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

Anonymous ACL submission

Abstract

Large Vision–Language Models (LVLMs) have shown strong potential as multilingual Graphical User Interface (GUI) agents, as evidenced by existing GUI benchmarks. However, these benchmarks exhibit two primary limitations: (1) although Perception and Reasoning (P&R) capabilities are fundamental for GUI agents, current benchmarks lack fine-grained diagnostics to identify which specific capabilities lead to task failures, hindering targeted improvements; (2) existing benchmarks fail to provide a strictly aligned cross-lingual evaluation environment, introducing confounding factors that prevent isolating the language impact on GUI agent performance. To address these issues, we propose the **M**ultilingual **P&R GUI Bench**mark (**MPR-GUI-Bench**), featuring strictly aligned environments across six languages and eight fine-grained P&R tasks. Our benchmark reveals consistent P&R gaps between English and non-English settings, particularly on reasoning-intensive tasks. To leverage the superior English P&R capabilities for bridging cross-lingual gaps, we identify layers sensitive to language and propose **GUI-XLI**, a **GUI Cross-Lingual I**ntervention method that aligns non-English hidden states with their English counterparts at these layers during inference. Experiments show that GUI-XLI effectively reduces the cross-lingual gaps, with an average gain of 6.5% in non-English settings.

1 Introduction

Rapidly evolving Large Vision-Language Models (LVLMs) have shown potential as multilingual GUI agents, as demonstrated by recent benchmarks. Existing GUI benchmarks, broadly categorized into interactive and static ones, suffer from two critical limitations. First, despite GUI Perception and Reasoning (P&R) capabilities being fundamental to real-world end-to-end competence (Zhang et al., 2025a; Xie et al., 2025; Qin et al., 2025), current benchmarks lack systematic and fine-grained assessment of these capabilities. Interactive benchmarks rely on holistic task success rates, obscuring failure causes, while static benchmarks lack structured P&R analysis. Second, existing benchmarks lack strictly aligned cross-lingual evaluation. Linguistic factors are underexplored in static benchmarks, while interactive benchmarks such as MacOSWorld (Yang et al., 2025a) inevitably introduce language-irrelevant variations (e.g., UI layouts and interaction trajectories), preventing isolation of language effects.

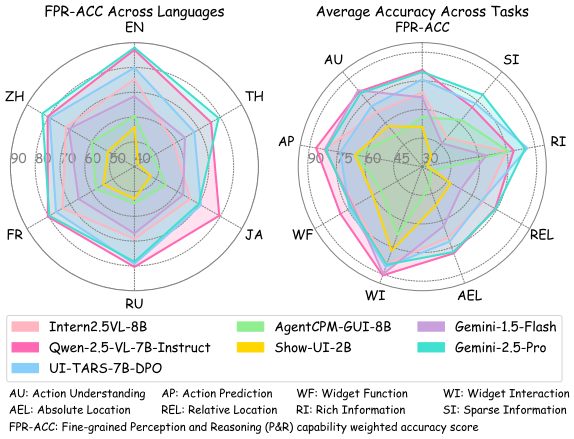


Figure 1: **Performance evaluation on MPR-GUI-Bench. (Left)** Multilingual comparison revealing a consistent gap between English and non-English settings across all baselines. **(Right)** Fine-grained P&R capability analysis across specific dimensions.

To bridge these gaps, we introduce the **M**ultilingual **P&R GUI Bench**mark (**MPR-GUI-Bench**). MPR-GUI-Bench is strictly aligned across six languages carefully chosen to balance common GUI languages and orthogonal coverage of different language families, spanning 39 scenarios and six device types with 12,936 samples. Building upon key perception and reasoning tasks proposed by prior works, we further incorporate end-to reasoning tasks that reflect real-world end-to-end scenarios, resulting in tasks hierarchically

Dataset	Languages							Cross-lingual alignment	Fine-grained Dimensions	Platform			textbfSize	Type
	EN	ZH	FR	RU	JA	TI	AR			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 with related benchmarks. Abbreviations: **Web./Mob./Desk.** (Website/Mobile/Desktop). Under **Type**, *env.* represents interactive environments vs. static *dataset*. **Cross-lingual alignment** indicates whether the benchmark provides strictly aligned tasks across different languages for controlled evaluation.

organized into three levels and eight fine-grained P&R dimensions. As shown in Figure 1, evaluations across seven advanced LVLMs reveal consistent non-English performance gaps relative to English, particularly in reasoning tasks.

To leverage the superior P&R capabilities of English for improving non-English performance, we identify critical layers most sensitive to linguistic factors during inference and propose a **GUI Cross-L**ingual **I**ntervention method (**GUI-XLI**), which steers non-English representations toward their English counterparts. GUI-XLI achieves a 6.5% average performance gain with negligible inference latency. Consistent improvements under explicit reasoning-chain generation indicate that GUI-XLI aligns cross-lingual reasoning patterns at the representational level rather than acting as a prompting artifact.

Our contributions are summarized as follows:

- We present **MPR-GUI-Bench**, the first multilingual benchmark to systematically evaluate fine-grained GUI P&R capabilities.
- We present a comprehensive analysis of the LVLMs for GUI agent from the perspectives of P&R and cross-lingual capabilities.
- We propose **GUI-XLI**, a training-free representation engineering method that effectively mitigates cross-lingual P&R capability gaps.

2 Related Work

GUI agent benchmarks As presented in Table 1, existing GUI agent benchmarks generally fall into two categories: interactive environments and static datasets (Nguyen et al., 2024a). Interactive environments are widely regarded as better reflections of real-world end-to-end scenarios (Rawles et al., 2024; Wang et al., 2024; Chen et al., 2025). However, they typically evaluate performance through task completion rates, treating each trajectory as a single unit. Such a coarse-grained metric obscures the underlying P&R skills required for success, making it difficult to diagnose failures or guide targeted improvements. Conversely, static datasets such as ScreenSpot (Cheng et al., 2024), GUI-World (Chen et al., 2024a) and GUI-Odessey (Lu et al., 2024) isolate particular capabilities and demonstrate the value of decomposing GUI tasks into foundational skills, but they lack a systematic and fine-grained taxonomy. Moreover, multilingual capability, a prerequisite for global deployment, remains largely under-explored. While interactive benchmarks like MacOSWorld (Yang et al., 2025b) include language factors, their dynamic nature introduces complexity, hindering controlled, aligned comparison across languages. These limitations highlight the need for a systematic, multilingual, fine-grained P&R evaluation setup, which motivates the design of MPR-GUI-Bench.

Cross-lingual Representation Alignment Recent studies in LVLMs have investigated the internal mechanisms of cross-lingual transfer. Research indicates that inputs with identical semantics but different languages often yield substantially different distributions within the model's latent space (Chang et al., 2022; Peng et al., 2025; Zhao et al., 2024). This divergence correlates with performance inconsistencies across languages. Meanwhile, representation engineering research demonstrates that controlled interventions on hidden states can effectively steer model behavior (Andy Zou, 2023; Li et al., 2024; Turner et al.,

Figure 2: Hierarchical composition of MPR-GUI-Bench. The structure expands from P&R (inner) to four domains (middle) and eight fine-grained dimensions (outer). Gray numbers indicate image counts.

2024). Inspired by these findings, our proposed GUI-XLI treats the cross-lingual discrepancy vector as an explicit optimization direction, aligning under-performing languages toward the distribution of better-performing ones, thereby enhancing multilingual GUI P&R consistency.

3 MPR-GUI-Bench

Existing GUI benchmarks have mostly neglected fine-grained P&R capabilities, leading to difficulties in their development. Moreover, even fewer studies have focused on these capabilities in multilingual settings. To this end, We propose **MPR-GUI-Bench**, the first benchmark to systematically evaluate the fine-grained P&R capabilities required by GUI tasks in multilingual environments.

3.1 Data Source

As shown in Figure 3, we collect parallel screenshots across 6 languages (English, Chinese, French, Russian, Japanese and Thai), spanning 39 distinct real-world GUI scenarios on two operating systems (iOS and Android) and 6 mobile device models.

3.2 Task Definitions

As shown in Figure 2, we define 8 fine-grained dimensions derived from key P&R capabilities mentioned in prior works, supplemented by two dimensions reflecting end-to-end agent performance, organized into two primary categories: (1) **perception capabilities**, covering the perception of interactive components (widgets) and user actions; and (2) **reasoning capabilities**, which encompass spatial reasoning and end-to-end reasoning based on integrating fundamental P&R capabilities. The eight 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 the capabilities to synthesize long interaction histories, and integrate all fine-grained P&R dimensions to infer user intention. It simulates real-world interactive end-to-end GUI tasks.
- **Sparse Information (SI)** evaluates the capability to infer user intent from shorter screenshot sequences with minimal cues. It simulates more challenging real-world end-to-end GUI tasks, reflecting the model's P&R capability upper bound in more complex GUI tasks.

