GUI KNOWLEDGE BENCH: REVEALING THE KNOWLEDGE GAP BEHIND VLM FAILURES IN GUI TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision–language models (VLMs) have advanced graphical user interface (GUI) task automation but still lag behind humans. We hypothesize this gap stems from missing core GUI knowledge, which existing training schemes (such as supervised fine-tuning and reinforcement learning) alone cannot fully address. By analyzing common failure patterns in GUI task execution, we distill GUI knowledge into three dimensions: (1) interface perception, knowledge about recognizing widgets and system states; (2) interaction prediction, knowledge about reasoning action–state transitions; and (3) instruction understanding, knowledge about planning, verifying, and assessing task completion progress. We further introduce GUI Knowledge Bench, a benchmark with multiple choice and yes/no questions across six platforms (Web, Android, MacOS, Windows, Linux, iOS) and 292 applications. Our evaluation shows that current VLMs identify widget functions but struggle with perceiving system states, predicting actions, and verifying task completion. Experiments on real world GUI tasks further validate the close link between GUI knowledge and task success. By providing a structured framework for assessing GUI knowledge, our work supports the selection of VLMs with greater potential prior to downstream training and provides insights for building more capable GUI agents.

1 INTRODUCTION

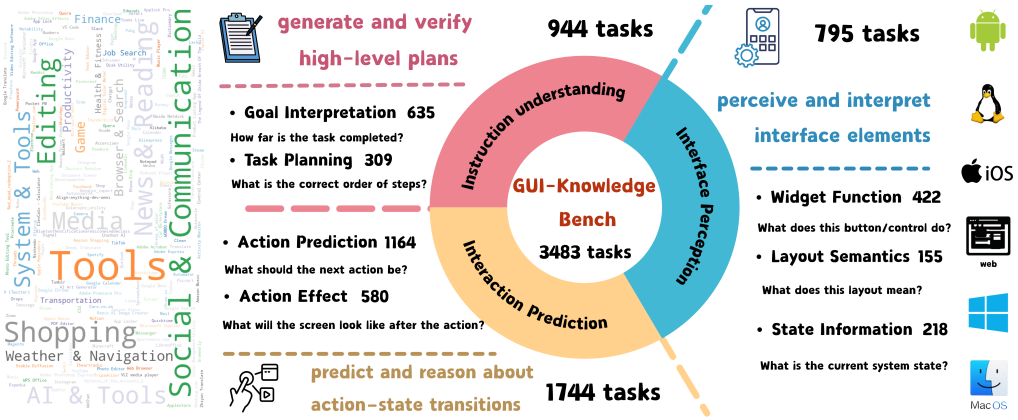

Figure 1: GUI Knowledge Bench: A benchmark evaluating VLMs on GUI knowledge across six platforms (Web, Android, MacOS, Windows, Linux, iOS). It measures three types of knowledge: Interface Perception, which evaluates understanding of GUI components, layout, and system state; Interaction Prediction, which assesses the ability to anticipate user actions and foresee their effects on the interface; and Instruction Understanding, which tests whether a model can grasp task goals and plan correct execution steps.

Graphical User Interface (GUI) task automation, such as booking a flight, editing a presentation, or configuring system settings, poses unique challenges for AI agents Wu et al. (2024a); Hong et al.

(2024); Xu et al. (2024a); He et al. (2024). Recent approaches have leveraged large vision–language models (VLMs) with techniques such as prompt engineering Agashe et al. (2025); Xie et al. (2025a), supervised fine-tuning (SFT) Wu et al. (2024b); Hong et al. (2024); Lin et al. (2025); Liu et al. (2025); Xu et al. (2024b), and reinforcement learning (RL) Lian et al. (2025); Luo et al. (2025), achieving strong task performance in many applications. However, GUI agents still fail in many real-world scenarios Xie et al. (2025c). For example, agents may misinterpret widget functions in unfamiliar applications, fail to predict correct action parameters, or struggle with multi-step planning and error recovery in long-horizon GUI tasks. Our analysis suggests that a primary reason for these failures is that the used VLMs lack the necessary GUI knowledge. While prompt engineering, SFT, and RL can improve reasoning, grounding, and planning abilities, they contribute little to injecting new GUI knowledge Ovadia et al. (2024), which also plays important roles in solving GUI tasks.

Different from most existing benchmarks that primarily evaluate task success, which mainly focus on the grounding Li et al. (2025); Cheng et al. (2024); Jurmu et al. (2008), reasoning, and planning Lin et al. (2024) capabilities of GUI agents, our work targets the missing dimension of knowledge evaluation. To systematically examine these knowledge gaps, we introduce GUI-Knowledge Bench, a benchmark designed to assess the extent of GUI knowledge encoded in VLMs prior to downstream tasks, while also serving as a diagnostic tool to guide the design of VLM-based agent systems. The benchmark is constructed from over 40,000 screenshots and 400 execution trajectories spanning 292 applications across six platforms (Web, Android, MacOS, Windows, Linux, iOS)). Through a combination of automated generation and manual annotation, we derive a set of 3483 knowledge-centric questions that systematically test VLMs' knowledge in GUI.

We categorize the GUI knowledge into three complementary aspects derived from common agent failure modes: (1) interface perception, which involves recognizing widget functions, layout semantics, and state cues (e.g., enabled/disabled, selected/focused); (2) interaction prediction, which involves anticipating action outcomes and preconditions (e.g., what changes after toggling a switch or submitting a form, and which parameters are required); and (3) instruction understanding, which focuses on grounding natural-language instructions into executable, multi-step action sequences with coherent plans. This categorization enables a systematic examination of which components of GUI knowledge are already present in current models and which remain underdeveloped.

Our evaluation reveals that current VLM models still are still short of enough knowledge in these three categories for completing real world GUI tasks. First, although VLMs perform well at discerning different widget functions and layout semantics but struggle to reason about current states; they may recognize interface components but fail to infer the system's actual state information. Second, VLMs underperform in interaction prediction, showing difficulties in anticipating correct action outcomes and required parameters. They frequently confuse click actions with other types of actions, a behavior commonly observed in many models. Third, VLMs struggle with judging task completion states and understanding human instructions. Some tasks are easy to complete, yet they still fail because the models do not understand the goals of the tasks. These findings highlight critical gaps in the internal GUI knowledge of current VLMs, revealing that while they can perceive interface elements, their understanding about state dynamics and interaction outcomes remains limited. Our contributions are as followed:

- We introduce GUI-Knowledge bench, designed to evaluate GUI knowledge in base VLMs. Experiments on real world GUI environment further validates the close link between GUI knowledge and task success.
- Our evaluation identifies key gaps in state reasoning, action prediction, and judging task completion, providing guidance for selecting or training VLMs prior to downstream GUI tasks.

2 Related Work

2.1 GUI Agent

Progress in GUI task automation has largely relied on pretrained vision–language models (VLMs), with improvements driven by supervised fine-tuning (SFT), reinforcement learning (RL), and synthetic data generation. SFT-based methods train VLMs on large-scale GUI datasets to enhance element grounding and action prediction, as seen in OS-Atlas Wu et al. (2024b), CogAgent Hong et al.

