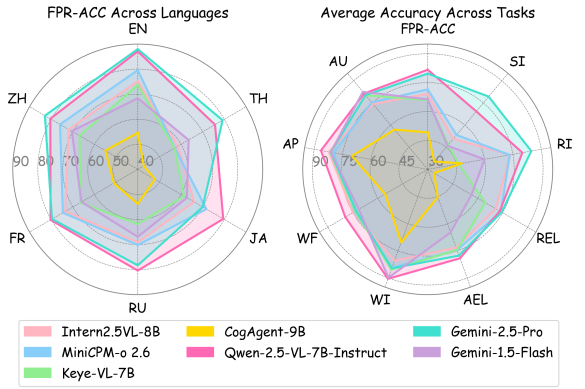


Figure 1: Performance of GUI agents on our **MPR-GUI-Bench** benchmark. The left figure illustrates that all GUI agents exhibit the strongest performance in English, while the right one exhibits their fine-grained P&R capabilities across multiple dimensions.

Dataset	Languages							Capability		Fine-grained	Platform			Size	Type
	EN	ZH	FR	RU	JA	TI	AR	P	R		Web.	Mob.	Desk.		
GUI-WORLD	✔	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✔	✔	12,379	dataset
AndroidWorld	✔	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✗	116	env.
Mobile-Agent-Bench	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	100	env.
ScreenSpot	✔	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✔	✔	1200+	dataset
GUI-Odyssey	✔	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✔	✗	7735	dataset
SPA-Bench	✔	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	340	env.
MacOSWorld	✔	✔	✔	✔	✗	✗	✔	✗	✗	✗	✗	✗	✔	201+29	env.
MPR-GUI-Bench	✔	✔	✔	✔	✔	✔	✗	✔	✔	✔	✔	✔	✗	12,936	dataset

Table 1: Comparison between MPR-GUI-Bench and other GUI benchmarks. In **Capability** category, P means the benchmark involves perception capabilities and R means reasoning capabilities. The **Fine-grained** category means the benchmark involves fine-grained analysis on GUI agents' P&R capabilities. In **Platform** catagory, Web. means website, Mob. means mobile devices and Desk means computer desktop. In the **Type** catagory, env. means interactive environments to evaluate GUI agents' real-time performances, while dataset evaluates task-specific performances through images the same as traditional multimodal benchmarks.

Building on previous works (Nguyen et al., 2024b; Peng et al., 2025; Chang et al., 2022) which have shown that the hidden state distribution of LVLMs' English input differs from other languages, and by aligning the distribution of other languages with English the competence gaps can be migrated, we propose **GUI Cross-Lingual Intervention** (**GUI-XLI**) method. **GUI-XLI** applies intervention to the hidden state of non-English inputs at P&R capability related layers to migrating the cross-lingual gaps. Experimental results demonstrate that our **GUI-XLI** method significantly improves the performance of open-source LVLMs on non-English GUI tasks by 6.5% in average with low inference latency.

Our contributions are summarized as follows:

- we propose a semi-automatic pipeline that leverages compositional prompting with GPT-4o to construct multilingual GUI datasets to reduce manual effort while ensuring data quality.
- We present **MPR-GUI-Bench**, the first multilingual benchmark to evaluate GUI agents' fine-grained P&R capabilities on mobile devices.
- We propose **GUI-XLI**, a training-free representation engineering method that mitigates LVLMs' Scross-lingual P&R capability gaps.

2 Related Work

Multimodal LLM-based Agents The continuous advancement of LVLMs in P&R capabilities reveals their potential as Multimodal LLM-based Agents (MLAs). In GUI scenarios, they can be grouped into three categories: (1) Closed-source LVLM-based GUI agents relying on standardized protocols (Yan et al., 2025) in GUI scenarios; (2) Open-source LVLMs-based GUI agents strengthened by incorporating GUI data into training corpora (Chen et al., 2024b; Yao et al., 2024). These two catagories intend to directly transfer the general competence of foundational LVLMs to real-time GUI scenarios, overlooking the intrinsic properties of GUI tasks. (3) Other GUI agents (Hong et al., 2023; Qin et al., 2025; Yang et al., 2024) trained on GUI datasets with stronger instruction following and GUI-grounding capabilities while reduced generalization and reasoning capabilities.

GUI Agent Benchmarks As presented in Table 1, existing GUI agent benchmarks generally fall into two types: interactive environments and static datasets (Nguyen et al., 2024a). Environment-based benchmarks (Rawles et al., 2024; Wang et al., 2024; Chen et al., 2025) treat each Status-Action-Operation cycle as a whole, providing limited analysis on the agents' P&R capabilities. Dataset-based benchmarks are composed of static screenshots (Cheng et al., 2024; Lu et al., 2024; Chen et al., 2024a), present limited analysis in GUI agents' perception process. Most of recent benchmarks only focus on English. However, with the increasing demand from users in different linguistic environments, GUI agents must possess balanced P&R capabilities across multilingual contexts to achieve broader applications.(Tang et al., 2025; Nguyen et al., 2024a). There have been benchmarks (Yang et al., 2025; Chen et al., 2025) that involve multilingual settings; however,

none of them systematically evaluate fine-grained P&R capabilities, resulting in a lack of targeted explainability. To address this gap, we propose **MPR-GUI-Bench**, the first of its kind to systematically evaluate GUI agents' fine-grained P&R capabilities in multilingual GUI scenarios.

3 MPR-GUI-Bench

Existing studies for GUI applications on LVLMs have mostly neglected fine-grained P&R capabilities, leading to disparities in their development. Moreover, even fewer studies have focused on these capabilities in multi-lingual contexts. As technology advances, users from diverse linguistic backgrounds have an increasingly urgent demand for LVLMs. Therefore, to achieve broader applications in GUI scenarios, it is crucial for LVLMs to eliminate multi-lingual bias. To this end, we introduce the **M**ulti-lingual fine-grained **P**erception and **R**easoning **GUI Bench**mark (**MPR-GUI-Bench**), a benchmark evaluating these capabilities in diverse multilingual GUI tasks.

3.1 Data Source

As shown in Figure 2, parallel screenshots in 6 languages: English, Chinese, French, Russian, Japanese and Thai (EN, ZH, FR, RU, JA, TH); spanning 39 distinct real-world GUI scenarios under two operating systems: iOS and Android on 6 mobile device models are collected by annotators.

3.2 Task Definitions

As shown in Figure 3, we design eight fine-grained dimensions to evaluate LVLMs': (1) **perception capabilities** including perception of the interactive components (widgets) within the screenshots and users' interative actions; (2) **reasoning capabilities** including spacial reasoning capabilities on the location or to clarify spatial relationship between elements and integrated reasoning capabilities on synthesized perception information. The eight fine-grained dimensions are defined as follows:

Perception Capabilities Evaluation Dimensions

- **Widget Function Comprehension (WF)** evaluates LVLMs' perception of the function of GUI elements and the meaning of visual cues.
- **Widget Interaction Comprehension (WI)** evaluates LVLMs' perception of the most suitable way for users to interaction with widgets.
- **Action Understanding (AU)** evaluates LVLMs' perception of the consequences of executed actions, including interface changes, system feedback, and impacts on future interactions.
- **Action Prediction (AP)** evaluates LVLMs' perception of action organization (e.g., types, targets, order, input content) to accomplish goals.

Reasoning Capabilities Evaluation Dimensions

- **Absolute Element Location (AEL)** evaluates LVLMs' reasoning capability to correctly locate UI elements and analyze their global positions.
- **Relative Element Location (REL)** evaluates LVLMs' reasoning capability in relative spatial relationships between GUI elements.
- **Rich information (RI)** evaluates LVLMs' capability to synthesize and reason based on long sequential screenshots with rich perception information, which requires a strong grasp of the fine-grained perception capabilities.
- **Sparse Information (SI)** evaluates LVLMs' capability to synthesize and reason about users' intentions based on shorter screenshot sequences and less information and visual cues compared to RI, leading to higher difficulty.

Notably, the first six dimensions involve samples based on single screenshots, while the last two involve those based on sequential screenshots.

3.3 Benchmark Construction Pipeline

Figure 2 illustrates our automatic dataset construction pipeline leveraging GPT-4o (OpenAI et al., 2024) to reduce human effort. Annotators are required to collect screenshots, design prompts and check GPT-4o's output to ensure data quality.

Step 1: Screenshot Collection Annotators are required to collect parallel screenshots across six languages and distinct GUI scenarios following strict guidelines shown in Appendix B.

Step 2: Candidate VQA Lists Construction Following the definitions in Section 3.2, annotators are required to carefully design prompts for each dimension and leverage GPT-4o to generate English VQAs, forming the candidate VQA lists.

Step 3: Manually Checking Each candidate VQA in the lists is independently reviewed by the annotators according to the following principles:

- Questions should be challenging and answerable based on the reference screenshots, strictly adhering to the definition of corresponding dimension.
- Given the checked questions and answers, the distractors should also be based on the reference

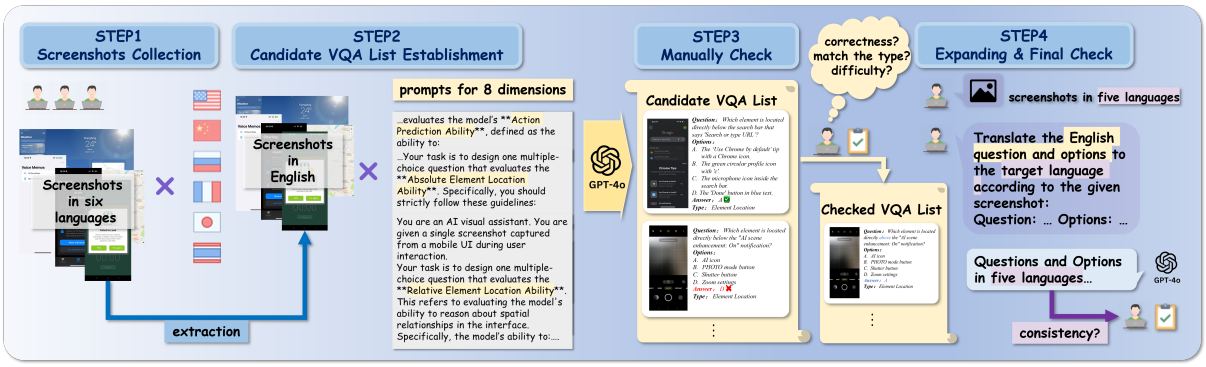


Figure 2: An overview of the MPR-GUI-Bench construction pipeline in §3.3. **Step 1 Screenshot Collection**: Annotators collect parallel screenshots across 6 languages and diverse GUI scenarios. **Step 2 Candidate VQA Lists Construction**: Manually designed prompts and the English screenshots are fed to GPT-4o to construct candidate VQA lists. **Step 3 Manually Checking**: Annotators manually check the candidate VQA lists to ensure quality. **Step 4 Expansion & Final Check**: English QAs and Non-English screenshots are provided to GPT-4o for multi-lingual parallel VQA expansion, followed by annotators' manual check to ensure cross-lingual consistency.