3.3 Benchmark Construction Pipeline

As illustrated in Figure 3, to balance scalability with accuracy, we adopt a semi-automated approach followed by human verification:

Step 1: Screenshot Collection Annotators curated a rigorously aligned dataset across six languages. While end-to-end reasoning tasks feature fewer samples due to the complexity of aligning multi-image sequences, they offer deeper diagnostic insights than standard single-image tasks. Detailed information about the annotators' backgrounds and guidelines are provided in the Appendix A.1, A.2.

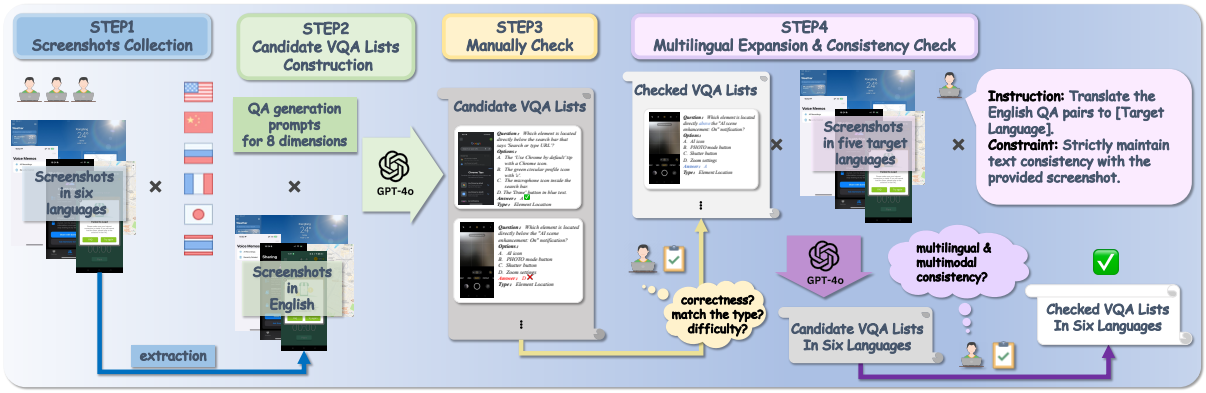


Figure 3: Construction pipeline of MPR-GUI-Bench, as described in §3.3: **Step 1**: Collect parallel screenshots in six languages; **Step 2**: Generate candidate English VQA lists via GPT-4o; **Step 3**: Verify quality manually; **Step 4**: Expand multilingual data with GPT-4o and perform cross-lingual and cross-modal consistency checks.

Step 2: Candidate Visual Question Answering(VQA) Lists Construction Following the definitions in Section 3.2, we construct prompts that enforce strict structural, lexical constraints, limiting the model's generative freedom and reducing the imprinting of model-specific phrasing style. GPT-4o (OpenAI et al., 2024) is then used to generate the English VQA candidates based on the prompts, which are provided in Appendix A.3.

Step 3: Manually Check Six annotators independently verify the candidate VQA list across three dimensions: question formulation, answer correctness, and distractor quality. To mitigate GPT-4o's stylistic artifacts, annotators manually reorder options and rephrase question patterns. We further validate the removal of source-specific bias by comparing VQA items generated by Gemini-2.5-Pro and GPT-4o. Evaluation with three models reveals no consistent performance discrepancies between the two sources, suggesting the benchmark is robust to generator choice. Detailed guidelines, verification results, and inter-rater agreement statistics are provided in the Appendix A.5.

Step 4: Multilingual Expansion & Consistency Check We employ GPT-4o for translation. Unlike text-only machine translation systems, which often generate synonyms causing lexical mismatches and subsequent GUI grounding failures, GPT-4o's multimodal capabilities ensure the translated text aligns faithfully with the on-screen text. Analyses of translation fidelity and inter-rater consistency are provided in the Appendix A.4, A.5.

3.4 Evaluation Metrics

Samples follow a standard four-option single-choice format, evaluated via exact match following (Chen et al., 2024a). We eschew free-form responses because real-world GUI tasks inherently require precise understanding and selection within a constrained action space; free-form outputs often introduce unnecessary evaluation noise. To represent the overall performance considering dimension difficulty, we define a fine-grained P&R weighted accuracy score FPR-ACC:

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

where α_i and w_i denote the accuracy and weight for dimension i, respectively. Detailed descriptions are provided in Appendix A.7.

3.5 Experiment Setup

Baseline We select baselines from three model types: (1) **Open-source LVLMs**: Intern2.5VL-8B (Chen et al., 2024b), Qwen-2.5-VL-7B-Instruct (Bai et al., 2025); (2) **Closed-source LVLMs**: Gemini-1.5-flash (Gemini Team et al., 2024) and Gemini-2.5-Pro (Comanici et al., 2025); (3) **Multimodal GUI agents**: UI-TARS-7B-DPO (Qin et al., 2025), AgentCPM-GUI (Zhang et al., 2025b), Show-UI-2B (Lin et al., 2024). Among models with known parameter sizes, we select versions smaller than 8B, as lightweight models better support on-device deployment for GUI agents, which is critical for preserving user privacy. The evaluation results for three extra models are provided in Appendix A.9.

Implementation Details Our evaluation is conducted on 8 × NVIDIA A100 GPUs.

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
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
Multimodal GUI Agents										
UI-TARS-7B-DPO	EN	76.2	**79.5**	**80.8**	**88.8**	74.0	71.3	80.0	**84.0**	**79.9**
	ZH	**77.9**	74.9	72.6	88.5	**77.9**	66.1	**96.0**	76.0	79.4
	FR	67.5	70.8	77.0	85.5	69.7	70.5	84.0	80.0	76.5
	RU	72.1	75.1	73.7	87.4	71.0	**72.7**	88.0	**84.0**	79.2
	JA	67.2	73.5	72.3	85.8	70.5	63.1	84.0	58.0	71.0
	TH	67.2	67.5	69.0	79.2	67.8	51.9	84.0	58.0	67.9
AgentCPM-GUI-8B	EN	**64.8**	**71.6**	**67.1**	**82.8**	**57.1**	**39.3**	80.0	72.0	**60.4**
	ZH	49.2	64.8	58.1	70.5	41.8	32.8	**88.0**	**88.0**	59.9
	FR	54.6	64.5	57.3	69.1	40.0	36.1	72.0	60.0	57.9
	RU	43.4	62.3	53.7	62.6	46.0	32.2	72.0	56.0	54.8
	JA	51.4	68.6	50.1	66.4	33.8	37.4	76.0	48.0	54.5
	TH	45.6	59.3	42.7	56.0	27.9	32.0	68.0	48.0	48.6
Show-UI-2B	EN	**67.8**	**73.8**	**64.7**	**84.7**	57.1	**62.3**	32.0	36.0	**55.8**
	ZH	62.8	66.1	61.4	82.2	**58.5**	45.6	32.0	36.0	52.3
	FR	59.0	67.2	61.6	79.0	55.5	48.4	**44.0**	**40.0**	54.4
	RU	55.7	68.9	61.6	77.6	50.3	48.1	40.0	**40.0**	52.9
	JA	54.6	64.2	57.3	75.4	44.5	42.6	28.0	36.0	47.7
	TH	46.7	54.6	53.2	62.6	41.5	36.1	20.0	36.0	41.8
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 English, Chinese, France, Russian, Japanese and Thai (EN, ZH, FR, RU, JA, TH). Background colors indicate the **best** and worst performance per setting.

Metric	Pearson r	Strength
RI vs. FPR-ACC	0.7373	High
SI vs. FPR-ACC	0.7152	High
Avg. vs. FPR-ACC	**0.7795**	**V. High**

Table 3: Correlation Analysis of End-to-End Reasoning Scores and FPR-ACC; Avg. denotes the mean of RI and SI scores.**V. High**: Very High.

3.6 Evaluation Result

From the evaluation result presented in Table 2, we draw conclusions in three key aspects:

Performance gap across languages & models Regarding languages, while English and Chinese yield dominant results across all baselines, performance degrades in low-resource settings (e.g., Thai), revealing a severe generalization bottleneck in current LVLMs for GUI tasks.

Performance gap across dimensions A significant capability imbalance exists across the eight dimensions. Most models achieve near-saturation in basic perception tasks (e.g., WI). However, performance diverges sharply in spatial reasoning tasks.