Benchmark	Evaluation Scope	OS	Applications	Data Scale
ScreenSpot-Pro Li et al. (2025)	Action	3	23	1581
SeeClick Cheng et al. (2024)	Action	5	20+	1272
VideoGUI Lin et al. (2024)	Task	1	11	463
OSWorld Xie et al. (2025c)	Task	1	9	369
MacOSworld Yang et al. (2025)	Task	1	30	202
AndroidWorld Rawles et al. (2024)	Task	1	20	116
MMBench-GUI Wang et al. (2025)	Knowledge	6	-	8000+
Web-CogBench Guo et al. (2025b)	Knowledge	1	14	876
GUI-Knowledge-Bench	Knowledge	**6**	**292**	**3483**

Table 1: Comparison of existing GUI benchmarks and our proposed benchmark across evaluation scope, operating system coverage, application diversity, and data scale. Our benchmark systematically spans multiple OS and applications with a comprehensive scope of GUI knowledge evaluation.

(2024), and ShowUI Lin et al. (2025), while multi-stage pipelines such as InfiGUIAgent Liu et al. (2025) and Aguvis Xu et al. (2024b) further inject reasoning and planning abilities with synthetic data. RL approaches, including UI-AGILE Lian et al. (2025) and GUI-R1 Luo et al. (2025), refine action selection through long-horizon rewards or policy optimization, sometimes achieving superior performance with less training data. To address data scarcity, OS-Genesis and UI-Genie Sun et al. (2024) generate high-quality synthetic trajectories, while multi-agent systems such as GUI-OWL and Mobile-Agent-v3 Wanyan et al. (2025) decompose perception, reasoning, and planning across modules to improve robustness in long-horizon tasks.

Despite these advances, most approaches primarily optimize execution strategies—whether through imitation of expert trajectories, reward shaping, or modular design—without fundamentally enriching the model's internal GUI knowledge. The trained models still fall short in interacting with unfamiliar applications or understanding complex system states. To address this gap, our work systematically evaluates these foundational knowledge deficiencies and introduces a benchmark that identifies missing GUI knowledge in base VLMs prior to downstream training, providing insights into how future approaches may extend beyond standard fine-tuning paradigms.

2.2 GUI Benchmark

Evaluating GUI agents is essential for advancing their capabilities, and existing benchmarks generally fall into three categories. Action-level benchmarks focus on the precision of low-level operations such as mouse and keyboard inputs and accurate element grounding. Examples include ScreenSpot-Pro Li et al. (2025) highlights grounding challenges in professional high-resolution interfaces, SeeClick Cheng et al. (2024) and ScreenSpot Jurmu et al. (2008) for cross-environment grounding. Plan-level evaluations extend beyond single actions to hierarchical execution. VideoGUI Lin et al. (2024), for instance, evaluates GUI agents with high-level and mid-level planning. Task-level benchmarks emphasize end-to-end task success in simulated environments, such as OSWorld Xie et al. (2025c), OSWorld-Verified Xie et al. (2025b), MacOSworld Yang et al. (2025), and AndroidWorld Rawles et al. (2024). Beyond execution, a few recent efforts assess GUI knowledge, such as MMBench-GUI Wang et al. (2025), which tests content understanding and widget semantics, and Web-CogBench Guo et al. (2025b), which probes cognitive reasoning in web navigation. However, these benchmarks remain narrow in application scopes and domain knowledge coverage.

In contrast, we carefully categorize the GUI knowledge into three complementary aspects derived from common agent failure modes, interface perception, interaction prediction and goal interpretation. Our benchmark offers a systematic and comprehensive evaluation of GUI knowledge, spanning multiple platforms and applications, thereby providing a more complete evaluation of base model's GUI knowledge.

3 GUI KNOWLEDGE BENCH

3.1 BENCHMARK OVERVIEW

We introduce GUI Knowledge Bench, a benchmark for systematically evaluating the knowledge vision–language models (VLMs) need to complete graphical user interface (GUI) tasks. Based on common failure patterns in GUI task execution, we identify three complementary dimensions: interface perception, which covers recognizing GUI elements, their states, and layout semantics; interaction prediction, which tests whether models understand how actions change interface states and can anticipate outcomes; and instruction understanding, which examines whether models can interpret task goals and plan multi-step operations. Together, these dimensions capture the core knowledge required for reliable GUI task completion and form the foundation of our benchmark.

3.2 DATA SOURCES AND COLLECTION PIPELINE

To build GUI Knowledge Bench, we aggregate data from multiple sources to ensure both trajectory-level interaction coverage and diverse standalone screenshots.

We leverage existing benchmarks such as GUI-Odyssey and VideoGUI, which provide screenshots paired with tasks and action annotations. In addition, we collect new trajectories by running UITars agents in environments including OSWorld and MacOSWorld, capturing realistic interaction sequences across both mobile and desktop platforms.

To further increase visual diversity and cover a wider range of application interfaces and operating systems, we further gather standalone GUI screenshots. Specifically, we sample from ScreenSpot v2 and extract representative key frames from YouTube tutorials, ensuring coverage of real-world applications, operating systems, and interface layouts. For less common actions, we manually perform operations on MacOS, Linux, and Windows, recording screenshots and corresponding actions.

Together, these sources yield a heterogeneous pool of GUI images and trajectories. From this pool, we construct task-specific question–answer pairs for each evaluation dimension, ensuring sufficient diversity and coverage while minimizing redundancy.

3.3 INTERFACE PERCEPTION

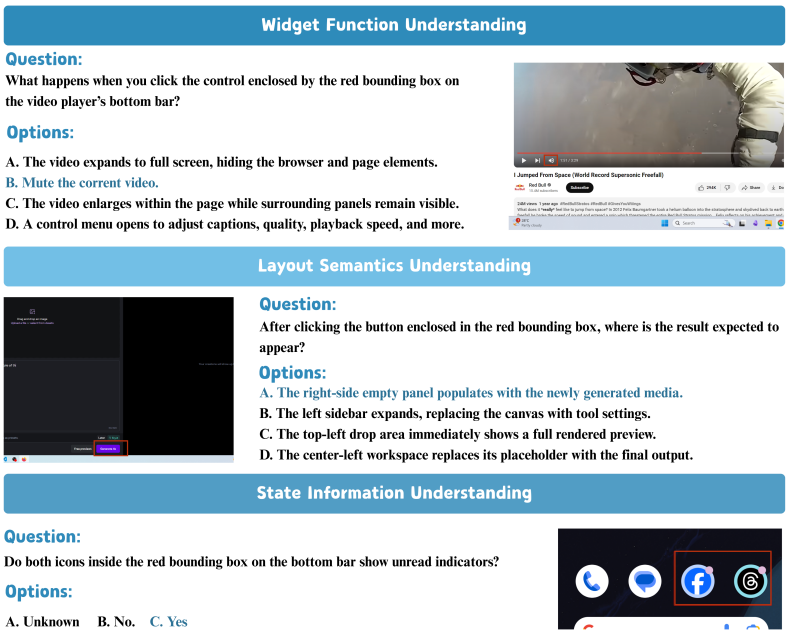

Figure 2: Example questions for Interface Perception. red bounding box

A fundamental requirement for solving GUI tasks is the ability to accurately perceive and interpret interactive elements in GUI. Without such perception, higher-level reasoning—such as predicting interactions or following task instructions—becomes infeasible. This motivates us to evaluate whether VLMs possess sufficient interface perception capabilities.