Figure 3: The composition of MPR-GUI-Bench. As shown with gray numbers, We generated 2156 samples for each language, specifically 351 samples for the first 6 dimensions, and 25 for each of the last 2 dimensions.

screenshots while being sufficiently misleading. Fleiss' Kappa (Fleiss, 1971) is computed to measure the inter-rater agreement between annotators with detailed analysis in Appendix A.1.

Step 4: Expansion & Final Check GPT-4o is used to expand the English-version checked VQA list to other five languages. All generated VQAs are checked by annotators for cross-language consistency in visual and textual modalities. We validated the translation quality using back-translation, with results presented in the Appendix A.2.

3.4 Evaluation Metrics

Each sample in our benchmark is a question with four options, including one correct answer and three incorrect distractors. We consider the answer right only when the model chooses exactly the correct option. To represent the overall performance considering dimension difficulty, We define a weighted accuracy metric FPR-ACC:

$$\text{FPR-ACC} = \frac{\sum_{i=1}^{8} w_i \, acc_i}{\sum_{i=1}^{8} w_i}.$$

We categorize the eight dimensions into three difficulty levels and assign different weights to each level. Details are presented in Appendix A.3.

3.5 Experiment Setup

Baseline We select baselines from three model types: (1) **Open-source LVLMs**: Intern2.5VL-8B (Chen et al., 2024b), MiniCPM-o 2.6 (Yao et al., 2024), Qwen-2.5-VL-7B-Instruct (Bai et al., 2025) and Keye-VL-7B (Team, 2025a,b); (2) **Closed-source LVLMs**: Gemini-1.5-flash (Team et al., 2024) and Gemini-2.5-Pro (Comanici et al., 2025) ; (3) **GUI agent**: CogAgent (Hong et al., 2023).

Implementation Details Our evaluation is conducted on 8 × NVIDIA A100 GPUs. For all baseline models, we use default parameter settings.

3.6 Evaluation Result

From the evaluation results illustrated in Table 2, we draw conclusions in two key aspects:

Performance gap across languages & models In terms of languages, GUI agents achieve the highest accuracy in EN followed by other high-resource languages, whereas their performances drop sharply in low-resource languages (e.g., TH

Model	Lang	Perception				Reasoning				FPR-ACC
		AU	AP	WF	WI	AEL	REL	RI	SI	
Open-source LVLMs										
Intern2.5VL-8B	EN	**81.2**	**89.9**	**79.5**	**92.1**	**82.0**	**82.0**	**80.0**	44.0	**75.2**
	ZH	72.4	85.5	75.1	88.0	78.4	67.8	64.0	**60.0**	71.9
	FR	77.1	83.9	75.6	88.5	72.7	76.5	**80.0**	52.0	73.5
	RU	70.2	81.4	70.4	83.3	68.3	66.9	**80.0**	48.0	69.1
	JA	64.2	82.8	72.9	80.6	73.2	69.1	64.0	44.0	66.0
	TH	57.9	67.5	52.6	72.7	42.9	38.3	**80.0**	52.0	58.5
MiniCPM-o 2.6	EN	**83.9**	**83.6**	**81.6**	**91.0**	**77.9**	**84.4**	**84.0**	**64.0**	**79.6**
	ZH	76.5	82.0	75.1	90.2	75.1	73.8	80.0	**64.0**	75.9
	FR	76.5	79.0	77.8	89.9	74.3	76.8	76.0	60.0	74.6
	RU	74.6	79.5	72.3	86.6	72.4	73.8	64.0	56.0	70.2
	JA	72.7	77.6	74.0	85.8	74.0	67.2	68.0	**64.0**	71.7
	TH	66.4	74.0	60.6	78.1	63.1	42.6	52.0	40.0	57.1
Qwen-2.5-VL-7B-Instruct	EN	**86.1**	**89.4**	**86.0**	**93.4**	**86.0**	**81.6**	**96.0**	**72.0**	**87.1**
	ZH	83.6	88.8	77.8	88.8	79.2	74.3	68.0	68.0	80.4
	FR	81.7	83.6	80.0	91.3	76.5	79.0	72.0	**72.0**	80.3
	RU	77.6	86.1	76.7	89.6	77.3	75.1	72.0	**72.0**	80.4
	JA	81.7	87.7	79.2	90.7	77.3	69.1	88.0	68.0	79.5
	TH	76.8	82.5	77.5	88.8	73.5	65.3	76.0	**72.0**	75.7
Keye-VL-7B	EN	**88.3**	**83.9**	**81.6**	**93.7**	**79.0**	**73.8**	**48.0**	**64.0**	**73.7**
	ZH	82.0	81.4	73.4	89.3	77.3	68.0	40.0	52.0	66.9
	FR	84.2	82.5	76.4	91.0	72.4	71.6	44.0	44.0	66.5
	RU	80.1	82.0	72.1	86.6	75.7	68.3	44.0	28.0	61.8
	JA	78.1	82.2	71.2	86.6	72.7	58.2	36.0	40.0	61.4
	TH	75.1	77.3	66.0	84.4	68.3	44.0	36.0	36.0	57.0
GUI Agents										
CogAgent-9B	EN	**63.8**	**78.1**	**63.8**	**81.9**	52.0	**40.8**	44.0	36.0	54.6
	ZH	62.8	74.3	59.3	78.1	**54.0**	31.6	**60.0**	36.0	**55.0**
	FR	56.9	69.7	58.9	68.0	43.2	38.2	52.0	36.0	51.0
	RU	55.2	72.4	52.4	67.2	43.9	31.6	56.0	**52.0**	53.8
	JA	48.3	73.5	54.8	68.0	36.0	26.3	40.0	44.0	47.9
	TH	50.0	69.7	54.8	68.6	32.0	31.1	40.0	20.0	42.8
Close-source LVLMs										
Gemini-1.5-Flash	EN	85.0	**85.8**	**76.2**	**93.4**	**71.6**	61.5	64.0	40.0	68.4
	ZH	**86.2**	81.4	68.5	89.9	64.5	49.2	**68.0**	**64.0**	**70.5**
	FR	84.4	80.6	74.0	90.4	64.8	**65.0**	64.0	36.0	66.0
	RU	80.1	81.2	72.6	89.9	66.1	59.3	60.0	48.0	66.9
	JA	80.3	82.8	71.2	88.8	52.0	44.7	64.0	40.0	62.7
	TH	77.9	79.5	67.7	86.3	59.0	40.4	64.0	48.0	63.5
Gemini-2.5-Pro	EN	85.0	**90.7**	**85.0**	**93.2**	**84.7**	**93.2**	**96.0**	80.0	**88.0**
	ZH	78.4	85.8	82.5	81.2	82.5	71.0	**92.0**	**84.0**	82.9
	FR	**86.9**	64.5	81.6	92.9	81.4	65.3	**88.0**	76.0	79.6
	RU	63.4	90.4	54.0	92.1	75.4	70.8	**92.0**	80.0	78.3
	JA	85.2	84.7	53.4	65.0	60.1	62.3	**68.0**	76.0	70.0
	TH	82.8	71.3	81.6	83.4	81.6	83.7	**72.0**	80.0	79.2

Table 2: Model Performance (%) Across six Languages (EN, ZH, FR, RU, JA, TH). For each dimension, the highest accuracy in each setting is bolded with a **green** background and the lowest accuracy with a yellow background.

and JA), which reveals GUI agents' multi-lingual bias. In terms of model types, open-source LVLMs perform comparably with closed-source ones in high-resource languages. Nevertheless, their performances in low-resource languages is inferior to those of closed-source LVLMs. Furthermore, GUI Agent (e.g., CogAgent) shows a more significant cross-lingual performance gap compared to the foundation models. Additionally, CogAgent's limited instruction-following capability inevitably degrades its performance across all languages.

Performance gap across dimensions GUI agents' varying performance across the 8 dimensions reveals an imbalance in their fine-grained P&R capabilities. While all GUI agents handle basic perception tasks (e.g., WI and WF) well, they exhibit significant weaknesses in dimensions requiring strong reasoning capabilities (e.g.,SI, REL and AEL), These weaknesses reveals their limitations in spatial reasoning capabilities, synthesizing perceived information across sequential GUI scenarios and reason about user intentions on it.

4 GUI-XL-Intervention

As we observed in Table 2, GUI agents demonstrate stronger P&R performance in English. Prior studies have illustrated that cross-lingual disparities can be reduced by aligning hidden state distributions. (Nguyen et al., 2024b; Peng et al., 2025; Chang et al., 2022). Inspired by this, we propose **GUI** **C**ross-**L**ingual **I**ntervention (**GUI-XLI**) method to align the hidden state distribution of non-English with English to migrate there P&R capability gaps. The overview of GUI-XLI method is presented in Figure 12 in Appendix B.