Correlation Between Fundamental P&R Capabilities and End-to-End Competence. As shown in Table 3, RI and SI—two tasks with varying difficulty levels that reflect end-to-end agent performance—show a high correlation with FPR-ACC. This indicates that FPR-ACC effectively reflects both the models' fundamental P&R capabilities and their advanced end-to-end performance.

4 GUI-XL-Intervention

Building on prior findings that representation alignment mitigates cross-lingual discrepancies (Nguyen et al., 2024b; Peng et al., 2025; Chang et al., 2022), we aim to leverage the superior P&R capabilities of English to bridge cross-

lingual P&R gaps. To this end, we identify the layers where the cross-lingual input representation distributions are most divergent, as exemplified in the left graph of Figure 4. We then propose **GUI Cross-Lingual Intervention (GUI-XLI)** to steer non-English representations toward their English counterparts (See Appendix B.2 for an overview).

4.1 Preliminaries

Given an LVLM parameterized by θ, visual and textual inputs are embedded and concatenated into an initial sequence $\boldsymbol{H}^{(0)}$. Processed through L Transformer (Vaswani et al., 2017) layers, the hidden state $h^{(l)}$ of the final token updates via:

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

where $a^{(l)}$ and $m^{(l)}$ denote attention and MLP outputs, respectively. Finally, $\boldsymbol{H}^{(L)}$ is projected by the language head for autoregressive prediction.

4.2 Cross-lingual Discrepancy Vector Construction

Given the autoregressive nature of LVLMs, the final token's hidden state aggregates the global context of the input sequence. Consequently, it serves as a natural anchor for cross-lingual alignment. Leveraging this property, we construct cross-lingual discrepancy vectors $\delta^{(l)}_{\textbf{en-tgt}}$ between parallel inputs $(X_{\textbf{en}}, I_{\textbf{en}})$ and $(X_{\textbf{tgt}}, I_{\textbf{tgt}})$ to faithfully capture the gaps in GUI P&R capabilities:

$$\delta^{(l)}_{\textbf{en-tgt}} = h^{(l)}(X_{\textbf{en}}, I_{\textbf{en}}) - h^{(l)}(X_{\textbf{tgt}}, I_{\textbf{tgt}}), \tag{2}$$

where the subscript **en** and **tgt** means English and the target language, respectively. The effectiveness of $\delta^{(l)}_{\textbf{en-tgt}}$ in isolating linguistically induced P&R discrepancies relies on three core design principles:

Explicit Reasoning Chains Text input X includes the question, a reasoning chain, and the final answer. This explicitly externalized reasoning chain ensures that $h^{(l)}(X, I)$ captures the underlying GUI P&R logic for choosing the final answer.

Positive–negative Sample Pairs Building on the observation that LVLMs use intermediate layers as an English-centric reasoning hub (Zhao et al., 2024), we select sample pairs where the model succeeds in English (X_{en}) but fails in the target language (X_{tgt}). This strategy contrasts the "successful English reasoning pathway" against the "failed target-language pathway" within these pivotal layers, maximizing the extraction of the capability gap while minimizing irrelevant linguistic noise.

Visual-Linguistic Isolation I_{en} and I_{tgt} represent identical GUI scenarios differing solely in language, which ensures that any visual variance in $\delta^{(l)}_{\text{en-tgt}}$ stems exclusively from visual-linguistic discrepancies, aligning with MPR-GUI-Bench.

4.3 GUI-XL-Memory

We construct the GUI Cross-Lingual Memory (GUI-XL-Memory) to store the cross-lingual discrepancy vectors, which enables LVLMs to adaptively retrieve and apply these vectors as optimization directions during inference, facilitating robust generalization to out-of-domain GUI scenarios.

Specifically, we extract the target language query and image representation as the retrieval key:

$$r^{(l)}_{\textbf{tgt}} = h^{(l)}(Q_{\textbf{tgt}}, I_{\textbf{tgt}}), \tag{3}$$

and the discrepancy vector as the value:

$$v^{(l)}_{\textbf{en-tgt}} = h^{(l)}(X_{\textbf{en}}, I_{\textbf{en}}) - h^{(l)}(X_{\textbf{tgt}}, I_{\textbf{tgt}}). \tag{4}$$

The $(r^{(l)}_{\textbf{tgt}}, v^{(l)}_{\textbf{en-tgt}})$ forms one entry of the memory.

4.4 Cross-lingual Representation Intervention

Given the current input $(Q^{\text{curr}}_{\text{tgt}}, I^{\text{curr}}_{\text{tgt}})$, we extract the hidden state $h^{(l,\text{curr})}_{\text{tgt}}$. We then identify the top-k semantically nearest entries in GUI-XL-Memory by maximizing cosine similarity:

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

The retrieved crossdiscrepancy vectors are averaged to form the intervention vector $\bar{v}^{(l)}_{\text{en-tgt}}$. During inference, we intervene on the residual stream by incorporating $\bar{v}^{(l)}_{\text{en-tgt}}$ with strength α, followed by magnitude-preserving normalization:

$$\tilde{h}^{(l,\text{curr})}_{\textbf{tgt}} = \frac{\|h^{(l,\text{curr})}_{\textbf{tgt}}\|_2 \cdot \left(h^{(l,\text{curr})}_{\textbf{tgt}} + \alpha\,\bar{v}^{(l)}_{\textbf{en-tgt}}\right)}{\|h^{(l,\text{curr})}_{\textbf{tgt}} + \alpha\,\bar{v}^{(l)}_{\textbf{en-tgt}}\|_2}, \tag{6}$$

which steers non-English representations toward English P&R patterns without modifying model parameters. Consequently, GUI-XLI is training-free and compatible with diverse LVLM architectures.

5 Experiment

5.1 Setup

Baseline Models We evaluate GUI-XLI's effectiveness on Intern2.5VL-8B (Chen et al., 2024b) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025).

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 4: Effectiveness of GUI-XLI across languages. Each cell reports accuracy (%) for the baseline (×) and our method (✓). Colored highlights denote absolute performance gains or drops. GXI denotes GUI-XLI.

Language Selection We select 4 languages (ZH, JA, RU, TH) as their non-Latin scripts and large linguistic distance from English pose the most challenging cross-lingual alignment conditions.

Memory Construction For basic P&R tasks, we build dimension-specific memories. For end-to-end reasoning tasks (RI and SI), we merge the entries from the six basic P&R dimensions so that the memory captures their combined capability. Details are provided in Appendix B.1.

5.2 Main Results

Table 4 presents the evaluation results on MPR-GUI-Bench before and after applying GUI-XLI. Our analysis yields three key conclusions:

(1) Effective Cross-Lingual Capability Transfer GUI-XLI significantly enhances fine-grained P&R capabilities in non-English settings, effectively aligning them with English-level proficiency. For high-resource languages like ZH, Intern2.5VL-8B and Qwen-2.5-VL-7B-Instruct achieve absolute FPR-ACC gains of 10.9% and 6.0%, respectively. Crucially, this improvement is consistent across lower-resource languages (e.g., TH, JA), bridging the average performance gap by 5.4% on average.

(2) Data Independence and Model Generalization We construct GUI-XL-Memory using data entirely disjoint from MPR-GUI-Bench, ensuring that performance gains stem from transferable P&R patterns. In addition, consistent improvements across heterogeneous open-source LVLMs demonstrate our GUI-XLI method's robustness and broad architectural applicability.

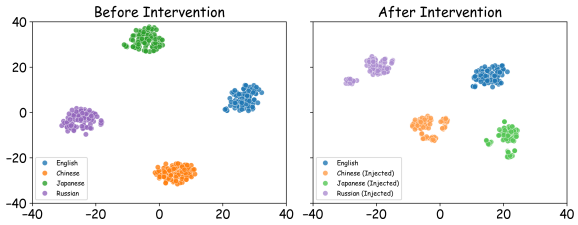


Figure 4: t-SNE Visualization of Multilingual Hidden State before and after applied GUI-XLI.

(3) Differential Gains across Task Dimensions Gains are most pronounced in action understanding and prediction tasks (AU, AP), with Intern2.5VL-8B and Qwen-2.5-VL-7B-Instruct achieving average FPR-ACC increases of 3.7% and 2.4%, respectively. Conversely, simpler dimensions (e.g., WI) exhibit marginal gains due to performance saturation. Improvements in spatial and end-to-end reasoning (REL, SI) are moderate; we attribute this to their weaker dependence on linguistic context—which limits the impact of cross-lingual alignment—and their inherent complexity, which remains a bottleneck beyond language barriers.

This result suggests that our approach is particularly effective in enhancing performance in tasks that require complex reasoning and prediction, but less so for tasks that are already well-performing or involve higher-order cognitive processing.