Specifically, this dimension encloses three aspects: (i) widget function understanding, i.e., recognizing the roles of common interface elements (e.g., three vertical dots for settings, speech bubbles for messaging apps); (ii) state information understanding, such as detecting whether a button is enabled/disabled, selected/focused, or toggled on/off; and (iii) layout semantics understanding, where spatial arrangement encodes critical information (e.g., distinguishing departure and arrival cities by their relative positions, identifying senders and receivers in an email, or inferring file hierarchy from indentation). Correctly perceiving these cues is essential for grounding subsequent reasoning and action.

Task Definition. We formalize the evaluation as a unified multiple-choice question-answering task. Given a question q, a set of candidate options O, and a screenshot S, the model is required to select the correct answer o^* and provide its reasoning in thought t: VLM : $(S, q, O) \mapsto (t, o^*)$.

Our questions include two types: (1) multiple-choice with four candidates, and (2) judgment with Yes/No/Unknown. To reduce the burden of visual grounding, the relevant regions in the screenshot S are highlighted using red dots or bounding boxes. This design ensures the evaluation focuses on whether the model possesses the required GUI knowledge rather than its grounding ability.

Task Collection and Curation. To construct the evaluation set, we first have human annotators design an initial set of seed questions based on the collected GUI screenshots. We then leverage GPT-5 to expand this pool with additional candidate questions, increasing diversity while maintaining relevance. Questions that can be answered based solely on the text, without viewing the screenshot, are removed using Qwen 2.5VL to ensure visual understanding is necessary. Finally, the remaining questions are manually verified for correctness, and relevant regions in the screenshots are annotated to support precise visual grounding. This pipeline ensures that the evaluation focuses on interface perception knowledge rather than being confounded by grounding or annotation errors.

3.4 Interaction Prediction

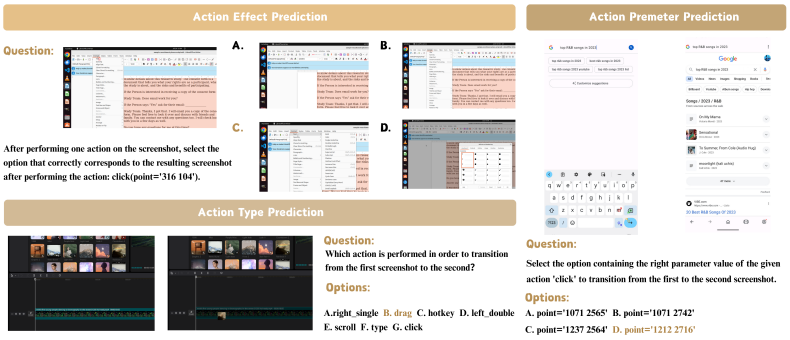

Figure 3: Example questions for Interaction Prediction.

A core requirement for solving GUI tasks is understanding how actions change the interface state. Unlike physical environments, GUI interactions follow symbolic and platform-specific rules (e.g., toggling a switch, typing text, dragging windows), which are often subtle and context-dependent. Without a proper understanding of these interaction dynamics, models cannot reliably predict the consequences of actions or predict right action types/parameters to complete a GUI task. This motivates our evaluation of whether VLMs can reason about action–state transitions in GUI environments.

Interaction prediction is evaluated through two complementary tasks: (i) Action effect prediction, where the model is provided with a current screenshot S and an action a, and must select the resulting screenshot S' from a set of candidate options; (ii) Action prediction, where the model is given two consecutive screenshots (s, s') and must infer the action a that caused the transition. Action

prediction is further divided into action type prediction, which identifies the correct action category, and action parameter prediction, which selects the appropriate arguments such as click coordinates, typed content, or drag vectors.

Task Definition. We formalize GUI interaction dynamics as a state–action transition $S + a \rightarrow S'$, where S represents the current screenshot, S' the consequent screenshots and a the action $a = (a_{\text{type}}, a_{\text{param}})$. (i) Action Effect Prediction. The model is given S and a, and is required to select the resulting screenshots from a set of candidate screenshot options O: VLM : $(S, a, O) \mapsto S'$. (ii) Action Prediction. The model is given two consecutive screenshots (S, S') and a set of candidate action types O_{type}, and must select the correct action type a_{type}: VLM : $(S, S', O_{\text{type}}) \mapsto a_{\text{type}}$. Given the correct action type a_{type} and the same state pair (S, S'), the model selects the correct action parameters from a candidate set O_{param}: VLM : $(S, S', a_{\text{type}}, O_{\text{param}}) \mapsto a_{\text{param}}$.

Task Collection and Curation. To construct challenging distractor options, we design task-specific strategies for different prediction settings. For action effect prediction, candidate screenshots include the preceding screenshots, the true next screenshot, and visually similar but different screenshots sampled from current trajectories. For action type prediction, the model must choose from the full set of possible actions defined for the platform, with a unified action space across different platforms. (seven actions for desktop platforms and four actions for mobile platforms) For action parameter prediction, For clicks, bounding boxes of candidate elements are identified using OmniParser, and nearby but incorrect coordinates are sampled; for drags, distractors include reversed directions, shortened distances, or swapped start and end points; for scrolls, distractors vary in direction (up, down, left, right); for typing, inputs are perturbed with case changes, partial deletions, or common typos; and for hotkeys, distractors are drawn from a predefined set of common shortcuts. This design ensures that solving the tasks requires precise reasoning about GUI action dynamics rather than relying on superficial visual or layout cues.

3.5 Instruction Understanding

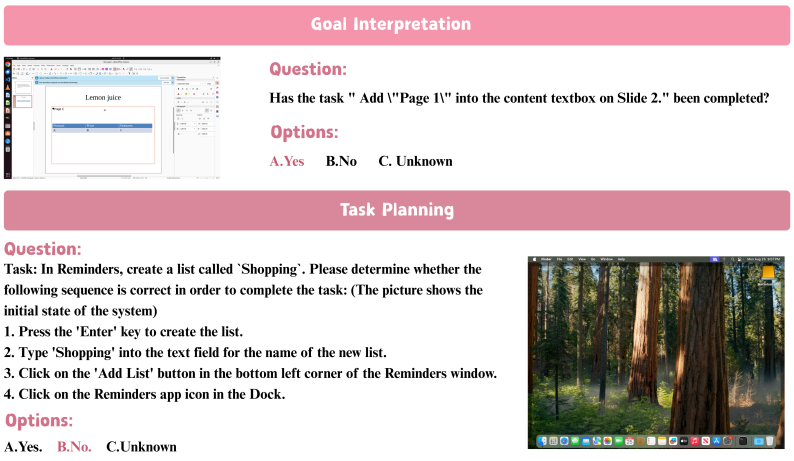

Figure 4: Example questions for Instruction Understanding.