4.1 Preliminary

We consider a LVLM parameterized by θ. It first embeds its visual input $\mathbf{V} = \{v_1, v_2, \ldots, v_m\}$ and input textual input $\mathbf{T} = \{t_1, t_2, \ldots, t_n\}$ into an initial representation $\boldsymbol{H}^1$, where m and n denote the text and visual sequence lengths. This sequence of token representations is then propagated through L Transformer (Vaswani et al., 2017) layers. For each token at a given position j in the sequence, its hidden state vector at layer l, denoted $h_j^{(l)}$, is updated via the residual stream:

$$\boldsymbol{h}_j^{(l)} = \boldsymbol{h}_j^{(l-1)} + \boldsymbol{a}_j^{(l)} + \boldsymbol{m}_j^{(l)}, \tag{1}$$

where $\boldsymbol{a}_j^{(l)}$ and $\boldsymbol{m}_j^{(l)}$ denote the outputs of the multi-head attention and the MLP block, respectively. The final hidden representation $\boldsymbol{H}^{(L)}$ is then fed into the autoregressive language head.

4.2 GUI-XL-Memory

Due to LLMs' autoregressive nature, the hidden state of the last token is particularly informative. Thus, for an input query Q and image I, we denote the hidden state of the last token at layer l as: $h^{(l)}(Q, I) \in \mathbb{R}^d$. To characterize cross-lingual discrepancies, we define a difference vector between two semantically equivalent inputs (Q_a, I_a) and (Q_b, I_b), expressed in languages a and b. This difference captures how the model encodes language-specific semantics in its latent space:

$$\Delta_{a\to b}^{(l)} = h^{(l)}(Q_a, I_a) - h^{(l)}(Q_b, I_b). \tag{2}$$

Intuitively, $\Delta_{a\to b}^{(l)}$ encapsulates the representational gap between two languages at the hidden-state level. To effectively utilize it, we construct a **GUI** **C**ross-**L**ingual **Memory** (**GUI-XL-Memory**), which stores knowledge pairs of task-related representations and corresponding difference vectors.

Specifically, we collect semantically equivalent Visual Question Answering (VQA) pairs $(X_{\text{en}}, I_{\text{e}})$ and $(X_{\text{tgt}}, I_{\text{tgt}})$, where $X = (Q, A)$ denotes a QA input with reasoning chain. For each VQA, we extract the target language representation:

$$r_{\text{tgt}}^{(l)} = h^{(l)}(X_{\text{tgt}}, I_{\text{tgt}}), \tag{3}$$

and the cross-lingual difference vector:

$$u_{\text{en-tgt}}^{(l)} = h^{(l)}(X_{\text{en}}, I_{\text{en}}) - h^{(l)}(X_{\text{tgt}}, I_{\text{tgt}}). \tag{4}$$

We store $(r_{\text{tgt}}^{(l)}, u_{\text{en-tgt}}^{(l)})$ as one entry in the **GUI-XL-Memory**. Here, $r_{\text{tgt}}^{(l)}$ serves as the retrieval key, while $u_{\text{en-tgt}}^{(l)}$ provides a transferable direction that bridges the gap between target-language and English P&R capability in the latent space.

4.3 Cross-lingual Representation Intervention

We now introduce our intervention mechanism to enhance LVLMs' P&R capabilities in non-English languages. Given an input VQA pair $(Q_{\text{tgt}}, I_{\text{tgt}})$, we first extract its hidden state $h_{\text{tgt}}^{(l)}$ at layer l, then retrieve the top-k semantically nearest entries from **GUI-XL-Memory** by selecting the set of indices:

$$\mathcal{I} = \underset{\mathcal{J}\subseteq\{1,\ldots,N\},\,|\mathcal{J}|=k}{\arg\max} \sum_{i\in\mathcal{J}} \frac{(h_{\text{tgt}}^{(l)})^\top r_i^{(l)}}{\|h_{\text{tgt}}^{(l)}\|_2\,\|r_i^{(l)}\|_2}. \tag{5}$$

Here, N is the total number of stored entries. The corresponding difference vectors are retrieved and averaged to form the intervention signal:

$$\bar{u}_{\text{en-tgt}}^{(l)} = \frac{1}{k}\sum_{j\in\mathcal{I}} u_j^{(l)}. \tag{6}$$

During decoding, we intervene on the hidden state of the first generated token by injecting the averaged difference vector and normalizing the result to preserve the scale of the original representation:

$$\tilde{h}_{\text{tgt}}^{(l)} = \|h_{\text{tgt}}^{(l)}\|_2 \cdot \frac{h_{\text{tgt}}^{(l)} + \alpha\,\bar{u}_{\text{en-tgt}}^{(l)}}{\|h_{\text{tgt}}^{(l)} + \alpha\,\bar{u}_{\text{en-tgt}}^{(l)}\|_2}. \tag{7}$$

Here, α is a tunable coefficient that controls the strength of intervention. By injecting $\bar{u}_{\text{en-tgt}}^{(l)}$ only at inference time, our method enhances non-English representations toward English-like reasoning patterns without modifying the model parameters. This makes GUI-XLI training-free, and transferable across different LVLM architectures.

Model	Lang	GXI	Perception				Reasoning				FPR-ACC
			AU	AP	WF	WI	AEL	REL	RI	SI	
Intern2.5VL-7B	ZH	×	72.4	85.5	75.1	88.0	78.4	67.8	64.0	60.0	71.9
		✓	81.9 ↑9.5	90.2 ↑4.7	80.3 ↑5.2	90.5 ↑2.5	81.9 ↑3.5	70.2 ↑2.4	80.0 ↑16.0	72.0 ↑12.0	82.8 ↑10.9
	TH	×	57.9	67.5	52.6	72.7	42.9	38.3	80.0	52.0	58.5
		✓	58.3 ↑0.4	69.4 ↑1.9	55.8 ↑3.2	71.7 ↓1.0	44.1 ↑1.2	39.6 ↑1.3	80.0 ↑0.0	40.0 ↓12.0	62.2 ↑3.7
	JA	×	64.2	82.8	72.9	80.6	73.2	69.1	64.0	44.0	69.1
		✓	67.5 ↑3.3	85.5 ↑2.7	75.1 ↑2.2	81.5 ↑0.9	72.4 ↓0.8	68.7 ↓0.4	72.0 ↑8.0	56.0 ↑12.0	77.2 ↑8.1
Qwen-2.5-VL-Instruct	ZH	×	83.6	88.8	77.8	88.8	79.2	74.3	68.0	68.0	77.1
		✓	86.2 ↑2.6	89.2 ↑0.4	78.1 ↑0.3	91.5 ↑2.7	79.6 ↑0.4	74.6 ↑0.3	84.0 ↑16.0	76.0 ↑8.0	83.1 ↑6.0
	RU	×	77.6	86.1	76.7	89.6	77.3	75.1	72.0	72.0	78.1
		✓	84.6 ↑7.0	87.9 ↑1.8	77.4 ↑0.7	88.6 ↓1.0	77.5 ↑0.2	74.5 ↓0.6	84.0 ↑12.0	84.0 ↑12.0	83.6 ↑5.5
	JA	×	81.7	87.7	79.2	90.2	77.3	69.1	88.0	68.0	79.9
		✓	83.6 ↑1.9	88.3 ↑0.6	81.7 ↑2.5	93.5 ↑3.3	77.1 ↓0.2	69.3 ↑0.2	92.0 ↑4.0	80.0 ↑12.0	84.4 ↑4.5

Table 3: Model Performance (%) Before and After Using the GUI-XLI (GXI) Method. The baseline (×) performance is shown first, followed by the performance with GUI-XLI enabled (✓). The gain or loss is indicated next to the score, with green background indicating performance improvement and red indicating decline.

5 Experiment

5.1 Setup

Baseline Models We evaluate the effectiveness of GUI-XLI on two advanced LVLMs: InternVL-8B (Chen et al., 2024b) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) on our benchmark.

Memory for RI&SI We combine the memories of the six single screenshot dimensions (i.e., AU, AP, AEL, REL, WF and WI) to form the memory for RI and SI, because these two dimensions assess the model's capabilities to integrate and utilize these capabilities from the first six dimensions.

5.2 Main Results

Table 3 presents the models' evaluation results on our **MPR-GUI-Bench** before and after applying **GUI-XLI**, leading to three key conclusions:

(1) Effective Multilingual P&R Capability Enhancement Our GUI-XLI method significantly improves the fine-grained P&R capabilities across multiple languages. For high-resource languages (e.g., ZH), we observe a notable improvement, with Intern2.5VL-8B achieving a 10.9% increase and Qwen-2.5-VL-7B-Instruct achieving a 6.0% increase in FPR-ACC. Additionally, in low-resource languages (e.g., TH, JA), the improvement is also substantial, further reducing the performance gaps between English and non-English languages.

(2) Data and Model Generalizability Our method proves to be data- and model-agnostic without depending on the specifics of the dataset or the model used. When constructing **GUI-XL-Memory**, we ensure that it does not overlap with our **MPR-GUI-Bench**. The improvements seen across non-English languages are due to the P&R capability they obtain from the difference vectors, allowing the approach to generalize effectively. The data-independent nature of the method is a key strength, making it highly adaptable to a wide range of tasks, models, and datasets.

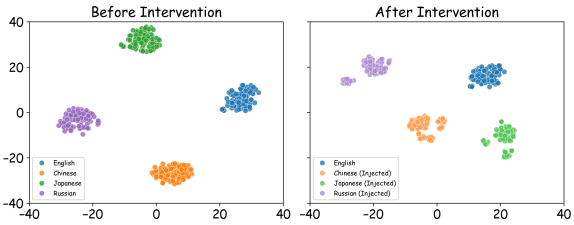


Figure 4: t-SNE Visualization of Multilingual Hidden State before and after applied GUI-XLI.

(3) Significant Improvements across Tasks We observe that the performance improvements are more pronounced in tasks involving complex action understanding and prediction (i.e.,AU and AP), with Intern2.5VL-8B achieving a 3.7% increase and Qwen-2.5-VL-7B-Instruct achieving a 2.4% increase averagely in accuracy on these two dimensions. In contrast, for simpler dimensions (e.g., WI) that already have high baseline performance, the improvements are relatively smaller. For spatial reasoning and integrated reasoning tasks (e.g., REL and SI), the improvements are also more modest.