6 Analysis

6.1 Cross-lingual Alignment Visualization

As mentioned in Section 4.2, where intermediate layers are shown to serve as English-centric reasoning hubs, cross-lingual distributional differences in intermediate layers $h^{(l)}$ reflect discrepancies

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 5: Performance comparison under two settings: **Reasoning Answering**, where the model generates reasoning chains before prediction, and **Reasoning + GUI-XLI**, which further applies GUI-XLI. GXI denotes GUI-XLI.

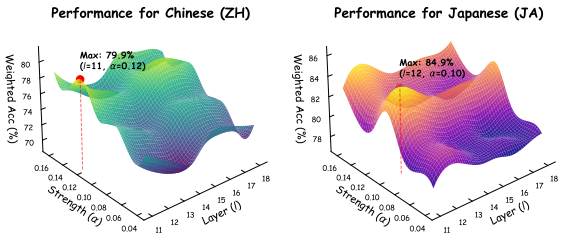


Figure 5: 3D performance landscape on MPR-GUI-Bench. Variations in weighted average accuracy are shown across intervention layers l and strengths α for Chinese and Japanese. Red dot markers denote the global optima, identifying the sweet spot for model performance in both language settings.

in GUI P&R capabilities across languages. To gain mechanistic insight into how GUI-XLI improves non-English P&R capabilities, we apply t-SNE to the final-token's hidden states at the intervention layer l for English inputs and their non-English counterparts (ZH, RU, JA), before and after applying GUI-XLI. As shown in Figure 4, without GUI-XLI, representations form distinct language-specific clusters, indicating substantial cross-lingual divergence. After applying GUI-XLI, non-English representations become more concentrated and aligned with their English counterparts, providing qualitative evidence that GUI-XLI effectively bridges cross-lingual GUI P&R gaps.

6.2 Ablation Studies

Figure 5 presents a systematic study to identify **P&R-dominant layers** and determine the appropriate intervention strength α. Motivated by our analysis in Section 6.1, which shows that intermediate layers function as an English-centric reasoning hub, we believe that cross-lingual P&R discrepancies are most salient in this layer range. We conduct a two-stage ablation. First, fixing $\alpha = 0.1$, we sweep the intervention layer l across intermediate layers (11–18) to identify the optimal layer l^*. Second, conditioned on l^*, we vary α to assess sensitivity to intervention strength. For each model and language setting, we select the configuration yielding the highest weighted accuracy.

6.3 Reasoning Enhancement of GUI-XLI

Table 5 shows that beyond improving non-English P&R performance, GUI-XLI also yields substantial gains when required to generate reasoning chains before answering. Notably, GUI-XLI consistently improves performance even under explicit reasoning supervision, indicating that the gains arise from a genuine enhancement of underlying P&R capability rather than prompting effects. From a mechanistic perspective, reasoning chains externalize internal P&R trajectories; the improved performance therefore suggests that GUI-XLI effectively reshapes non-English P&R patterns toward more coherent and task-relevant reasoning.

7 Conclusion

In this paper, we introduce MPR-GUI-Bench, the first benchmark designed to evaluate the fundamental fine-grained perception and reasoning (P&R) capabilities in GUI agents across strictly aligned cross-lingual environments spanning six languages. Our evaluation of current LVLMs reveals a consistent performance bottleneck in non-English settings, particularly in reasoning-intensive tasks. To leverage the superior P&R capabilities of English, we identify critical layers that are most sensitive to linguistic discrepancies. Based on these insights, we propose GUI-XLI, an inference-time intervention method that aligns non-English representations with superior English counterparts at these critical layers. Experimental results demonstrate that GUI-XLI significantly bridges the cross-lingual gap with an absolute average performance gain of 6.5% in non-English settings.

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, although our MPR-GUI-Bench is effective for benchmarking their fine-grained P&R capabilities.

References

Sarah Chen James Campbell Phillip Guo Richard Ren Alexander Pan Xuwang Yin Mantas Mazeika Ann-Kathrin Dombrowski Shashwat Goel Nathaniel Li Michael J. Byun Zifan Wang Alex Mallen Steven Basart Sanmi Koyejo Dawn Song Matt Fredrikson Zico Kolter Dan Hendrycks Andy Zou, Long Phan. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

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.

Domenic V Cicchetti and Alvan R Feinstein. 1990. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

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.

Gemini Team, Tom Brown, Jan Leike, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2024. Showui: One vision-language-action model for gui visual agent. *Preprint*, arXiv:2411.17465.

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.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

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.

Bin Xie, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Jie Liu, Min Zhang, and Liqiang Nie. 2025. GUI-explorer: Autonomous exploration and mining of transition-aware knowledge for GUI agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5650–5667, Vienna, Austria. Association for Computational Linguistics.

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

Pei Yang, Hai Ci, and Mike Zheng Shou. 2025b. macosworld: A multilingual interactive benchmark for gui agents. *Preprint*, arXiv:2506.04135.

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*.

Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, Jie Xie, Wei Zhou, Wang Xu, Yuanheng Zhang, Zhou Su, Zhongwu Zhai, Xiaoming Liu, Yudong Mei, Jianming Xu, and 6 others. 2025a. AgentCPM-GUI: Building mobile-use agents with reinforcement fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 155–180, Suzhou, China. Association for Computational Linguistics.

Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, Jie Xie, Wei Zhou, Wang Xu, Yuanheng Zhang, Zhou Su, Zhongwu Zhai, Xiaoming Liu, Yudong Mei, Jianming Xu, and 6 others. 2025b. AgentCPM-GUI: Building mobile-use agents with reinforcement fine-tuning. *arXiv preprint arXiv:2506.01391*.

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 Annotator Background

The annotation process was conducted by one of the authors and five trained volunteers, all of whom hold at least a bachelor's degree and possess expertise in GUI applications. To ensure high-quality multilingual dataset construction, we implemented a **Lead-Supporting Annotator** framework. For each language, we ensured that at least two **Lead Annotators** were assigned based on their certified proficiency (e.g., TOEFL, HSK, or JLPT), with their specific roles detailed in Table 6.

To prioritize annotator well-being and maintain high data quality, the **workload was strictly controlled** at 1–2 hours per day over a period of two weeks. Annotators were compensated at an average rate of 15 USD per hour, following a comprehensive training phase that included task-specific tutorials and alignment on quality criteria.

Privacy and Content Safety. During data collection, annotators were explicitly instructed to avoid collecting any GUI content that contains personal or sensitive information. In particular, screenshots and annotations that name or uniquely identify individual people (e.g., real names, user IDs, email addresses, phone numbers, or account-related information) were excluded. After collection, all samples were manually reviewed, and any content containing personally identifiable information or potentially offensive material was filtered out. As a result, the final dataset does not include personal identifiers or offensive content.

A.2 Data Collection Guidelines

As shown in Table 7, we provide guidelines for annotators on data collection and verification to ensure data quality and consistency across annotators.

A.3 Prompts For Candidate VQA Lists Construction

In this section, we list all prompts used during the process of constructing **MPR-GUI-Bench**, which include VQA generation for eight dimensions (Table 8 - Table 14). 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.4 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 15.

A.5 Comprehensive Inter-Rater Reliability Analysis for All Languages

We report the inter-rater reliability analysis for all six languages included in MPR-GUI-Bench (EN, ZH, FR, RU, JA, TH). In the English setting, annotators verified the compliance of generated questions with task requirements. For the remaining languages, the evaluation focused on the linguistic faithfulness of translated text relative to the visual content in the respective screenshots. This analysis covers all 2,156 samples per language, with six annotators classifying each as "Compliant" or "Non-compliant."

The annotators who have proficiency in respective languages provided direct judgments on the faithfulness of GPT-4o translations relative to the visual content. Other annotators served as **Supporting Annotators**, utilizing auxiliary tools (e.g., DeepL and Google Translate) to translate model outputs back into English and cross-reference them with the original visual prompts for auxiliary semantic verification.

Distribution of Rater Agreement. Table 16 presents the distribution of rater agreement across all languages. Across the entire benchmark, a high level of consensus was observed, with the majority of samples achieving perfect (6:0) or near-perfect (5:1) agreement. Notably, the high proportion of "6 vs. 0" cases confirms the effectiveness of our Lead-Supporting annotator framework and the clarity of our annotation guidelines.