Instruction understanding evaluates whether a model can interpret natural-language tasks and map them to a sequence of GUI operations. This ability is critical because many failures in GUI automation stem from misunderstanding task goals or misinterpreting user intent. Accurately understanding instructions and producing a feasible sequence of steps is essential for performing high-level tasks in GUI environments.

We assess two complementary abilities: (i) goal interpretation, which evaluates whether a model can determine if a task has been successfully completed based on history screenshots; and (ii) task planning, which evaluates whether a model can reorder a set of candidate option steps into the correct sequence required to achieve the task goal. Together, these tasks test the model's ability to both verify and generate high-level plans.

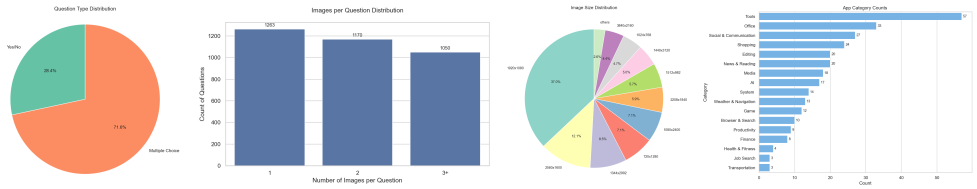

Figure 5: Statistics of GUI Knowledge Bench, including question type distribution, images per question, image size distribution, and app category counts.

Task Definition. For goal interpretation, the model receives a natural-language task description t and history screenshots, and must select the correct option $o^* \in \{\text{Yes}, \text{No}, \text{Unknown}\}$ indicating whether the task is completed: VLM : $(S_{1:T}, t, O) \mapsto o^*$. For task planning, the model is given a natural-language task description t, current screenshot S and a set of candidate orderings $O = \{\pi_1, \pi_2, \ldots, \pi_m\}$, where each π_i is a possible permutation of the operation steps. The model must select the correct ordering π^* from O: VLM : $(t, S, O) \mapsto \pi^*$.

Task Collection and Curation. For goal interpretation, human annotators label each trajectory as successful or unsuccessful based on last one to five screenshots of the trajectory, and some successful trajectories are augmented by removing the final screenshot to create unsuccessful ones. For task planning, operation plans are first generated by Chat-GPT-5 and then verified by annotators. The annotated steps are automatically shuffled to form multiple-choice ordering questions, with longer sequences retaining initial steps and only permuting later steps. For shorter sequences, additional question formats are created by converting the shuffled sequence into Yes/No/Unknown questions, or into operation-level fill-in-the-blank questions with distractor steps. Tasks solvable without observing screenshots are filtered out using Qwen-VL-2.5-7B ensuring that the evaluation emphasizes high-level reasoning over task goals.

3.6 Benchmark Statistics

Figure 5 summarizes key statistics of our GUI Knowledge Bench. The benchmark covers a wide variety of application types and question formats, with diverse image counts and resolutions across tasks.

4 Benchmarking VLMs

4.1 Settings

We evaluate a diverse set of both open- and closed-source models on the GUI Knowledge Bench. The closed-source set includes Claude-Sonnet-4 Anthropic (2025), Doubao-V-Pro (Doubao-1.5-Thinking-Vision-Pro-250428) Guo et al. (2025a), Gemini-2.5-Pro Comanici et al. (2025), GPT-5-chat OpenAI (2025a), O3OpenAI (2025b), and GLM-4.5 Team et al. (2025). The open-source set covers Qwen2.5-72B (Qwen2.5-VL-72B-Instruct), Qwen2.5-7B (Qwen2.5-VL-7B-Instruct) Bai et al. (2025), UITARS-1.5-7B Qin et al. (2025). Apart from necessary model-specific settings, all other parameters (e.g., temperature, top-p) were kept consistent across evaluations. Please refer the appendix for the detailed message template for each knowledge categories.

4.2 Benchmarking Results

Table 2 highlights results with several key observations. First, o3 achieves strong performance across multiple metrics, consistent with its high success rate in real GUI tasks; notably, in the OS-World benchmark under the Agent framework category, four of the top five agents leverage o3 (e.g., Agent-S2.5 w/ O3 50-step version and 100-step version, Jedi-7B w/ O3 w/ 50-step version and 100-step version). Agashe et al. (2025); Xie et al. (2025a) This is likely because o3 effectively replaces the auxiliary modules that were removed or made optional. Second, UITARS-1.5-7B, trained on Qwen2.5VL-7B, shows improvements in instruction understanding and goal reasoning but a decline in interface perception. Third, smaller models retain only limited knowledge, suggesting that

retrieval-augmented generation or knowledge-base integration may be a viable approach to enhance GUI agent performance.

Table 2: Performance on GUI Knowledge Bench across three dimensions. Bold numbers indicate the best results in each sub-task.

Model	Interface Perception			Interaction Prediction			Instruction Understanding	
	state	widget	layout	effect	type	parameter	goal	plan
GLM-4.5 Team et al. (2025)	49.54%	48.10%	53.55%	27.07%	17.55%	38.13%	30.87%	91.91%
Claude-Sonnet-4 Anthropic (2025)	70.18%	78.44%	78.06%	41.55%	62.52%	45.75%	63.78%	94.82%
Doubao-V-Pro Guo et al. (2025a)	72.48%	83.65%	81.29%	68.10%	75.64%	52.17%	33.07%	94.17%
Gemini-2.5-Pro Comanici et al. (2025)	81.19%	**84.36%**	87.10%	71.72%	73.25%	**60.31%**	66.46%	92.56%
GPT-5-Chat OpenAI (2025a)	78.90%	84.12%	**88.39%**	73.10%	71.55%	57.89%	67.40%	91.26%
O3 OpenAI (2025b)	**83.03%**	84.12%	**88.39%**	**76.55%**	**75.98%**	**60.31%**	**67.56%**	**95.47%**
Qwen2.5VL-7B Bai et al. (2025)	53.21%	67.77%	60.00%	52.59%	50.60%	44.54%	17.01%	48.87%
Qwen2.5VL-72B Bai et al. (2025)	69.72%	77.49%	80.00%	62.93%	64.91%	44.89%	61.89%	85.44%
UITARS-1.5-7B Qin et al. (2025)	49.54%	59.48%	59.35%	22.76%	59.11%	43.50%	39.84%	55.34%

4.3 Error Analysis and Discussion

Most models handle widget functions and layout semantics well but struggle with system state perception. As shown in Figure 6, in Safari, an update notification is often mistaken for a blocking pop-up, leading to incorrect predictions, while in PowerPoint, models can understand the effect of delete action but not which element is selected. Our benchmark reveals that current models underperform in system state perception, despite its crucial role in GUI tasks.

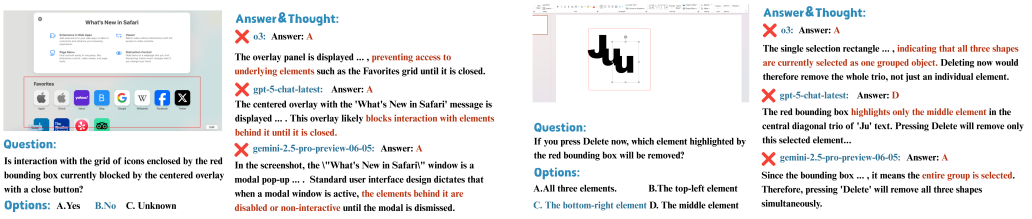

Figure 6: Example failure cases of interface perception questions.