6 Analysis

6.1 Visualization of Multilingual Hidden State

To better understand how **GUI-XLI** improves multilingual performance, we use t-SNE to visualize

Model	Lang	GXI	Perception				Reasoning				FPR-ACC
			AU	AP	WF	WI	AEL	REL	RI	SI	
Qwen-2.5-VL-Instruct	ZH	×	77.8	72.9	69.8	67.8	73.7	80.6	84.0	68.0	74.2
		✓	82.1 ↑4.3	83.5 ↑10.5	74.4 ↑4.6	67.7 ↓0.1	78.3 ↑4.6	85.5 ↑4.8	80.0 ↑0.3	72.0 ↑4.0	77.4 ↑3.2
	RU	×	83.5	87.9	76.9	68.4	77.4	87.7	84.0	80.0	80.8
		✓	88.9 ↑5.4	90.7 ↑2.8	77.5 ↑0.6	68.4 ↑0.0	76.4 ↓1.0	87.5 ↓0.2	88.0 ↑4.0	80.0 ↑0.0	82.3 ↑1.5
	JA	×	82.4	86.3	69.5	65.2	72.6	84.9	88.0	72.0	77.6
		✓	83.6 ↑1.2	88.3 ↑2.0	70.6 ↑1.1	66.1 ↑0.9	71.5 ↓1.1	85.2 ↑0.3	92.0 ↑4.0	76.0 ↑4.0	79.5 ↑1.9

Table 4: Reasoning Effect (%) Before and After Using the GXI Method. The **Reasoning Answering** (×) performance is shown first, followed by the performance of **Reasoning + GUI-XLI** (✓). The gain or loss is indicated next to the score, with green background meaning performance improvement and red meaning decline.

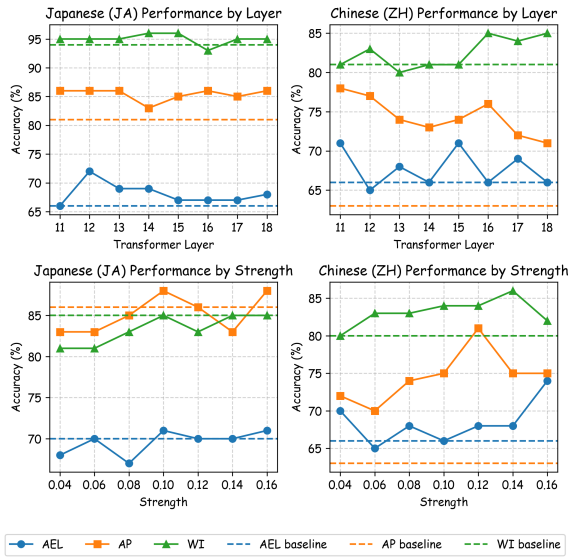


Figure 5: Line chart of grid search on **MPR-GUI-Bench** for intervention strength α and layer l on Zh and JA. The upper two figures present the grid search results for l, and the lower two present those for α.

the last token' hidden states of English inputs and their semantically parallel non-English counterparts (i.e., ZH, RU, JA and TH) in Figure 4, before and after applying GUI-XLI. Without GUI-XLI, the hidden states of semantically parallel questions in different languages are scattered, forming distinct clusters by language. After applying GUI-XLI, these hidden state distributions are pulled closer together, forming clusters centered around their English counterparts. This phenomenon indicates that our method effectively reduces the cross-lingual gaps , making GUI agents to perform more consistently in multilingual GUI tasks.

6.2 Ablation Studies

Figure 5 illustrates our systematic grid search to select the optimal layer l and intervention strength α. Based on prior finding (Zhao et al., 2024) that non-English queries initially produce non-English embeddings but shift towards English-like representations in the middle layers, injecting the P&R capability difference vectors into the middle layers best facilitates alignment of non-English hidden state with English . We first fix α at 0.1 and conduct a grid search across layers 11 to 18. Based on the best layer configuration, we conduct a grid search on α. The optimal configurations vary across language settings, and we select the most suitable ones for each model based on weighted accuracy. Additionally, although experiments are conducted on all settings, we only present three representative dimensions (i.e., AP, AEL and WI) for Qwen2.5VL-8B-Instruct due to space limitations,.

6.3 Reasoning Enhancement of GUI-XLI

As present in Table 4, beyond its impact on non-English P&R capability enhancement, experimental results indicates that GUI-XLI also benefits tasks in our benchmark when required to output reasoning chain before answering their options. Specifically, we compare two settings: (i) **Reasoning Answering**, where the LVLM-based GUI agents is prompted to generate intermediate reasoning steps before selecting an answer; and (ii) **Reasoning + GUI-XLI**, where reasoning chain is combined with our GUI-XLI method.

7 Conclusion

In this paper, we introduce **MPR-GUI-Bench**, the first multilingual benchmark for evaluating GUI agents' fine-grained perception and reasoning capabilities. The evaluation results reveal that their capabilities in English is stronger than in other languages. To address this gap, we proposed **GUI-XLI**, a training-free representation engineering method which significantly reduces cross-lingual perception and reasoning capability gaps.

Limitations

Due to limited resources, our **MPR-GUI- Bench** only include mobile device models. We will work on expanding it to more platforms including website and desktop. Additionally, we are unable to extend our **GUI-XLI** to closed-source LVLMs.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, and 1 others. 2024a. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*.

Jingxuan Chen, Derek Yuen, Bin Xie, Yuhao Yang, Gongwei Chen, Zhihao Wu, Li Yixing, Xurui Zhou, Weiwen Liu, Shuai Wang, Kaiwen Zhou, Rui Shao, Liqiang Nie, Yasheng Wang, Jianye HAO, Jun Wang, and Kun Shao. 2025. Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In *The Thirteenth International Conference on Learning Representations*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for gui agents. *Preprint*, arXiv:2312.08914.

Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*.

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, and 10 others. 2024a. Gui agents: A survey. *Preprint*, arXiv:2412.13501.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, and Ranjay Krishna. 2024b. Multilingual diversity improves vision-language representations. In *Advances in Neural Information Processing Systems*, volume 37, pages 91430–91459. Curran Associates, Inc.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qiwei Peng, Guimin Hu, Yekun Chai, and Anders Søgaard. 2025. Debiasing multilingual llms in cross-lingual latent space. *Preprint*, arXiv:2508.17948.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

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. 2024. Androidworld: A dynamic benchmarking environment for autonomous agents. *Preprint*, arXiv:2405.14573.

Fei Tang, Haolei Xu, Hang Zhang, Siqi Chen, Xingyu Wu, Yongliang Shen, Wenqi Zhang, Guiyang Hou,

Zeqi Tan, Yuchen Yan, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueting Zhuang. 2025. A survey on (m)llm-based gui agents. *Preprint*, arXiv:2504.13865.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Kwai Keye Team. 2025a. Kwai keye-vl 1.5 technical report. *Preprint*, arXiv:2509.01563.

Kwai Keye Team. 2025b. Kwai keye-vl technical report. *Preprint*, arXiv:2507.01949.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Luyuan Wang, Yongyu Deng, Yiwei Zha, Guodong Mao, Qinmin Wang, Tianchen Min, Wei Chen, and Shoufa Chen. 2024. Mobileagentbench: An efficient and user-friendly benchmark for mobile llm agents. *Preprint*, arXiv:2406.08184.

Yunhe Yan, Shihe Wang, Jiajun Du, Yexuan Yang, Yuxuan Shan, Qichen Qiu, Xianqing Jia, Xinge Wang, Xin Yuan, Xu Han, Mao Qin, Yinxiao Chen, Chen Peng, Shangguang Wang, and Mengwei Xu. 2025. Mcpworld: A unified benchmarking testbed for api, gui, and hybrid computer use agents. *Preprint*, arXiv:2506.07672.

Pei Yang, Hai Ci, and Mike Zheng Shou. 2025. macosworld: A multilingual interactive benchmark for gui agents.

Yuhao Yang, Yue Wang, Dongxu Li, Ziyang Luo, Bei Chen, Chao Huang, and Junnan Li. 2024. Aria-ui: Visual grounding for gui instructions. *arXiv preprint arXiv:2412.16256*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Additional Details of MPR-GUI-Bench

A.1 Inter-Rater Reliability Analysis using Fleiss' Kappa

1. Data Summary

An inter-rater reliability analysis is conducted to determine the consistency of agreement among 6 annotators for 2,156 samples in each languages. For all six languages we conducted certain analysis, here we take English as an example. Each VQA sample is classified into one of two nominal categories: "Compliant" or "Non-compliant". The distribution of ratings is summarized in Table 5.

Rater Agreement Distribution (Comp. vs. Non-Comp.)	Frequency (Items Num.)
6 vs. 0	1693
5 vs. 1	291
4 vs. 2	110
3 vs. 3	42
2 vs. 4	16
1 vs. 5	1
0 vs. 6	3
Total	**2156**

Table 5: Summary of Rater Agreement Distribution for English Candidate VQA Lists Checking. Comp. means Complaint and Non-Comp. means Non-complaint.

2. Calculation of Fleiss' Kappa

Fleiss' Kappa (κ) is calculated to assess the degree of agreement beyond what would be expected by chance.[1, 2] The calculation followed three steps.

Step 1: Overall Observed Agreement ($\bar{P}$)

The proportion of observed agreement for each item (P_i) is calculated using the formula:

$$P_i = \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right)$$

where $n = 6$ is the number of raters and $k = 2$ is the number of categories. The average of these proportions across all $N = 2156$ items yielded the overall observed agreement:

$$\bar{P} = 0.9120$$

Step 2: Agreement Expected by Chance ($\bar{P}_e$)

The proportion of all ratings assigned to the "Compliant" category (p_1) and the "Non-compliant" category (p_2) is calculated from the total of 12,936 ratings (2156 items $\times$ 6 raters).