Comparative Reliability Metrics. As shown in Table 17, the Fleiss' Kappa coefficients across all languages fall within a relatively low range (0.15–0.22). This is a well-documented phenomenon known as the *prevalence paradox* (Cicchetti and Feinstein, 1990), which occurs when the distribution of categories is highly skewed. In our case, since over 94% of samples are classified as "Com-

Annotator	EN	ZH	FR	RU	JA	TH
1	Lead	Lead	Lead	–	–	–
2	Lead	Lead	–	Lead	–	–
3	Lead	–	Lead	–	Lead	–
4	Lead	–	–	–	–	Lead
5	Lead	Lead	–	–	Lead	–
6	Lead	–	Lead	–	–	Lead
Role	Expert	Expert/Tool	Expert/Tool	Expert/Tool	Expert/Tool	Expert/Tool

Table 6: Annotator roles and language distribution. "Lead" indicates annotators with certified proficiency who provided direct linguistic judgments. Supporting annotators (denoted by tool-assisted roles) utilized back-translation for verification.

pliant," the probability of agreement by chance ($\bar{P}_e$) is naturally very high. Consequently, even with a high observed agreement ($\bar{P} > 0.90$), the Kappa coefficient remains low because it only measures the marginal improvement over an already high baseline of chance.

To provide a more robust assessment that accounts for this imbalance, we also report **Gwet's AC1** coefficient, which is mathematically less sensitive to the prevalence problem. As shown in the table, our AC1 values consistently exceed 0.85 across all languages, indicating "almost perfect" agreement according to standard benchmarks and validating the reliability of our annotation process.

Adjudication and Quality Control. For each language, all samples that failed to reach a 6:0 consensus were subjected to a final **Adjudication Phase**. The respective Lead Annotators (native or proficient speakers) resolved all disagreements to ensure the final ground-truth labels were accurate.

This rigorous multi-stage verification process guarantees that the linguistic and visual alignment of MPR-GUI-Bench remains reliable despite the inherent difficulty of multilingual GUI evaluation.

A.6 Quality Assurance and Refinement to Eliminate Model-specific Semantic Style

As mentioned in Appendix A.3, to safeguard the professional rigor and stylistic neutrality of MPR-GUI-Bench, we developed dimension-specific constrained prompts. This framework decomposes complex GUI interactions into fundamental sub-abilities—such as requiring multi-cue synthesis for Widget Function perception and strict sequence validation for Action Prediction. By incorporating Negative Linguistic Constraints to prohibit meta-comments and conversational fillers, we minimize the stylistic artifacts of the generator, ensuring the benchmark targets genuine visual reasoning over linguistic shortcuts.

Furthermore, we implemented rigorous annotation guidelines to refine all GPT-4o-generated samples as shown in Table 18.

To verify that MPR-GUI-Bench is generator-agnostic, we conducted a cross-model evaluation using task sets independently generated by GPT-4o and Gemini-2.5-Pro. The results, summarized in Table 19, provide strong evidence that our evaluation pipeline is resilient to model-specific bias.

Absence of Generator Dominance Empirical data show that the generator backbone does not inherently dominate its own test set. On the GPT-4o-generated subset, Qwen-2.5-VL-7B-Instruct achieves the highest scores in AEL (78%) and WF (86%), surpassing the generator (GPT-4o). Conversely, on the Gemini-2.5-Pro-generated subset, GPT-4o maintains a performance lead in 5 out of 8 tasks (e.g., AU, WI, WF, SI), despite the tasks being curated by Gemini. This lack of "home-field advantage" confirms that the diagnostic data represents objective P&R challenges rather than stylistic artifacts.

High Ranking Consistency Despite variations in absolute scores—likely due to differing task difficulty distributions—the relative performance hierarchy remains remarkably stable. We observe a 100% Top-1 ranking consistency in key perception and reasoning dimensions, including AEL, REL, AU, and WI. For instance, GPT-4o consistently outranks Gemini-2.5-Pro in WI across both generation sources (94% vs. 79% and 92% vs. 89%). The addition of RI and SI tasks further reinforces

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 7: Data Collecting Guidelines

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 8: 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 $\rightarrow$ step2 $\rightarrow$ ... }
B. {... }
C. {... }
D. {... }
answer: {Correct option letter}
type: Action Prediction

Table 9: 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 questions in this exact format. Only include questions that have definite answers.

Table 10: 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 questions in this exact format. Only include questions that have definite answers.

Table 11: 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 12: 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 13: 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 14: Prompt for RI & SI Dimensions

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

Table 15: 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.

Agmt.	EN	ZH	FR	RU	JA	TH
6 vs. 0	1693	1650	1621	1634	1602	1588
5 vs. 1	291	312	340	325	355	360
4 vs. 2	110	120	135	128	140	145
3 vs. 3	42	54	45	50	48	52
2 vs. 4	16	15	12	14	10	9
1 vs. 5	1	3	2	3	1	2
0 vs. 6	3	2	1	2	0	0
Total	2156	2156	2156	2156	2156	2156

Table 16: Summary of Rater Agreement Distribution for all six languages. **Agmt.** denotes the Agreement level; numbers represent the count of items for each agreement configuration (e.g., "6 vs. 0" denotes total consensus).

this stability, with Gemini and GPT-4o maintaining shared dominance in RI (96% and 92% respectively) regardless of the data source.

Capability-Driven vs. Style-Driven Results The robust performance of Qwen-2.5-VL-7B-Instruct, serving as a non-generator "third-party" model, provides additional validation. Qwen consistently achieves top-tier results in AEL (ranking 1st in both subsets with 78% and 81%), proving that the benchmark measures standardized GUI interaction skills that transcend specific LLM prompting styles.

Quantitative Verification Quantitatively, the Metric-wise Ranking Concordance remains high (averaging $\tau \approx 0.70$ across tasks, with perfect $\tau = 1.0$ in AEL). Given the narrow performance margins between these state-of-the-art models (averaging $< 2\%$), this degree of concordance is statistically significant. It indicates that the performance hierarchy is driven by the intrinsic P&R capabilities of the evaluated models rather than stylistic alignment with the generator, effectively neutralizing potential self-preference bias.

Lang.	P_o	P_e	Fleiss' κ	Gwet's AC1
EN	0.9120	0.8942	0.1682	0.8950
ZH	0.9055	0.8850	0.1783	0.8892
FR	0.8994	0.8790	0.1686	0.8810
RU	0.9021	0.8812	0.1752	0.8845
JA	0.8950	0.8710	0.1860	0.8780
TH	0.8920	0.8680	0.1818	0.8755

Table 17: Summary of statistical reliability metrics for all languages. **Lang.** is Language; P_o is Observed Agreement; P_e is Expected (Chance) Agreement. The high AC1 values across all settings validate the robustness of the MPR-GUI-Bench data collection process despite the prevalence paradox impacting the κ values.

A.7 Details about FPR-ACC

We use the FPR-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.8 Case Study

To gain deeper insights into the limitations of current LVLMs, we conduct a qualitative analysis of typical failure modes across the various dimensions of **MPR-GUI-Bench**. Figures 8 through 13 illustrate representative incorrect responses, highlighting the specific challenges they encounter in GUI perception and reasoning.

A.9 Evaluations results on additional models with broader size

As shown in Table 20, to verify the diagnostic power of MPR-GUI-Bench, we extend our evaluation to a broader spectrum of models, including an LVLM for GUI agent: CogAgent-9B (Hong et al., 2023) and two additional open-source baselines: MiniCPM-o 2.6 (Yao et al., 2024) and Keye-VL-7B (Kwai Keye Team, 2025a,b).

Quality Assurance and Refinement Guidelines

1. Objective.
Ensure that each VQA sample is factually accurate, strictly grounded in visual evidence, and presents a *human-solvable yet non-trivial* reasoning challenge without inheriting model-specific semantic styles. If two independent annotators disagree on the answer after inspection, the sample must be flagged and revised or removed. *Non-trivial* indicates that the question cannot be answered via a single salient cue and requires integration of at least two visual or semantic cues (e.g., icon + text, color + position).

2. Question Evaluation Criteria.
(i) *Visual Existence & Strict Grounding*: all referenced UI elements must be clearly identifiable in the screenshot;
(ii) *Factual Integrity*: zero incorrect assumptions about GUI state (e.g., disabled vs. enabled);
(iii) *Moderate Complexity*: avoid trivial perception-only questions; require contextual perception and reasoning;
(iv) *Human Answerability & Legibility*: discard samples with blurred, truncated elements or those requiring insider app knowledge.

3. Answer and Distractor Standards.
The correct answer must be the *only* logically sound option. Distractors must be grounded in the same interface and reflect plausible misinterpretations (e.g., similar icons in different regions), while remaining mutually exclusive. Distractors must not be partially or conditionally correct and should introduce subtle logical or spatial errors that penalize superficial pattern matching.