On desktop, models often confuse click, double-click, and right-click. This is partly because different operating systems and applications treat these interactions differently: in some contexts, single, double, or right clicks can substitute for each other, while in others the distinction is strict. Humans often try multiple actions to achieve a goal, but models predict a single action based on learned patterns. As a result, less frequent actions like double-click or right-click are more prone to misprediction, especially for smaller models. Please refer to appendix to see the confusion matrix of action type prediction.

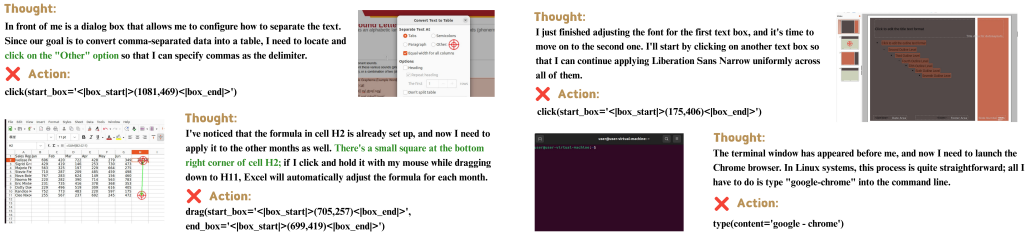

Figure 7: Example failure cases due to lack of interaction prediction knowledge.

4.4 Results on OSWorld

This section examines the role of three types of knowledge in enabling successful real-world GUI task execution. To this end, we run several experiments using UITARS-1.5-7B on the OSWorld benchmark and report our key findings here.

Interface Perception Knowledge. In our evaluation, we identified several tasks that the model consistently failed to solve even under the pass@32 setting. We attribute some of them to lack of knowledge of the interface. Two example failure cases are shown in Figure 8. When asked to add a note, the model repeatedly attempted to insert comments or text boxes, incorrectly treating these actions as equivalent to adding a note. In reality, adding a note requires first enabling the Notes pane through the View menu and then placing the note at the bottom of the slide. Similarly, in converting comma-separated text into a table, the model repeatedly failed because it did not specify the delimiter, a necessary step for correct execution. These cases suggest that the failures stem from missing application-specific knowledge rather than inherent reasoning limitations. Importantly, once the required knowledge was provided, the model was able to complete these tasks successfully.

Table 3: Effect of appending operation plans on UITARS-1.5-7B.

Metric	UITARS-1.5-7B (base)	+ GPT-4o plan	+ OSWorld-human plan	+ o3 plan
Pass@1 (%)	24.81	27.59	28.20	**30.79**

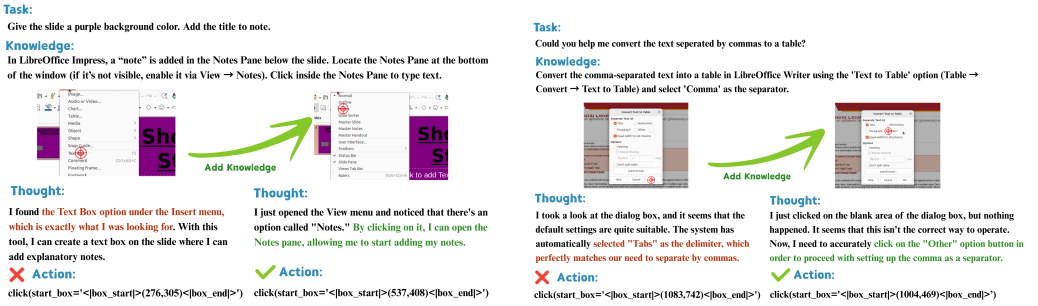

Figure 8: Example failure cases due to lack of interface perception knowledge.

Interaction Prediction Knowledge. Many errors of models occur because it lacks knowledge for localizing interface elements correctly. Prior work has improved localization using masks, accessibility trees, or APIs. Another promising approach is to leverage actions themselves for self-verification, using visual prompts to check if the executed action was correct.

Instruction Understanding Knowledge. We generated knowledge about operation plans from GPT-4o and o3 conditioned on task instructions, and used human-authored operation plans from OSWorld-human. Each plan was appended to the original task description as an additional knowledge for UITARS-1.5-7B. Results are summarized in Table 3. These results show that providing knowledge about operation plans improves task performance, highlighting the importance of instruction understanding for task completion. Notably, o3-generated plans achieve the largest gain, surpassing human-authored plans and aligning with o3's top performance across our benchmark evaluations.

5 Conclusion

We introduces GUI Knowledge Bench, a novel benchmark designed to evaluate the GUI knowledge encoded in vision-language models (VLMs) before downstream training. By analyzing common failure patterns in GUI task execution, the benchmark categorizes GUI knowledge into three dimensions: interface perception, interaction prediction, and instruction understanding. The evaluation reveals significant gaps in current VLMs' understanding of system states, action outcomes, and task completion verification. These findings highlight the necessity of enriching VLMs with domain-specific GUI knowledge to enhance their performance in real-world GUI tasks and provide insights to guide the development of more capable GUI agents.

REFERENCES

Saaket Agashe, Kyle Wong, Vincent Tu, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent s2: A compositional generalist-specialist framework for computer use agents. *arXiv preprint arXiv:2504.00906*, 2025.

Anthropic. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic PBC, May 2025. URL https://www.anthropic.com/claude-4-system-card. Accessed: 2025-09-25.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