- $p_1 = \frac{(1693\times6)+(291\times5)+\cdots+(1\times1)}{2156\times6} = 0.9440$
- $p_2 = \frac{(3\times6)+(1\times5)+\cdots+(291\times1)}{2156\times6} = 0.0560$

The probability of agreement by chance is the sum of the squared proportions:

$$\bar{P}_e = p_1^2 + p_2^2 = (0.9440)^2 + (0.0560)^2 = 0.8942$$

Step 3: Fleiss' Kappa (κ)

The final Kappa value is calculated using the standard formula:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{0.9120 - 0.8942}{1 - 0.8942} = 0.1682$$

The Fleiss' Kappa coefficient for this analysis is $\kappa = 0.17$. According to the benchmarks proposed by Landis and Koch (1977), a Kappa value in the range of 0.00–0.20 indicates slight agreement.

A.2 Validation on GPT-4o Translation

To validate the translation quality of GPT-4o, we adopt the back translation method. First, we randomly sample 500 English VQAs from our MPR-GUI-Bench. Then we leverage GPT-4o to translate these questions to other 5 languages according to řefstep 4, followed by translating them back to English. Finally, we evaluate the accuracy of Qwen 2.5VL-7B-Instruct on these samples and the evaluation result is present in Table 6.

Translation Path	Accuracy (%)
Original (EN)	87.2
ZH $\rightarrow$ EN	87.2
JA $\rightarrow$ EN	86.6
RU $\rightarrow$ EN	86.0
FR $\rightarrow$ EN	87.0
TH $\rightarrow$ EN	86.2

Table 6: Back-translation Accuracy (%) of Qwen 2.5VL-7B-Instruct on 500 VQA Samples. The first column shows accuracy on original English questions, while subsequent columns show accuracy on questions back-translated from the target language to English.

A.3 Details about MPR-ACC

We use the MPR-ACC parameter as the comprehensive score for the fine-grained P&R capabilities

of the model on our MPR-GUI-Bench. Specifically, we categorize the eight task dimensions into three difficulty levels. The six static dimensions (Table 2, d1–d6) involve only single-image perception and are assigned a base weight of $w_i = 1$. The RI dimension (d7), which benefits from temporal context across multiple screenshots, is assigned a medium weight of $w_7 = 1.5$. The SI dimension (d8), which requires inferring user intentions from minimal visual evidence and sparse information and represents the highest reasoning challenge, is assigned the largest weight of $w_8 = 2$.

A.4 Prompts

In this section, we list all prompts used during the process of constructing **MPR-GUI-Bench**, which include VQA generation for eight dimensions (Table 7 - Table 13). Note that for RI and SI demensions, we ask annotators to provide the goal for the screenshot sequences, so the corresponding prompt requires GPT-4o to only generate distractors.

A.5 Case Study

In this section, we present examples of incorrect responses generated by the baseline models for each dimension, along with the corresponding question categories included in our **MPR-GUI-Bench** benchmark. (Figure 8 - Figure 11)

A.6 Data Collection Guidelines

In this section, we provide guidelines for annotators on data collection and verification to ensure data quality and consistency across annotators.

B Overview of GUI-XLI

In this section, we present the overview illustration figure of our GUI-XLI method in Figure 12.

Prompt for AU Dimension
You are an AI visual assistant. You are given a single screenshot captured from a mobile UI during user interaction. Your task is to design ONE multiple-choice question that evaluates the model's **Action Understanding Ability**, defined as the ability to: Predict the immediate outcome and effects of performing a specific action given the current interface state. Focus on: 1. **Interface state changes** (e.g., navigating to a new page, opening or closing a popup, expanding or collapsing content areas, toggling an icon's opacity or color) 2. **Data state changes** (e.g., saving data, deleting an item) 3. **System feedback** (e.g., displaying a success message, an error warning, or a loading indicator) 4. **Impact on subsequent flow** (e.g., unlocking the next step, resetting to the initial state, reaching a terminal page, expiring a critical condition) **Notice**: Ensure that the questions you design for these tasks are answerable and the answers can be deduced from the GUI content. You must make each question as **difficult and nuanced** as possible, requiring careful visual perception and contextual reasoning. Avoid obvious or overly simple options. Include plausible distractors for each question to increase the difficulty. For each given screenshot, create one multiple-choice question that tests one of the abilities mentioned above. Each question should have four answer options: one correct answer and three that are incorrect but closely relevant. Distractors should be designed to be tempting yet contain subtle mistakes drawn from the interface that are difficult to detect. Your reply must be structured like this, with **no extra explanation**: question: {your question} options: A. {Option A} B. {Option B} C. {Option C} D. {Option D} answer: {Correct option letter} type: Action Understanding Please keep each answer as **concise and difficult** as possible, and only structured in this exact format. Only include questions that you can answer confidently based on the image content.

Table 7: Prompt for AU Dimension

Prompt for AP Dimension
You are an AI visual assistant. You are given a single screenshot captured from a mobile UI during user interaction. Below, you will be provided with a hypothetical user task goal. Your task is to design ONE multiple-choice question that evaluates the model's **Action Prediction Ability**, defined as the ability to: 1. **Action Type**: Select the correct interaction type from the set {tap, long press, swipe, type text, press home/back/recent}. 2. **Action Target**: Identify the precise UI element to interact with. 3. **Input Content**: If text input is required, specify the exact text. 4. **Action Sequence**: For multi-step tasks, determine the correct order of operations. Your question must: - **Embed** a clear user task goal (e.g., "The user wants to add a new contact with name X and phone Y"). - Ask: "To achieve this goal, which of the following description is true?" - Provide **four** answer options (A–D), at least one option should describe a full sequence of actions, it could be the correct one or a distractor. - **One** correct sequence. - **Three** distractors that each violate at least one of: - Wrong action type on a step. - Missing a critical step. - Steps in the incorrect order. - Wrong target element. - Make options concise but **nuanced**—avoid obvious mistakes. **Structure your reply with NO extra text**: question: {your question embedding the user goal} options: A. {step1 → step2 → . . . } B. {. . . } C. {. . . } D. {. . . } answer: {Correct option letter} type: Action Prediction

Table 8: Prompt for AP Dimension

Prompt for AEL Dimension

You are an AI visual assistant. You are given a single screenshot captured from a mobile UI during user interaction. Your task is to design one multiple-choice question that evaluates the **Absolute Element Location Ability**. Specifically, you should strictly follow these guidelines:
 	
1. **Question Description**:
- Clearly specify the element to be located (e.g., "Please determine the position of the blue button on the screen").
- Ask the model to analyze the general area of this element within the global coordinate system.
 	
2. **Reference Layout Structure**:
- Prompt the model to consider the overall interface structure (e.g., top navigation bar, central content area, bottom action bar) when making its determination.
- Guide the model to identify which section the element belongs to, such as status bar / toolbar / main content area / floating button area.
 	
3. **Absolute Position Description**:
- Require the model to use standardized regions:
  - Quadrant-based description: upper-left / lower-left / upper-right / lower-right;
  - Alternatively, a three-part division: top / middle / bottom.
- The question stem or options must explicitly use the above-mentioned descriptive terms to clearly define the location.
 	
Notice: Ensure that the questions you design for these tasks are answerable and the answers can be deduced from the GUI content. You must make each question as **difficult and nuanced** as possible, requiring careful visual and contextual reasoning. Avoid obvious or overly simple options. Minimize the repetition of the questioned objects as much as possible. Include plausible distractors for each question to increase the difficulty.
 	
For each given screenshot, create one multiple-choice question that tests one of the abilities mentioned above. Each question should have four answer options: one correct answer and three that are incorrect or irrelevant.
 	
Your reply must be structured like this, with **no extra explanation**:
 	
question: {your question}
options:
A. {Option A}
B. {Option B}
C. {Option C}
D. {Option D}
answer: {Correct option letter}
type: Absolute Element Location
 	
Please keep each answer as **concise and focused** as possible, and only include the five questions in this exact format. Only include questions that have definite answers.

Table 9: Prompt for AEL Dimension

Prompt for REL Dimension
You are an AI visual assistant. You are given a single screenshot captured from a mobile UI during user interaction. Your task is to design one multiple-choice question that evaluates the **Relative Element Location Ability**. This refers to evaluating the model's ability to reason about spatial relationships in the interface. Specifically, the model's ability to: - Determine the relative location of elements on the interface. **Notice**: Ensure that the questions you design for these tasks are answerable and the answers can be deduced from the GUI content. You must make each question as **difficult and nuanced** as possible, requiring careful visual and contextual reasoning. Avoid obvious or overly simple options. Minimize the repetition of the questioned objects as much as possible. Include plausible distractors for each question to increase the difficulty. For each given screenshot, create one multiple-choice question that tests one of the abilities mentioned above. Each question should have four answer options: one correct answer and three that are incorrect or irrelevant. Your reply must be structured like this, with **no extra explanation**: question: {your question} options: A. {Option A} B. {Option B} C. {Option C} D. {Option D} answer: {Correct option letter} type: Relative Element Location Please keep each answer as **concise and focused** as possible, and only include the five questions in this exact format. Only include questions that have definite answers.