4. Stylistic and Terminology Refinement.
Remove AI-typical reasoning traces and conversational fillers to prevent exploitation of linguistic priors. Enforce a concise, imperative tone consistent with real GUI interactions. Standardize platform-specific terminology (e.g., "Tap" for mobile) and ensure localization fidelity by matching the exact UI lexicon shown on the screen for non-English samples.

5. Failure Mode Filtering (Mandatory Veto).
Discard or revise samples exhibiting:
(i) spatial inconsistency (incorrect relative positioning);
(ii) interactional logic violations in action prediction tasks;
(iii) shortcut cues such as systematic length, tone, or punctuation bias in the correct option;
(iv) evidence insufficiency, where answers rely on brand- or app-specific prior knowledge rather than explicit visual evidence.

Table 18: Quality assurance and refinement guidelines used during data collection and annotation.

Evaluated Model	AEL	REL	AU	AP	WI	WF	RI	SI
Generated by GPT-4o								
Gemini-2.5-Pro	58	**92**	81	**91**	79	63	**96**	**80**
GPT-4o	72	83	**92**	**91**	**94**	84	**96**	76
Qwen-2.5-VL-7B-Instruct	**78**	90	84	**91**	91	**86**	**96**	72
Generated by Gemini-2.5-pro								
Gemini-2.5-Pro	72	**67**	75	**90**	89	79	**92**	**82**
GPT-4o	77	60	**80**	88	**92**	**86**	**92**	**82**
Qwen-2.5-VL-7B-Instruct	**81**	55	70	89	91	83	88	72

Table 19: Cross-verification of model bias using QAs generated by different backbones.

Model	Lang	Perception				Reasoning				FPR-ACC
		AU	AP	WF	WI	AEL	REL	RI	SI	
Open-source LVLMs										
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
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
Multimodal 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

Table 20: Performance results (%) for the **three additional baseline models** across all six languages. Within each dimension, the highest and lowest accuracies are highlighted in green and orange, respectively, to demonstrate the performance variance across different model scales.

B Details on GUI-XLI

B.1 Memory Construction Details

To ensure maximum inference efficiency and real-time responsiveness, we maintain an extremely compact memory for the foundational P&R capabilities, where each dimension (AU, AP, WF, WI, AEL, and REL) consists of only 10 entries.

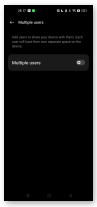
For the reasoning-intensive dimensions (RI and SI), relying solely on their own training samples might be insufficient as they inherently depend on foundational perception capabilities. Therefore, we construct the memory for RI and SI by aggregating entries from the six foundational dimensions: AU, AP, WF, WI, AEL, and REL. This aggregation is performed independently at each layer to maintain the layer-wise alignment structure defined in Section 4.3. This strategy ensures that the intervention for reasoning tasks is grounded in robust fine-grained perception cues.

B.2 Overview of GUI-XLI

In this subsection, we present the overview illustration figure of our GUI-XLI method in Figure 14.

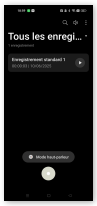
C Usage of LLMs

GPT-4o (OpenAI et al., 2024) was employed to assist in refining the language and enhancing the readability of the manuscript. All ideas, experiments, analyses, and conclusions were conceived, conducted, and verified by the authors.

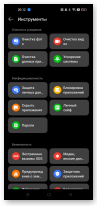


Question: *What immediate change will occur in the interface state if the user toggles the "Multiple users" option to its active state?*
Options:
A. The interface will display a list of currently set users along with options to add or delete users.
B. A warning message will appear indicating potential 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* ❌

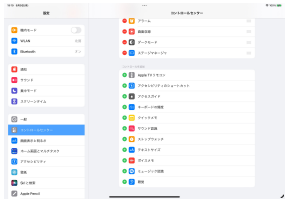
Question: *在当前界面状态下，关闭"私密保险箱"功能的直接效果是什么？*
Options:
A. "私密保险箱"内容将无法访问，除非输入密码。
B. 将出现确认对话框以验证用户禁用该功能的操作。
C. 系统将在更改生效前显示加载指示器。
D. 切换按钮将短暂闪烁以指示状态更新成功。
Answer: *A* ✅
Type: *Action Understanding*
Predict: *B* ❌



Question: *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* ❌

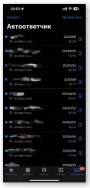
Question: *หากผู้ใช้แตะที่ไอคอน "ดาวน์โหลด" ผลลัพธ์การดำเนินการที่เป็นไปได้มากที่สุดคืออะไร?*
Options:
A. ระบบแสดงรายการไฟล์ที่ดาวน์โหลดทั้งหมดจัดเรียงตามประเภทไฟล์ B. อินเทอร์เฟซเปลี่ยนไปยังหน้าจอที่แสดงเฉพาะตัวเลือกการดาวน์โหลดจากอินเทอร์เน็ต
C. ระบบแสดงข้อความแสดงข้อผิดพลาดว่า "ไม่มีการดาวน์โหลด"
D. อินเทอร์เฟซเปิดไปยังหน้าการตั้งค่าเพื่อกำหนดค่าการดาวน์โหลด
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： *A* ✅
Type： *Action Prediction*
Predict： *C* ❌

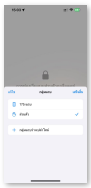
Question: *用户想切换到"视频"模式来录制视频。为实现此目标，以下哪项描述是正确的?*
Options:
A. 在"照片"模式上向左滑动以切换到"视频"模式。
B. 点击"视频"标签以切换到"视频"模式。
C. 长按快门按钮以切换到"视频"模式。
D. 在"照片"模式上向右滑动以切换到"视频"模式。
Answer: *A* ✅
Type: *Action Prediction*
Predict: *C* ❌

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: *C* ✅
Type: *Action Prediction*
Predict: *A* ❌



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: *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* ❌



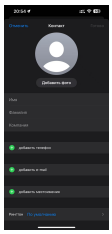
Question: *ผู้ใช้ต้องการสร้างกลุ่มเก็บว่างเปล่าใหม่ เพื่อบรรลุเป้าหมายนี้ คำอธิบายใดต่อไปนี้เป็นจริง?*
Options:
A. แตะ "กลุ่มเก็บว่างเปล่าใหม่" → แตะ "เสร็จสิ้น".
B. แตะ "แก้ไข" → แตะ "กลุ่มเก็บว่างเปล่าใหม่".
C. กดค้างที่ "ส่วนตัว" → แตะ "กลุ่มเก็บว่างเปล่าใหม่".
D. ปัดไปทางซ้ายบน "175 แอป" → แตะ "กลุ่มเก็บว่างเปล่าใหม่".
Answer: *A* ✅
Type: *Action Prediction*
Predict: *B* ❌

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

Question： *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： *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： *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： *C* ✅
Type： *Absolute Element Location*
Predict： *A* ❌

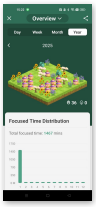


Question： *Пожалуйста, определите положение кнопки "Отменить" на экране.*
Options：
A. Верхний левый квадрант, строка состояния
B. Верхний левый квадрант, панель инструментов
C. Нижний левый квадрант, основная область контента
D. Средняя часть, панель инструментов
Answer： *B* ✅
Type： *Absolute Element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

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： *โปรดระบุตำแหน่งของสวิตช์สลับสำหรับ "แสดงอุปกรณ์บลูทูธที่ไม่มีชื่อ" ในระบบพิกัดทั่วโลก*
Options：
A. ไตรมาสบนซ้าย
B. ไตรมาสบนขวา
C. ไตรมาสล่างซ้าย
D. ไตรมาสล่างขวา
Answer： ***D*** ✅
Type： *Absolute Element Location*
Predict： ***B*** ❌

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



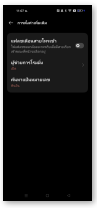
Question: *Which element is located at the top-left corner of the screenshot?*
Options:
A. Total focused time: 1467 mins
B. Year selection dropdown
C. Overview
D. Back arrow
Answer: *D* ✅
Type: *Relative Element Location*
Predict: *C* ❌

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* ❌

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* ❌

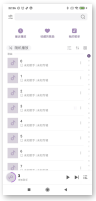
Question: *「録音の上限」オプションのすぐ上にある設定はどれですか？*
Options:
A. 指定した番号を録音
B. 不明な番号を記録
C. すべての通話を録音
D. 通話録音を表示
Answer: *C* ✅
Type: *Element Location*
Predict: *A* ❌



Question: *ตัวเลือกการตั้งค่าใดที่อยู่เหนือการตั้งค่า "ค้นหาแป้นหมายเลข" โดยตรง?*
Options:
A. ผู้ช่วยการโรมมิ่ง
B. แฟลชเตือนสายโทรเข้า
C. การตั้งค่าเพิ่มเติม
D. สวิตช์เปิดปิดสำหรับ "แฟลชเตือนสายโทรเข้า"
Answer: *A* ✅
Type: *Relative Element Location*
Predict: *B* ❌