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

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, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, Mu Cai, Mohammed Badawi, Corey Fry, Ale Hartman, Daniel Zheng, Eric Jia, James Keeling, Annie Louis, Ying Chen, Efren Robles, Wei-Chih Hung, Howard Zhou, Nikita Saxena, Sonam Goenka, Olivia Ma, Zach Fisher, Mor Hazan Taege, Emily Graves, David Steiner, Yujia Li, Sarah Nguyen, Rahul Sukthankar, Joe Stanton, Ali Eslami, Gloria Shen, Berkin Akin, Alexey Guseynov, Yiqian Zhou, Jean-Baptiste Alayrac, Armand Joulin, Efrat Farkash, Ashish Thapliyal, Stephen Roller, Noam Shazeer, Todor Davchev, Terry Koo, Hannah Forbes-Pollard, Kartik Audhkhasi, Greg Farquhar, Adi Mayrav Gilady, Maggie Song, John Aslanides, Piermaria Mendolicchio, Alicia Parrish, John Blitzer, Pramod Gupta, Xiaoen Ju, Xiaochen Yang, Puranjay Datta, Andrea Tacchetti, Sanket Vaibhav Mehta, Gregory Dibb, Shubham Gupta, Federico Piccinini, Raia Hadsell, Sujee Rajayogam, Jiepu Jiang, Patrick Griffin, Patrik Sundberg, Jamie Hayes, Alexey Frolov, Tian Xie, Adam Zhang, Kingshuk Dasgupta, Uday Kalra, Lior Shani, Klaus Macherey, Tzu-Kuo Huang, Liam MacDermed, Karthik Duddu, Paulo Zacchello, Zi Yang, Jessica Lo, Kai Hui, Matej Kastelic, Derek Gasaway, Qijun Tan, Summer Yue, Pablo Barrio, John Wieting, Weel Yang, Andrew Nystrom, Solomon Demmessie, Anselm Levskaya, Fabio Viola, Chetan Tekur, Greg Billock, George Necula, Mandar Joshi, Rylan Schaeffer, Swachhand Lokhande, Christina Sorokin, Pradeep Shenoy, Mia Chen, Mark Collier, Hongji Li, Taylor Bos, Nevan Wichers, Sun Jae Lee, Angéline Pouget, Santhosh Thangaraj, Kyriakos Axiotis, Phil Crone, Rachel Sterneck, Nikolai Chinaev, Victoria Krakovna, Oleksandr Ferludin, Ian Gemp, Stephanie Winkler, Dan Goldberg, Ivan Korotkov, Kefan Xiao, Malika Mehrotra, Sandeep Mariserla, Vihari Piratla, Terry Thurk, Khiem Pham, Hongxu Ma, Alexandre Senges, Ravi Kumar, Clemens Meyer, Ellie Talius, Nuo Wang Pierse, Ballie Sandhu, Horia Toma, Kuo Lin, Swaroop Nath, Tom Stone, Dorsa Sadigh, Nikita Gupta, Arthur Guez, Avi Singh, Matt Thomas, Tom Duerig, Yuan Gong, Richard Tanburn, Lydia Lihui Zhang, Phuong Dao, Mohamed Hammad, Sirui Xie, Shruti Rijhwani, Ben Murdoch, Duhyeon Kim, Will Thompson, Heng-Tze Cheng,

Daniel Sohn, Pablo Sprechmann, Qiantong Xu, Srinivas Tadepalli, Peter Young, Ye Zhang, Hansa Srinivasan, Miranda Aperghis, Aditya Ayyar, Hen Fitoussi, Ryan Burnell, David Madras, Mike Dusenberry, Xi Xiong, Tayo Oguntebi, Ben Albrecht, Jörg Bornschein, Jovana Mitrović, Mason Dimarco, Bhargav Kanagal Shamanna, Premal Shah, Eren Sezener, Shyam Upadhyay, Dave Lacey, Craig Schiff, Sebastien Baur, Sanjay Ganapathy, Eva Schnider, Mateo Wirth, Connor Schenck, Andrey Simanovsky, Yi-Xuan Tan, Philipp Fränken, Dennis Duan, Bharath Mankalale, Nikhil Dhawan, Kevin Sequeira, Zichuan Wei, Shivanker Goel, Caglar Unlu, Yukun Zhu, Haitian Sun, Ananth Balashankar, Kurt Shuster, Megh Umekar, Mahmoud Alnahlawi, Aäron van den Oord, Kelly Chen, Yuexiang Zhai, Zihang Dai, Kuang-Huei Lee, Eric Doi, Lukas Zilka, Rohith Vallu, Disha Shrivastava, Jason Lee, Hisham Husain, Honglei Zhuang, Vincent Cohen-Addad, Jarred Barber, James Atwood, Adam Sadovsky, Quentin Wellens, Steven Hand, Arunkumar Rajendran, Aybuke Turker, CJ Carey, Yuanzhong Xu, Hagen Soltau, Zefei Li, Xinying Song, Conglong Li, Iurii Kemaev, Sasha Brown, Andrea Burns, Viorica Patraucean, Piotr Stanczyk, Renga Aravamudhan, Mathieu Blondel, Hila Noga, Lorenzo Blanco, Will Song, Michael Isard, Mandar Sharma, Reid Hayes, Dalia El Badawy, Avery Lamp, Itay Laish, Olga Kozlova, Kelvin Chan, Sahil Singla, Srinivas Sunkara, Mayank Upadhyay, Chang Liu, Aijun Bai, Jarek Wilkiewicz, Martin Zlocha, Jeremiah Liu, Zhuowan Li, Haiguang Li, Omer Barak, Ganna Raboshchuk, Jiho Choi, Fangyu Liu, Erik Jue, Mohit Sharma, Andreea Marzoca, Robert Busa-Fekete, Anna Korsun, Andre Elisseeff, Zhe Shen, Sara Mc Carthy, Kay Lamerigts, Anahita Hosseini, Hanzhao Lin, Charlie Chen, Fan Yang, Kushal Chauhan, Mark Omernick, Dawei Jia, Karina Zainullina, Demis Hassabis, Danny Vainstein, Ehsan Amid, Xiang Zhou, Ronny Votel, Eszter Vértes, Xinjian Li, Zongwei Zhou, Angeliki Lazaridou, Brendan McMahan, Arjun Narayanan, Hubert Soyer, Sujoy Basu, Kayi Lee, Bryan Perozzi, Qin Cao, Leonard Berrada, Rahul Arya, Ke Chen, Katrina, Xu, Matthias Lochbrunner, Alex Hofer, Sahand Sharifzadeh, Renjie Wu, Sally Goldman, Pranjal Awasthi, Xuezhi Wang, Yan Wu, Claire Sha, Biao Zhang, Maciej Mikuła, Filippo Graziano, Siobhan Mcloughlin, Irene Giannoumis, Youhei Namiki, Chase Malik, Carey Radebaugh, Jamie Hall, Ramiro Leal-Cavazos, Jianmin Chen, Vikas Sindhwani, David Kao, David Greene, Jordan Griffith, Chris Welty, Ceslee Montgomery, Toshihiro Yoshino, Liangzhe Yuan, Noah Goodman, Assaf Hurwitz Michaely, Kevin Lee, KP Sawhney, Wei Chen, Zheng Zheng, Megan Shum, Nikolay Savinov, Etienne Pot, Alex Pak, Morteza Zadimoghaddam, Sijal Bhatnagar, Yoad Lewenberg, Blair Kutzman, Ji Liu, Lesley Katzen, Jeremy Selier, Josip Djolonga, Dmitry Lepikhin, Kelvin Xu, Jacky Liang, Jiewen Tan, Benoit Schillings, Muge Ersoy, Pete Blois, Bernd Bandemer, Abhimanyu Singh, Sergei Lebedev, Pankaj Joshi, Adam R. Brown, Evan Palmer, Shreya Pathak, Komal Jalan, Fedir Zubach, Shuba Lall, Randall Parker, Alok Gunjan, Sergey Rogulenko, Sumit Sanghai, Zhaoqi Leng, Zoltan Egyed, Shixin Li, Maria Ivanova, Kostas Andriopoulos, Jin Xie, Elan Rosenfeld, Auriel Wright, Ankur Sharma, Xinyang Geng, Yicheng Wang, Sam Kwei, Renke Pan, Yujing Zhang, Gabby Wang, Xi Liu, Chak Yeung, Elizabeth Cole, Aviv Rosenberg, Zhen Yang, Phil Chen, George Polovets, Pranav Nair, Rohun Saxena, Josh Smith, Shuo yiin Chang, Aroma Mahendru, Svetlana Grant, Anand Iyer, Irene Cai, Jed McGiffin, Jiaming Shen, Alanna Walton, Antonious Girgis, Oliver Woodman, Rosemary Ke, Mike Kwong, Louis Rouillard, Jinmeng Rao, Zhihao Li, Yuntao Xu, Flavien Prost, Chi Zou, Ziwei Ji, Alberto Magni, Tyler Liechty, Dan A. Calian, Deepak Ramachandran, Igor Krivokon, Hui Huang, Terry Chen, Anja Hauth, Anastasija Ilić, Weijuan Xi, Hyeontaek Lim, Vlad-Doru Ion, Pooya Moradi, Metin Toksoz-Exley, Kalesha Bullard, Miltos Allamanis, Xiaomeng Yang, Sophie Wang, Zhi Hong, Anita Gergely, Cheng Li, Bhavishya Mittal, Vitaly Kovalev, Victor Ungureanu, Jane Labanowski, Jan Wassenberg, Nicolas Lacasse, Geoffrey Cideron, Petar Dević, Annie Marsden, Lynn Nguyen, Michael Fink, Yin Zhong, Tatsuya Kiyono, Desi Ivanov, Sally Ma, Max Bain, Kiran Yalasangi, Jennifer She, Anastasia Petrushkina, Mayank Lunayach, Carla Bromberg, Sarah Hodkinson, Vilobh Meshram, Daniel Vlasic, Austin Kyker, Steve Xu, Jeff Stanway, Zuguang Yang, Kai Zhao, Matthew Tung, Seth Odoom, Yasuhisa Fujii, Justin Gilmer, Eunyoung Kim, Felix Halim, Quoc Le, Bernd Bohnet, Seliem El-Sayed, Behnam Neyshabur, Malcolm Reynolds, Dean Reich, Yang Xu, Erica Moreira, Anuj Sharma, Zeyu Liu, Mohammad Javad Hosseini, Naina Raisinghani, Yi Su, Ni Lao, Daniel Formoso, Marco Gelmi, Almog Gueta, Tapomay Dey, Elena Gribovskaya, Domagoj Ćevid, Sidharth Mudgal, Garrett Bingham, Jianling Wang, Anurag Kumar, Alex Cullum, Feng Han, Konstantinos Bousmalis, Diego Cedillo, Grace Chu, Vladimir Magay, Paul Michel, Ester Hlavnova, Daniele Calandriello, Setareh Ariafar, Kaisheng Yao, Vikash Sehwag, Arpi Vezer, Agustin Dal Lago, Zhenkai Zhu, Paul Kishan Rubenstein, Allen Porter, Anirudh Baddepudi, Oriana Riva, Mihai Dorin Istin, Chih-Kuan Yeh, Zhi Li, Andrew Howard, Nilpa Jha, Jeremy Chen, Raoul de Liedekerke, Zafarali Ahmed, Mikel