Table 10: Prompt for REL Dimension

Prompt for WF Dimension

You are an AI visual forensics analyst specializing in mobile UI screenshots. Design ONE expert-level multiple-choice question that rigorously tests **Widget Function Perception Ability** with these constraints:

Strict visual evidence requirements:
- All answers MUST be provable from explicit visual evidence
- Absolutely NO speculation beyond what's visible
- Correct answers require synthesizing $\geq$3 distinct visual cues
- Reject any interpretation not confirmed by:
  1. Standard platform conventions
  2. Explicit visual affordances (shadows, highlights, depth cues)
  3. State indicators (color coding, iconography, text labels)
  4. Spatial relationships to adjacent elements

Core ability focus (evidence-based):
- MUST synthesize $\geq$3 distinct visual cues:
  1. Primary text labels (e.g., "Weather", "Reminders")
  2. Icon semantics (standard meanings only)
  3. Data representations (charts, progress bars)
  4. Contextual positioning (status bar vs. home screen)
- BANNED:
  1. Speculation beyond visible elements
  2. Prior knowledge of specific apps

Question design requirements:
- Ambiguous but decodable visual patterns (e.g., semi-transparent overlay on a search icon requiring icon shape, faded color, and nearby label)
- Compound state indicators (e.g., lock icon + greyed-out button requiring icon meaning and color state)
- Conflicting affordances requiring prioritization (e.g., send arrow and trash icon in the same area)
- are platform-specific edge cases (e.g., Android 11 share button long-press reveals hidden menu)

Your reply must be structured like this, with no extra explanation:

question: {your question}
options:
A. {Option A}
B. {Option B}
C. {Option C}
D. {Option D}
answer: {Correct option letter}
type: Widget Function

Table 11: Prompt for WF Dimension

Prompt for WI Dimension

You are an AI visual assistant. You are given a single screenshot captured from a mobile UI during user interaction. Your task is to design ONE multiple-choice question that evaluates the model's **Widget Interaction Perception Ability**, defined as inferring how users can interact with visible widgets by analyzing the given mobile UI screenshot. Specifically, the ability to:

1. **Identify Interactive Elements**
   Recognize actionable widgets (buttons, sliders, toggles, input fields, etc.) and distinguish them from static elements.
2. **Predict Interaction Methods**
   Determine valid operation types for each widget (tap, double-tap, long-press, swipe, pinch, etc.).
3. **Anticipate Interaction Outcomes**
   Foresee the immediate results of interactions, including:
   - Interface transitions (e.g., opening a settings panel)
   - State changes (e.g., toggle switching)
   - Function executions (e.g., alarm creation)
4. **Understand Practical Utility**
   Explain how the interaction solves real-world problems or enhances convenience, such as:
   - "Clicking '+' on clock widget enables quick alarm setting"
   - "Swiping down the corner slider adjusts screen brightness"
   - "Tapping screen time widget reveals detailed usage analytics"

Notice: Ensure that the questions you design for these tasks are answerable and the answers can be deduced from the GUI content. You must make each question as difficult and nuanced as possible, requiring careful visual perception and contextual reasoning. Avoid obvious or overly simple options. Include plausible distractors for each question to increase the difficulty. For each given screenshot, create one multiple-choice question that tests one of the abilities mentioned above. Each question should have four answer options: one correct answer and three that are incorrect but closely relevant. Distractors should be designed to be tempting yet contain subtle mistakes drawn from the interface that are difficult to detect. The options should include at least one non-interactive distractor (static element misuse). Your reply must be structured like this, with **no extra explanation**:

question: {your question}
options:
  A. {Option A}
  B. {Option B}
  C. {Option C}
  D. {Option D}
answer: {Correct option letter}
type: Widget Interaction

Please keep each answer as concise and difficult as possible, and only structured in this exact format. Only include questions that you can answer confidently based on the image content.

Table 12: Prompt for WI Dimension

Prompt for RI & SI Dimensions
You are an AI assistant generating multiple-choice questions to evaluate understanding of mobile UI task flows. The following screenshots capture a short interaction sequence in a mobile app. The correct user goal is: "{correct_goal}" Your task is to generate **three incorrect but plausible alternative user goals** that could reasonably be mistaken for what the user is trying to do, based on the visual context. **Guidelines**: 1. Each option should **look like a real user task** — it doesn't need to match the exact phrasing or grammar of the correct goal, but should feel natural and fit within the app's context (e.g., settings, messaging, shopping, file management). 2. Focus on **plausible misinterpretations**: the user might think the person is doing something related but different — changing a setting instead of deleting, sharing instead of saving, searching for a contact instead of calling, etc. 3. Vary the **action**, **target**, or **intent**: use different verbs (edit, find, enable, share, create, view, check, etc.) or objects (a message, a photo, an account, a notification, etc.) that appear or could appear in the interface. 4. It's okay if the grammar is slightly informal or simplified — real users don't always phrase tasks perfectly. 5. Do **not** include explanations, reasoning, or meta-comments (e.g., no "attempt to", "mistake", "analyze"). 6. Make sure the options are clearly different from the correct goal, but still **contextually grounded** in the screenshots. Only output the three distractors in the following format: A. ... B. ... C. ...

Table 13: Prompt for RI & SI Dimensions

Data Collecting Guidelines

Annotators are required to collect screenshots in the following languages: Chinese (ZH), English (EN), French (FR), Russian (RU), Thai (TH), and Japanese (JA).

a. First, check whether each app/website supports the above languages.

b. Select an app and begin capturing screenshots. For each app/website, capture as many different screens as possible, each corresponding to one of the six language environments listed above. Try to ensure that the screens represent different scenarios.

c. Next, to maintain consistency, check the initial screenshots based on the following three guidelines:

   i. It is recommended to select as many screenshots as possible, as many might be discarded after checking. Ensure that the final dataset has at least 10 different scenes for each app.

   ii. Consistency must be maintained, meaning that apart from the language, the visual style, background coherence, and text formatting should remain consistent. For example, in a weather app, the temperature unit should be the same across different languages. If the app includes recommended content (e.g., search recommendations in a browser), ensure that the recommendations remain consistent when switching languages. Additionally, if searching within a browser, the search terms should be translated according to the language (e.g., searching "apple" in English should correspond to "pingguo" in Chinese). Make sure the input method is set to the correct language as well.

   iii. An example of a valid scenario is as follows: ...the 6 screenshots from the same scenario show only differences in language, while the layout remains almost identical. Such data should be retained. An example of invalid data that should be discarded: the screenshots show significant issues that hinder interface comprehension, such as inconsistent text content across languages, mixed languages, and layout issues that obstruct understanding.

d. After checking, use GPT-4o to automatically generate questions. The recommended prompt template can be found at the end of the task instructions.

e. After generating the questions, manually review and check them based on the following aspects, then either ask GPT-4o to regenerate the questions or design them manually:

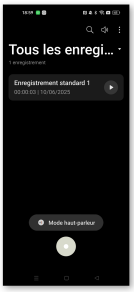
   i. For question design, check out Q&A pairs that are factually incorrect, have mismatched objects/screenshots, or have questions that are too easy or too difficult.

   ii. For option design, check out Q&A pairs where the correct option is inaccurate, incorrect options lack sufficient distractor quality, or the options are misleading.

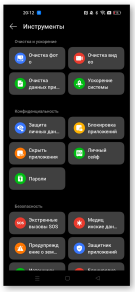
f. Next, annotators need to input the semantically parallel non-English screenshots along with the translation prompts into GPT-4o, allowing the model to translate all the candidate VQA pairs from the English list into the target languages.

g. Annotators must check each translated VQA pair in the target language to ensure cross-lingual consistency. If any discrepancies are found, the translation should be redone or the example discarded. This process will result in the creation of a complete dataset.

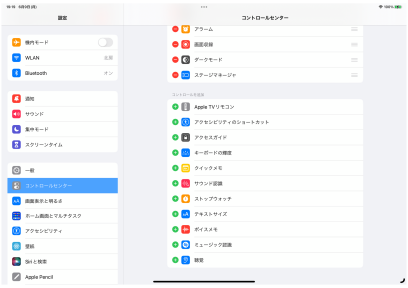
Table 14: Data Collecting Guidelines



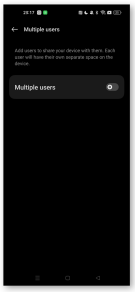
Question: *Quel est le résultat immédiat le plus probable de l'appui sur le bouton de lecture à côté de "Enregistrement standard 1" ?*
Options:
A. L'enregistrement commencera à jouer et la synchronisation cloud reprendra.
B. L'enregistrement commencera à jouer avec un retour système indiquant que la synchronisation est toujours en pause.
C. L'enregistrement commencera à jouer et le système passera automatiquement en "Mode haut-parleur".
D. L'enregistrement commencera à jouer mais sans retour visuel ou audio.
Answer: *B* ☑
Type: *Action Understanding*
Predict: *C* ✗



Question: *Каков непосредственный результат выбора опции "Очистка видео" на этом экране?*
Options:
A. Переход на страницу, где видео можно удалять по отдельности.
B. Открывается всплывающее окно для настройки пределов размера видеофайлов для очистки.
C. Запускается автоматический процесс оптимизации видео.
D. Инициируется очистка видео по всей системе без запросов пользователя.
Answer: *A* ☑
Type: *Action Understanding*
Predict: *D* ✗



Question: *「アクセスガイド」の横にある「+」アイコンをタップするとどうなりますか？*
Options:
A. 「アクセスガイド」が「コントロールを追加」セクションのトップに移動します。
B. 「アクセスガイド」がコントロールセンターに追加されたことを確認する成功メッセージが表示されます。
C. 「アクセスガイド」が「コントロールを追加」セクションに追加され、コントロールセンターからアクセスできるようになります。
D. 「アクセスガイド」の横にある「+」アイコンが「-」に変わり、すでに追加されていることを示します。
Answer: *D* ☑
Type: *Action Understanding*
Predict: *C* ✗



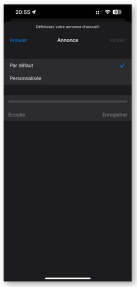
Question: *What immediate change will occur in the interface state if the user toggles the "Multiple users" option to its active state?*
Options:
A. The interface will display a list of currently set users along with options to add or delete users.
B. A warning message will appear indicating potential privacy risks when using multiple users.
C. The screen will navigate to a user setup wizard to configure new user profiles.
D. An error message will be displayed indicating that user setup is unavailable at the moment.
Answer: *A* ☑
Type: *Action Understanding*
Predict: *C* ✗

Question: *在当前界面状态下，关闭"私密保险箱"功能的直接效果是什么？*
Options:
A. "私密保险箱"内容将无法访问，除非输入密码。
B. 将出现确认对话框以验证用户禁用该功能的操作。
C. 系统将在更改生效前显示加载指示器。
D. 切换按钮将短暂闪烁以指示状态更新成功。
Answer: *A* ☑
Type: *Action Understanding*
Predict: *B* ✗

Question: *[illegible]*
Options:
A. [illegible]
B. [illegible]
C. [illegible]
D. [illegible]
Answer: *A* ☑
Type: *Action Understanding*
Predict: *B* ✗

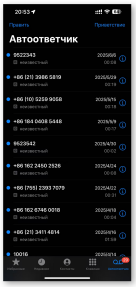
Figure 6: Examples of incorrect responses by LVLMs in AU dimension across 6 language settings.