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

Question: *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: *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* ❌

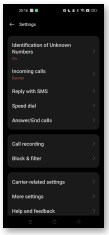
Question: *En se basant sur les indices visuels présentés dans la capture d'écran, laquelle des fonctions suivantes peut être directement activée à partir des widgets visibles sans aucune navigation supplémentaire ?*
Options:
A. Régler une nouvelle alarme dans le widget "Horloge"
B. Changer la fréquence 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: *A* ✅
Type: *Widget Function*
Predict: *B* ❌

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* ❌

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: *C* ✅
Type: *Widget Function*
Predict: *A* ❌

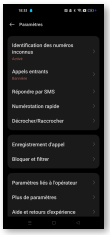
Question: *จากภาพหน้าจอ การกระทำใดที่ต้องเลือกก่อนจึงจะเปิดใช้งานได้*
Options:
A. แบ่งปัน
B. เปลี่ยนชื่อ
C. ตั้งเป็น
D. ลบ
Answer: *D* ✅
Type: *Widget Function*
Predict: *C* ❌

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



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* ❌

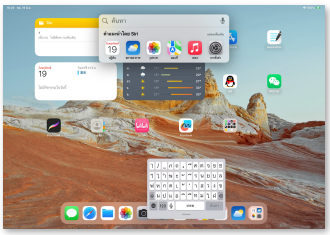
Question: *用户点击此移动界面截图中的"取消"按钮时会发生什么？*
Options:
A. 应用程序将继续收集数据并与云同步。
B. 退回到上一界面。
C. 用户将被重定向到详细说明隐私政策的帮助页面。
D. 将出现错误消息，指出该操作不允许。
Answer: *B* ✅
Type: *Widget Interaction*
Predict: *D* ❌



Question: *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* ❌

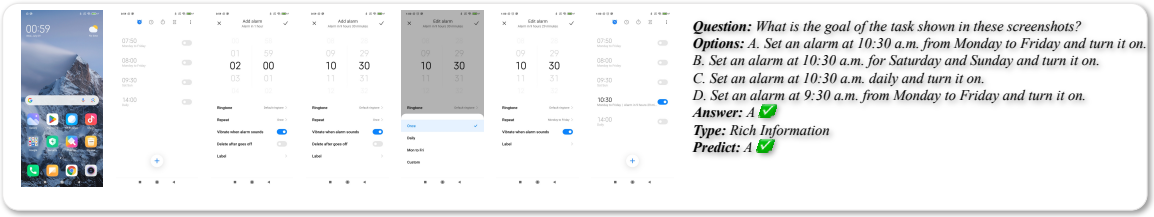
Question: *В музыкальном виджете, какое действие вы выполните, чтобы начать воспроизведение музыки с начала?*
Options:
A. Нажмите кнопку воспроизведения
B. Нажмите кнопку повтора
C. Удерживайте значок музыки
D. Дважды нажмите кнопку пропуска назад
Answer: *D* ✅
Type: *Widget Interaction*
Predict: *A*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

Question: *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* ❌



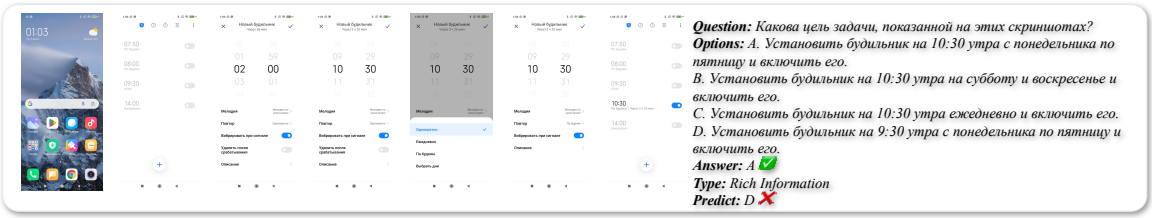
Question: *ผลลัพธ์ที่คาดหวังจากการแตะไอคอน "สภาพอากาศ" ภายใต้คำแนะนำโดย Siri คืออะไร?*
Options:
A. เปิดการพยากรณ์อากาศโดยละเอียด อัปเดตตำแหน่งตามข้อมูล GPS ปัจจุบัน
B. ส่งการแจ้งเตือนสภาพอากาศให้ผู้ใช้สำหรับสัปดาห์หน้า
C. แสดงหน้าต่างซ้อนทับพร้อมการปรับอุณหภูมิสำหรับสถานที่ต่างๆ
D. เพิ่มวิดเจ็ตสภาพอากาศไปยังหน้าจอหลักเพื่อการเข้าถึงอย่างรวดเร็ว
Answer: *A* ✅
Type: *Widget Interaction*
Predict: *C* ❌

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

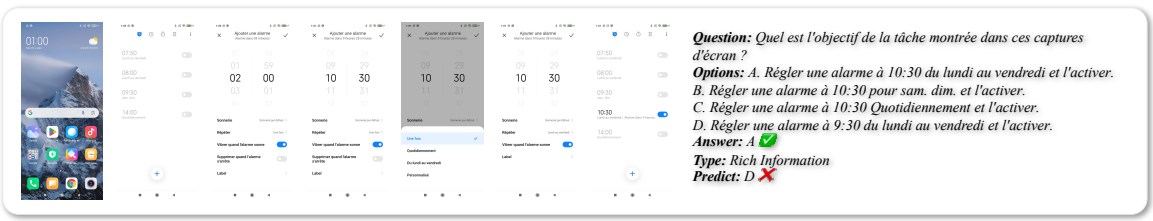


Question: *What is the goal of the task shown in these screenshots?*
Options: *A. Set an alarm at 10:30 a.m. from Monday to Friday and turn it on.*
B. Set an alarm at 10:30 a.m. for Saturday and Sunday and turn it on.
C. Set an alarm at 10:30 a.m. daily and turn it on.
D. Set an alarm at 9:30 a.m. from Monday to Friday and turn it on.
Answer: *A* ✅
Type: *Rich Information*
Predict: *A* ✅

Question: *这些截图中任务的目标是什么？*
Options: *A. 设置一个10:30的闹钟，周一至周五，并开启。*
B. 设置一个10:30的闹钟，周六周日，并开启。
C. 设置一个10:30的闹钟，每天，并开启。
D. 设置一个9:30的闹钟，周一至周五，并开启。
Answer: *A* ✅
Type: *Rich Information*
Predict: *D* ❌

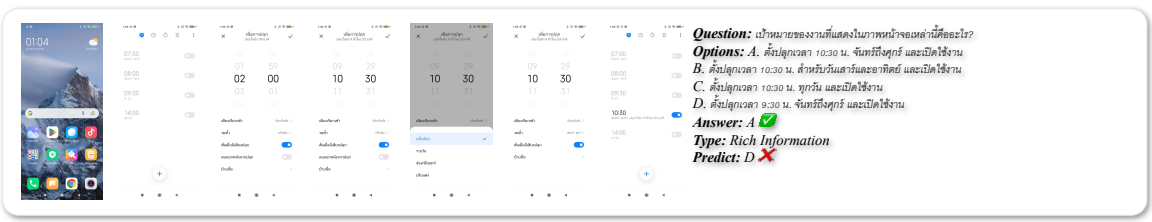


Question: *Какова цель задачи, показанной на этих скриншотах?*
Options: *A. Установить будильник на 10:30 утра с понедельника по пятницу и включить его.*
B. Установить будильник на 10:30 утра на субботу и воскресенье и включить его.
C. Установить будильник на 10:30 утра ежедневно и включить его.
D. Установить будильник на 9:30 утра с понедельника по пятницу и включить его.
Answer: *A* ✅
Type: *Rich Information*
Predict: *D* ❌



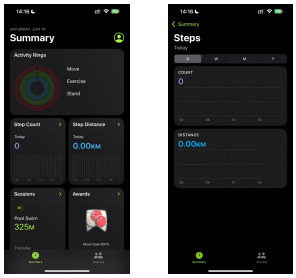
Question: *Quel est l'objectif de la tâche montrée dans ces captures d'écran ?*
Options: *A. Régler une alarme à 10:30 du lundi au vendredi et l'activer.*
B. Régler une alarme à 10:30 pour sam. dim. et l'activer.
C. Régler une alarme à 10:30 Quotidiennement et l'activer.
D. Régler une alarme à 9:30 du lundi au vendredi et l'activer.
Answer: *A* ✅
Type: *Rich Information*
Predict: *D* ❌