Rodriguez, Tanuj Bhatia, Bangju Wang, Ali Elqursh, David Klinghoffer, Peter Chen, Pushmeet Kohli, Te I, Weiyang Zhang, Zack Nado, Jilin Chen, Maxwell Chen, George Zhang, Aayush Singh, Adam Hillier, Federico Lebron, Yiqing Tao, Ting Liu, Gabriel Dulac-Arnold, Jingwei Zhang, Shashi Narayan, Buhuang Liu, Orhan Firat, Abhishek Bhowmick, Bingyuan Liu, Hao Zhang, Zizhao Zhang, Georges Rotival, Nathan Howard, Anu Sinha, Alexander Grushetsky, Benjamin Beyret, Keerthana Gopalakrishnan, James Zhao, Kyle He, Szabolcs Payrits, Zaid Nabulsi, Zhaoyi Zhang, Weijie Chen, Edward Lee, Nova Fallen, Sreenivas Gollapudi, Aurick Zhou, Filip Pavetić, Thomas Köppe, Shiyu Huang, Rama Pasumarthi, Nick Fernando, Felix Fischer, Daria Ćurko, Yang Gao, James Svensson, Austin Stone, Haroon Qureshi, Abhishek Sinha, Apoorv Kulshreshtha, Martin Matysiak, Jieming Mao, Carl Saroufim, Aleksandra Faust, Qingnan Duan, Gil Fidel, Kaan Katircioglu, Raphaël Lopez Kaufman, Dhruv Shah, Weize Kong, Abhishek Bapna, Gellért Weisz, Emma Dunleavy, Praneet Dutta, Tianqi Liu, Rahma Chaabouni, Carolina Parada, Marcus Wu, Alexandra Belias, Alessandro Bissacco, Stanislav Fort, Li Xiao, Fantine Huot, Chris Knutsen, Yochai Blau, Gang Li, Jennifer Prendki, Juliette Love, Yinlam Chow, Pichi Charoenpanit, Hidetoshi Shimokawa, Vincent Coriou, Karol Gregor, Tomas Izo, Arjun Akula, Mario Pinto, Chris Hahn, Dominik Paulus, Jiaxian Guo, Neha Sharma, Cho-Jui Hsieh, Adaeze Chukwuka, Kazuma Hashimoto, Nathalie Rauschmayr, Ling Wu, Christof Angermueller, Yulong Wang, Sebastian Gerlach, Michael Pliskin, Daniil Mirylenka, Min Ma, Lexi Baugher, Bryan Gale, Shaan Bijwadia, Nemanja Rakićević, David Wood, Jane Park, Chung-Ching Chang, Babi Seal, Chris Tar, Kacper Krasowiak, Yiwen Song, Georgi Stephanov, Gary Wang, Marcello Maggioni, Stein Xudong Lin, Felix Wu, Shachi Paul, Zixuan Jiang, Shubham Agrawal, Bilal Piot, Alex Feng, Cheolmin Kim, Tulsee Doshi, Jonathan Lai, Chuqiao, Xu, Sharad Vikram, Ciprian Chelba, Sebastian Krause, Vincent Zhuang, Jack Rae, Timo Denk, Adrian Collister, Lotte Weerts, Xianghong Luo, Yifeng Lu, Håvard Garnes, Nitish Gupta, Terry Spitz, Avinatan Hassidim, Lihao Liang, Izhak Shafran, Peter Humphreys, Kenny Vassigh, Phil Wallis, Virat Shejwalkar, Nicolas Perez-Nieves, Rachel Hornung, Melissa Tan, Beka Westberg, Andy Ly, Richard Zhang, Brian Farris, Jongbin Park, Alec Kosik, Zeynep Cankara, Andrii Maksai, Yunhan Xu, Albin Cassirer, Sergi Caelles, Abbas Abdolmaleki, Mencher Chiang, Alex Fabrikant, Shravya Shetty, Luheng He, Mai Giménez, Hadi Hashemi, Sheena Panthaplackel, Yana Kulizhskaya, Salil Deshmukh, Daniele Pighin, Robin Alazard, Disha Jindal, Seb Noury, Pradeep Kumar S, Siyang Qin, Xerxes Dotiwalla, Stephen Spencer, Mohammad Babaeizadeh, Blake JianHang Chen, Vaibhav Mehta, Jennie Lees, Andrew Leach, Penporn Koanantakool, Ilia Akolzin, Ramona Comanescu, Junwhan Ahn, Alexey Svyatkovskiy, Basil Mustafa, David D'Ambrosio, Shiva Mohan Reddy Garlapati, Pascal Lamblin, Alekh Agarwal, Shuang Song, Pier Giuseppe Sessa, Pauline Coquinot, John Maggs, Hussain Masoom, Divya Pitta, Yaqing Wang, Patrick Morris-Suzuki, Billy Porter, Johnson Jia, Jeffrey Dudek, Raghavender R, Cosmin Paduraru, Alan Ansell, Tolga Bolukbasi, Tony Lu, Ramya Ganeshan, Zi Wang, Henry Griffiths, Rodrigo Benenson, Yifan He, James Swirhun, George Papamakarios, Aditya Chawla, Kuntal Sengupta, Yan Wang, Vedrana Milutinovic, Igor Mordatch, Zhipeng Jia, Jamie Smith, Will Ng, Shitij Nigam, Matt Young, Eugen Vušak, Blake Hechtman, Sheela Goenka, Avital Zipori, Kareem Ayoub, Ashok Popat, Trilok Acharya, Luo Yu, Dawn Bloxwich, Hugo Song, Paul Roit, Haiqiong Li, Aviel Boag, Nigamaa Nayakanti, Bilva Chandra, Tianli Ding, Aahil Mehta, Cath Hope, Jiageng Zhang, Idan Heimlich Shtacher, Kartikeya Badola, Ryo Nakashima, Andrei Sozanschi, Iulia Comşa, Ante Žužul, Emily Caveness, Julian Odell, Matthew Watson, Dario de Cesare, Phillip Lippe, Derek Lockhart, Siddharth Verma, Huizhong Chen, Sean Sun, Lin Zhuo, Aditya Shah, Prakhar Gupta, Alex Muzio, Ning Niu, Amir Zait, Abhinav Singh, Meenu Gaba, Fan Ye, Prajit Ramachandran, Mohammad Saleh, Raluca Ada Popa, Ayush Dubey, Frederick Liu, Sara Javanmardi, Mark Epstein, Ross Hemsley, Richard Green, Nishant Ranka, Eden Cohen, Chuyuan Kelly Fu, Sanjay Ghemawat, Jed Borovik, James Martens, Anthony Chen, Pranav Shyam, André Susano Pinto, Ming-Hsuan Yang, Alexandru Ţifrea, David Du, Boqing Gong, Ayushi Agarwal, Seungyeon Kim, Christian Frank, Saloni Shah, Xiaodan Song, Zhiwei Deng, Ales Mikhalap, Kleopatra Chatziprimou, Timothy Chung, Toni Creswell, Susan Zhang, Yennie Jun, Carl Lebsack, Will Truong, Slavica Andačić, Itay Yona, Marco Fornoni, Rong Rong, Serge Toropov, Afzal Shama Soudagar, Andrew Audibert, Salah Zaiem, Zaheer Abbas, Andrei Rusu, Sahitya Potluri, Shitao Weng, Anastasios Kementsietsidis, Anton Tsitsulin, Daiyi Peng, Natalie Ha, Sanil Jain, Tejasi Latkar, Simeon Ivanov, Cory McLean, Anirudh GP, Rajesh Venkataraman, Canoee Liu, Dilip Krishnan, Joel D'sa, Roey Yogev, Paul Collins, Benjamin Lee, Lewis Ho, Carl Doersch, Gal Yona, Shawn Gao, Felipe Tiengo Ferreira, Adnan Ozturel, Hannah Muckenhirn, Ce Zheng, Gargi Balasubramaniam, Mudit Bansal, George van den Driessche, Sivan Eiger, Salem Haykal, Vedant Misra, Abhimanyu Goyal, Danilo Martins,