Question: *The user wants to share their weekly step and distance summary on social media. To achieve this goal, which of the following description is true?*
Options:
A. Tap "Sharing" → Choose social media app → Tap "Share"
B. Tap "Summary" → Tap the day with the highest step count → Tap "Sharing"
C. Tap the "W" tab → Tap "Sharing" → Choose social media app → Tap "Post"
D. Tap "Distance" under by day view → Tap "Sharing" → Choose contact → Tap "Send"
Answer: *C* ☑
Type: *Action Prediction*
Predict: *A* ✗

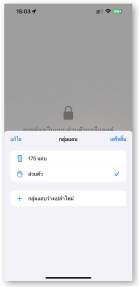


Question: *L'utilisateur souhaite définir un message sortant personnalisé en l'enregistrant. Pour atteindre cet objectif, laquelle des descriptions suivantes est vraie ?*
Options:
A. Appuyez sur "Personnalisée", appuyez sur "Enregistrer", enregistrez le message, appuyez sur "Valider".
B. Appuyez sur "Par défaut", appuyez sur "Écouter", appuyez sur "Valider".
C. Appuyez sur "Personnalisée", tapez le message, appuyez sur "Valider".
D. Appuyez sur "Personnalisée", appuyez sur "Enregistrer", appuyez sur "Annuler".
Answer: *A* ☑
Type: *Action Prediction*
Predict: *C* ✗

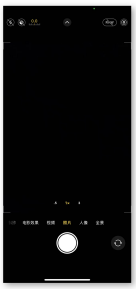
Question: *ユーザーは水曜日の身体活動の詳細な履歴を見たいと考えています。この目的を達成するために、次の説明のうちどれが正しいですか？*
Options:
A. *「水曜日」をハイライトセクションでタップ→活動の詳細を表示。*
B. *ハイライトセクションの「水曜日」項目を左にスワイプして詳細にアクセス。*
C. *下部の「概要」ボタンをタップ→ハイライトセクションで「水曜日」をタップ。*
D. *ハイライトセクションで「水曜日」を長押し→活動の詳細を表示。*
Answer: *A* ☑
Type: *Action Prediction*
Predict: *B* ✗



Question: *Пользователь хочет удалить голосовую почту с номера +86 (21) 3986 5819 с устройства. Чтобы достичь этой цели, какое из следующих описаний является верным?*
Options:
A. Нажмите кнопку "Править" → Проведите влево по +86 (21) 3986 5819 → Нажмите "Удалить".
B. Нажмите кнопку "Править" → Нажмите значок "i" рядом с +86 (21) 3986 5819 → Нажмите "Удалить".
C. Длительное нажатие на +86 (21) 3986 5819 → Нажмите "Удалить".
D. Проведите влево по +86 (21) 3986 5819 → Нажмите "Удалить".
Answer: *D* ☑
Type: *Action Prediction*
Predict: *C* ✗

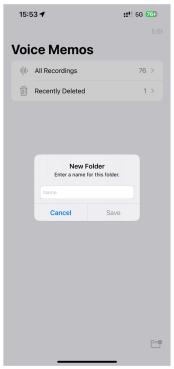


Question: *[illegible]*
Options:
A. [illegible] → [illegible].
B. [illegible] → [illegible].
C. [illegible] → [illegible].
D. [illegible] "175 [illegible]" → [illegible].
Answer: *A* ☑
Type: *Action Prediction*
Predict: *B* ✗



Question: *用户想切换到“视频”模式来录制视频。为实现此目标，以下哪项描述是正确的？*
Options:
A. 在“照片”模式上向左滑动以切换到“视频”模式。
B. 点击“视频”标签以切换到“视频”模式。
C. 长按快门按钮以切换到“视频”模式。D. 在“照片”模式上向右滑动以切换到“视频”模式。
Answer: *A* ☑
Type: *Action Prediction*
Predict: *B* ✗

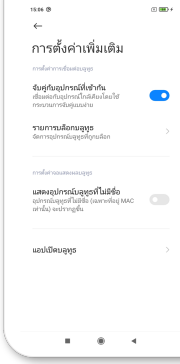
Figure 7: Examples of incorrect responses by LVLMs in AP dimension across 6 language settings.



Question: *Please determine the position of the "Save" button on the screen.*
Options:
A. Upper-left quadrant in the floating button area
B. Lower-right quadrant in the main content area
C. Bottom section in the toolbar area
D. Lower-right quadrant in the floating button area.
Answer: *D* ☑
Type: *Absolute Element Location*
Predict: *B* ✗

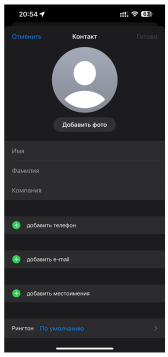
Question: *Veuillez déterminer la position du bouton d'information bleu "i" sur l'écran*
Options:
A. Quadrant supérieur gauche dans la zone de contenu principal
B. Quadrant inférieur droit dans la zone de bouton flottant
C. Quadrant supérieur droit dans la barre d'outils
D. Quadrant inférieur gauche dans la zone de contenu principal
Answer: *A* ☑
Type: *Absolute Element Location*
Predict: *C* ✗

Question: *緑色の「サイクリング（屋外）」ボタンの画面上の位置を決定してください。*
Options:
A. 左上の四分円、ツールバーエリア
B. 右下の四分円、下部アクションバー
C. 中央領域、メインコンテンツエリア
D. 右上の四分円、トップナビゲーションバー
Answer: *C* ☑
Type: *Absolute Element Location*
Predict: *D* ✗



Question: *[illegible]*
Options:
A. [illegible]
B. [illegible]
C. [illegible]
D. [illegible]
Answer: *B* ☑
Type: *Absolute Element Location*
Predict: *D* ✗

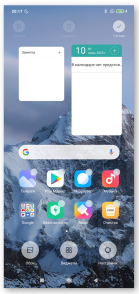
Question: *请确定屏幕上“扫描条码”按钮的位置。*
Options:
A. 右上象限
B. 中间部分
C. 左下象限
D. 底部部分
Answer: *D* ☑
Type: *Absolute Element Location*
Predict: *B* ✗



Question: *Пожалуйста, определите положение кнопки "Отменить" на экране.*
Options:
A. Верхний левый квадрант, строка состояния
B. Верхний левый квадрант, панель инструментов
C. Нижний левый квадрант, основная область контента
D. Средняя часть, панель инструментов
Answer: *B* ☑
Type: *Absolute Element Location*
Predict: *A* ✗

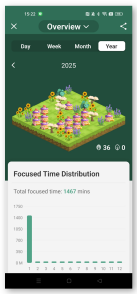
Figure 8: Examples of incorrect responses by LVLMs in AEL dimension across 6 language settings.

Question: *Quelle est la position relative de l'icône 'Play Store' par rapport à l'icône 'Chrome' ?*
Options:
A. Directement au-dessus
B. À gauche immédiate
C. Directement en dessous
D. À droite immédiate
Answer: *A* ☑
Type: *Relative Element Location*
Predict: *B* ✗



Question: *В отношении значка приложения 'Безопасность', где находится значок приложения 'Play Маркет' на интерфейсе?*
Options:
A. Прямо над значком приложения 'Безопасность'.
B. Прямо слева от значка приложения 'Безопасность'.
C. Диагонально выше и справа от значка приложения 'Безопасность'.
D. Прямо справа от значка приложения 'Безопасность'.
Answer: *A* ☑
Type: *Relative Element Location*
Predict: *B* ✗

Question: *「録音の上限」オプションのすぐ上にある設定はどれですか？*
Options:
A. 指定した番号を録音
B. 不明な番号を記録
C. すべての通話を録音
D. 通話録音を表示
Answer: *C* ☑
Type: *Element Location*
Predict: *A* ✗

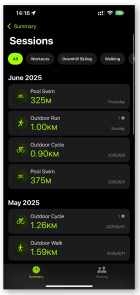


Question: *Which element is located at the top-left corner of the screenshot?*
Options:
A. Total focused time: 1467 mins
B. Year selection dropdown
C. Overview
D. Back arrow
Answer: *C* ☑
Type: *Relative Element Location*
Predict: *D* ✗

Question: *[illegible]?*
Options:
A. [illegible]
B. [illegible]
C. [illegible]
D. [illegible]
Answer: *A* ☑
Type: *Relative Element Location*
Predict: *B* ✗

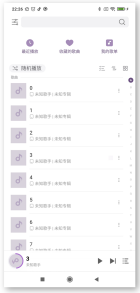
Question: *“左滑删除”开关相对于“恶意网址识别”开关的位置是什么？*
Options:
A. 在同一垂直列中直接在其下方。
B. 在同一垂直列中直接在其上方。
C. 在同一水平行中直接在其右侧。
D. 在同一水平行中直接在其左侧。
Answer: *A* ☑
Type: *Relative Element Location*
Predict: *D* ✗

Figure 9: Examples of incorrect responses by LVLMs in REL dimension across 6 language settings.