Question: *これらのスクリーンショットに示されているタスクの目的は何ですか？*
Options: *A. 月曜日から金曜日の午前10:30にアラームを設定してオンにする。*
B. 土曜日と日曜日の午前10:30にアラームを設定してオンにする。
C. 毎日午前10:30にアラームを設定してオンにする。
D. 月曜日から金曜日の午前9:30にアラームを設定してオンにする。
Answer: *A* ✅
Type: *Rich Information*
Predict: *D* ❌

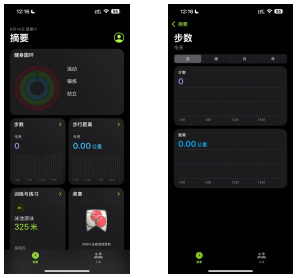


Question: *เป้าหมายของงานที่แสดงในภาพหน้าจอเหล่านี้คืออะไร?*
Options: *A. ตั้งปลุกเวลา 10:30 น. จันทร์ถึงศุกร์ และเปิดใช้งาน*
B. ตั้งปลุกเวลา 10:30 น. สำหรับวันเสาร์และอาทิตย์ และเปิดใช้งาน
C. ตั้งปลุกเวลา 10:30 น. ทุกวัน และเปิดใช้งาน
D. ตั้งปลุกเวลา 9:30 น. จันทร์ถึงศุกร์ และเปิดใช้งาน
Answer: *A* ✅
Type: *Rich Information*
Predict: *D* ❌

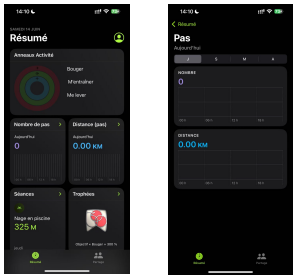
Figure 12: Examples of responses by LVLMs in RI dimension across 6 language settings.



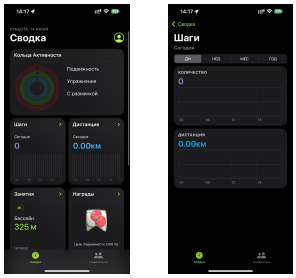
Question: *What is the goal of the task shown in these screenshots?*
Options: *A. Open the sport app.*
B. Check today's step count in the health app.
C. View the details of your latest swimming session.
D. Review your activity rings to assess your exercise goals.
Answer: *B* ✅
Type: *Sparse Information*
Predict: *D* ❌



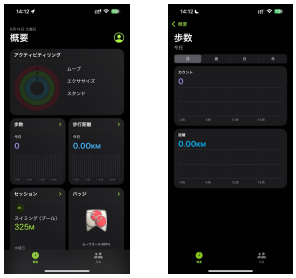
Question: *这些截图中显示的任务目标是什么？*
Options: *A. 打开运动应用。*
B. 查看健康应用中的今日步数。
C. 查看您最近一次游泳的详细信息。
D. 查看您的健身圆环以评估您的锻炼目标。
Answer: *B* ✅
Type: *Sparse Information*
Predict: *B* ✅



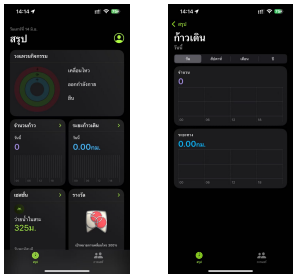
Question: *Quel est l'objectif de la tâche montrée dans ces captures d'écran ?*
Options: *A. Ouvrir l'application de sport.*
B. Vérifier le Nombre de pas d'aujourd'hui dans l'application de santé.
C. Voir les détails de votre dernière séance de Nage en piscine.
D. Examiner vos Anneaux Activité pour évaluer vos objectifs d'exercice.
Answer: *B* ✅
Type: *Sparse Information*
Predict: *B* ✅



Question: *Какова цель задачи, показанной на этих скриншотах?*
Options: *A. Открыть спортивное приложение.*
B. Проверить количество шагов за сегодня в приложении здоровья.
C. Просмотреть детали вашей последней тренировки по плаванию.
D. Просмотреть Кольца Активности, чтобы оценить ваши цели по упражнениям.
Answer: *B* ✅
Type: *Sparse Information*
Predict: *B* ✅



按照您要求的格式，为您转换该日语 Case 如下:
Question: *これらのスクリーンショットに示されているタスクの目標は何ですか？*
Options: *A. スポーツアプリを開く。*
B. ヘルスアプリで今日の歩数を確認する。
C. 最新のスイミングセッションの詳細を表示する。
D. アクティビティリングを確認してエクササイズ目標を評価する。
Answer: *B* ✅
Type: *Sparse Information*
Predict: *B* ✅



Question: *เป้าหมายของงานที่แสดงในภาพหน้าจอเหล่านี้คืออะไร?*
Options: *A. เปิดแอปกีฬา*
B. ตรวจสอบจำนวนก้าววันนี้ในแอปสุขภาพ
C. ดูรายละเอียดของเซสชั่นว่ายน้ำล่าสุดของคุณ
D. ตรวจสอบวงแหวนกิจกรรมของคุณเพื่อประเมินเป้าหมายการออกกำลังกาย
Answer: *B* ✅
Type: *Sparse Information*
Predict: *D* ❌

Figure 13: Examples of responses by LVLMs in SI dimension across 6 language settings.

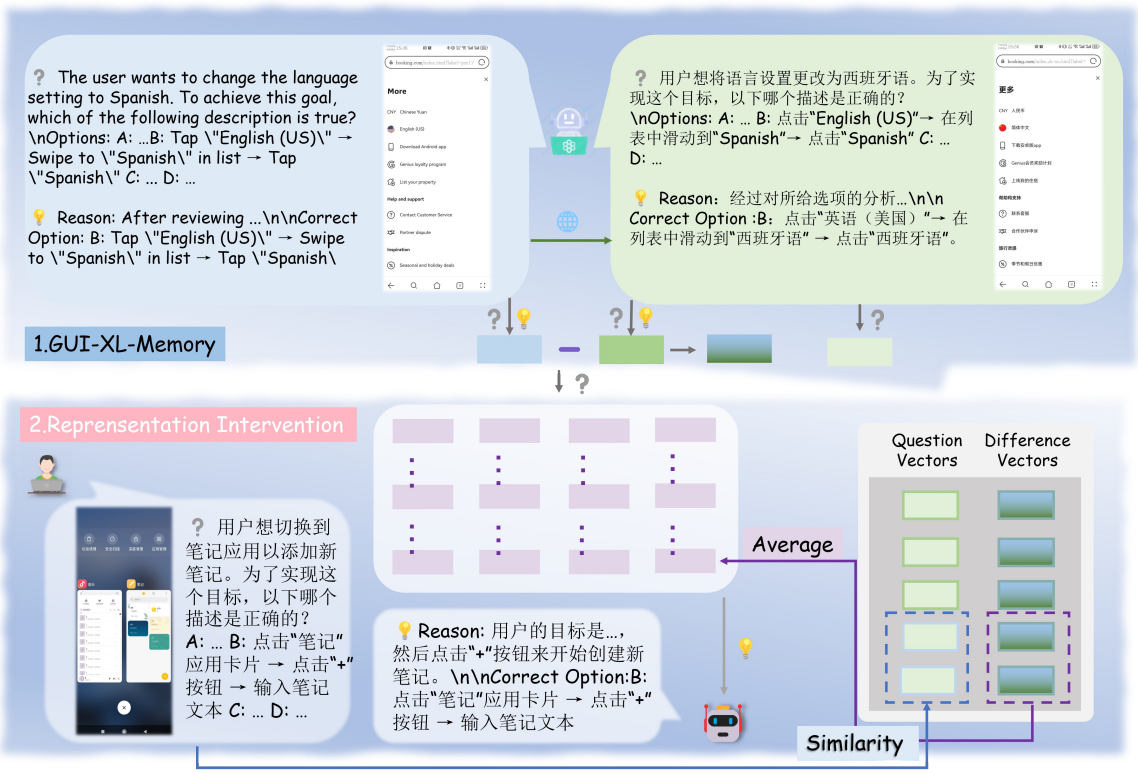


Figure 14: An Overview of our **GUI-XLI** method. **Step 1 GUI-XL-Memory**: We sample semantically parallel VQA pairs to form entries in **GUI-XL-Memory**. GUI-XLI uses semantically parallel entries from Step 1 as reference anchors to steer non-English states in Step 2. **Step 2 Cross-lingual Representation Intervention**: When answering non-English questions, related entries are retrieved to calculate cross-lingual discrepancy vectors and then injected to certain layer as intervention to add P&R capabilities to non-English settings.