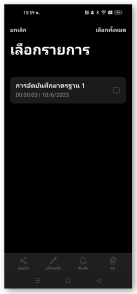
Question: *Based on the visible cues from the screenshot, what is the primary function of the "Summary" tab highlighted at the bottom of this mobile UI?*
Options:
A. To present detailed statistics about individual workout sessions
B. To provide a timeline view of various activities in chronological order
C. To offer a selection of recommended workout sessions
D. To display a summary of total distances and activities grouped by months
Answer: *B* ☑
Type: *Widget Function*
Predict: *D* ✗

Question: *En se basant sur les indices visuels présentés dans la capture d'écran, laquelle des fonctions suivantes peut être directement activée à partir des widgets visibles sans aucune navigation supplémentaire ?*
Options:
A. Régler une nouvelle alarme dans le widget "Horloge"
B. Changer la fréquence sur le widget "Radio FM"
C. Calculer une expression mathématique dans le widget "Calculatrice"
D. Gérer les autorisations des applications dans la fonctionnalité "Gérer les applications"
Answer: *B* ☑
Type: *Widget Function*
Predict: *A* ✗

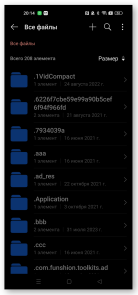


Question: *根据手机界面截图中的视觉证据，当前播放界面的状态指示了什么小部件功能？*
Options:
A. 歌曲播放已暂停，启用了随机播放模式。
B. 正在播放，播放编号为"3"的歌曲，随机播放模式已禁用。
C. 随机播放模式已激活，播放已停止，目前没有歌曲正在播放。
D. 播放编号为"2"的歌曲，随机播放模式已禁用，重复模式已启用。
Answer: *B* ☑
Type: *Widget Function*
Predict: *C* ✗

Question: *このUIで、表示されている状態インジケーターに基づいて現在無効になっている設定はどれですか？*
Options:
A. 非表示のファイルを表示する
B. スクリーンキャプチャーとカメラファイルを「最近のファイル」に表示
C. ファイルを開くためのデフォルトアプリを設定する
D. 機能の管理
Answer: *A* ☑
Type: *Widget Function*
Predict: *C* ✗

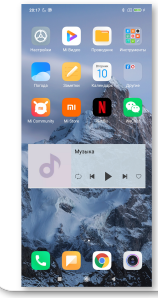


Question: [illegible]
Options:
A. [illegible]
B. [illegible]
C. [illegible]
D. [illegible]
Answer: *D* ☑
Type: *Widget Function*
Predict: *C* ✗



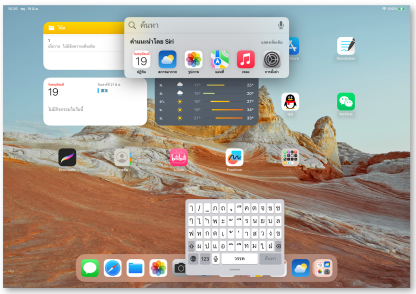
Question: *Какая функция виджета может быть выведена из скриншота интерфейса на основе визуальных доказательств из текстовых меток, семантики иконок и контекстного расположения?*
Options:
A. Функция поиска активна, что видно по выделенной иконке лупы, расположенной в "панели действий".
B. Функция сортировки по размеру активна, что видно по метке "Размер" и стрелке рядом с заголовком.
C. Функция добавления активна, что видно по выделенной иконке плюса в "панели действий" и её стандартной функции.
D. Функция удаления активна, что видно по иконке "контекстного меню", предлагающей дополнительные варианты действий.
Answer: *C* ☑
Type: *Widget Function*
Predict: *B* ✗

Figure 10: Examples of incorrect responses by LVLMs in WF dimension across 6 language settings.



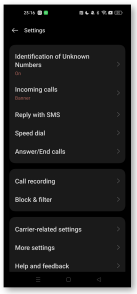
Question: *В музыкальном виджете, какое действие вы выполните, чтобы начать воспроизведение музыки с начала?*
Options:
A. Нажмите кнопку воспроизведения
B. Нажмите кнопку повтора
C. Удерживайте значок музыки
D. Дважды нажмите кнопку пропуска назад
Answer: *D* ☑
Type: *Widget Interaction*
Predict: *A* ✗

Question: *用户点击此移动界面截图中的“取消”按钮时会发生什么？*
Options:
A. 应用程序将继续收集数据并与云同步。
B. 提现将被确认，停止转录和云同步功能。
C. 用户将被重定向到详细说明隐私政策的帮助页面。
D. 将出现错误消息，指出该操作不允许。
Answer: *B* ☑
Type: *Widget Interaction*
Predict: *D* ✗



Question: *Ω≈ᵇ⁄∫ʒÁ¥ÃÂ“®¢¢“›≠µ„ŒŒ∫”À ”øŒ¢“»” ”¬µ¢⁄”∓∫⁄∫⁄”·¥¬Siri •ŒŒ„∿?*
Options:
A. ƒ°¥¢“›b¬“¢≥ŒŒ¢“»·¥∼≈ƒŒ≠ŒŒ%µ∫”∓Ã∫¿¡ ƒŒŒ≈GPS °®⁄”∫
B. Àœ“›≠®ƒ⁄ŒŒÀ ”øŒ¢“»„ÀŒŒ®⁄”Ã°Ã°¥“ÃÂ∫¢
C. ∓A¥∼Ã∫¢⁄∓ ŒŒ∑°øŒŒ¢“›Ρ›°ŒÈÃ¿Ã⁄Ã°Ã°Àª∫∑∑ÂÁ·Â
D. ƒb⁄A≠ƒŒÀ ”øŒ¢“»„°¬“Ã∫ç®ŒÈ≈¢ƒb⁄Œ“›£Ε¢øŒŒ∫¢›≠ƒ⁄Ê Answer: A ☑
Type: *Widget Interaction*
Predict: *C* ✗

Question: *Quelle interaction avec le paramètre "Identification des numéros inconnus" changerait probablement son statut actuel ?*
Options:
A. Appuyer n'importe où sur le texte "Identification des numéros inconnus".
B. Balayer vers la droite sur le libellé "Activé" pour désactiver la fonctionnalité.
C. Appui long sur l'option "Appels entrants" pour des paramètres étendus.
D. Appuyer sur la flèche à côté de "Répondre par SMS" pour accéder aux paramètres détaillés.
Answer: *A* ☑
Type: *Widget Interaction*
Predict: *B* ✗



Question: *Which interaction with the "Identification of Unknown Numbers" setting would most likely change its current status?*
Options:
A. Tapping anywhere on the "Identification of Unknown Numbers" text.
B. Swiping right over the "On" label to disable the feature.
C. Long-pressing the "Incoming calls" option for expanded settings.
D. Tapping the arrow next to "Reply with SMS" to access detailed settings.
Answer: *A* ☑
Type: *Widget Interaction*
Predict: *B* ✗

Question: *表示されているUIに従って、カメラモードを「ポートレート」に適切に変更するインタラクション方法はどれですか？*
Options:
A. 下部のモード選択オプションを左にスワイプする。
B. 「写真」ラベルを切り替わるまで繰り返しタップする。
C. 画面をピンチインしてビューを縮小する。
D. 中央のキャプチャボタンを変更されるまで長押しする。
Answer: *A* ☑
Type: *Widget Interaction*
Predict: *B* ✗

Figure 11: Examples of incorrect responses by LVLMs in WI dimension across 6 language settings.

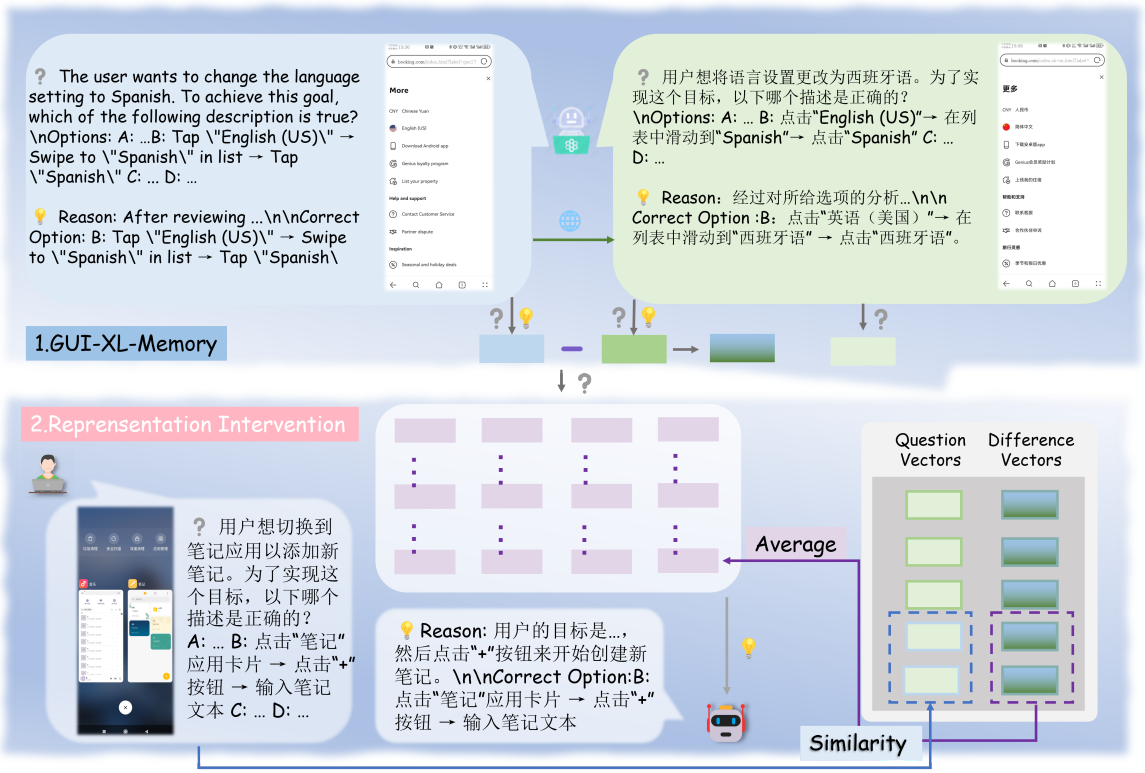


Figure 12: An Overview of our **GUI-XLI** method in §4.2. **Step 1 GUI-XL-Memory**: We sample semantically parallel VQA pairs to form entries in **GUI-XL-Memory**. **Step 2 Cross-lingual Representation Intervention**: When answering non-English questions, related entries are retrieved to calculate difference vectors and then injected to certain layer as intervention to add P&R capabilities to non-English settings.