Gary Leung, Jonas Valfridsson, Four Flynn, Will Bishop, Chenxi Pang, Yoni Halpern, Honglin Yu, Lawrence Moore, Yuvein, Zhu, Sridhar Thiagarajan, Yoel Drori, Zhisheng Xiao, Lucio Dery, Rolf Jagerman, Jing Lu, Eric Ge, Vaibhav Aggarwal, Arjun Khare, Vinh Tran, Oded Elyada, Ferran Alet, James Rubin, Ian Chou, David Tian, Libin Bai, Lawrence Chan, Lukasz Lew, Karolis Misiunas, Taylan Bilal, Aniket Ray, Sindhu Raghuram, Alex Castro-Ros, Viral Carpenter, CJ Zheng, Michael Kilgore, Josef Broder, Emily Xue, Praveen Kallakuri, Dheeru Dua, Nancy Yuen, Steve Chien, John Schultz, Saurabh Agrawal, Reut Tsarfaty, Jingcao Hu, Ajay Kannan, Dror Marcus, Nisarg Kothari, Baochen Sun, Ben Horn, Matko Bošnjak, Ferjad Naeem, Dean Hirsch, Lewis Chiang, Boya Fang, Jie Han, Qifei Wang, Ben Hora, Antoine He, Mario Lučić, Beer Changpinyo, Anshuman Tripathi, John Youssef, Chester Kwak, Philippe Schlattner, Cat Graves, Rémi Leblond, Wenjun Zeng, Anders Andreassen, Gabriel Rasskin, Yue Song, Eddie Cao, Junhyuk Oh, Matt Hoffman, Wojtek Skut, Yichi Zhang, Jon Stritar, Xingyu Cai, Saarthak Khanna, Kathie Wang, Shriya Sharma, Christian Reisswig, Younghoon Jun, Aman Prasad, Tatiana Sholokhova, Preeti Singh, Adi Gerzi Rosenthal, Anian Ruoss, Françoise Beaufays, Sean Kirmani, Dongkai Chen, Johan Schalkwyk, Jonathan Herzig, Been Kim, Josh Jacob, Damien Vincent, Adrian N Reyes, Ivana Balazevic, Léonard Hussenot, Jon Schneider, Parker Barnes, Luis Castro, Spandana Raj Babbula, Simon Green, Serkan Cabi, Nico Duduta, Danny Driess, Rich Galt, Noam Velan, Junjie Wang, Hongyang Jiao, Matthew Mauger, Du Phan, Miteyan Patel, Vlado Galić, Jerry Chang, Eyal Marcus, Matt Harvey, Julian Salazar, Elahe Dabir, Suraj Satishkumar Sheth, Amol Mandhane, Hanie Sedghi, Jeremiah Willcock, Amir Zandieh, Shruthi Prabhakara, Aida Amini, Antoine Miech, Victor Stone, Massimo Nicosia, Paul Niemczyk, Ying Xiao, Lucy Kim, Sławek Kwasiborski, Vikas Verma, Ada Maksutaj Oflazer, Christoph Hirnschall, Peter Sung, Lu Liu, Richard Everett, Michiel Bakker, Ágoston Weisz, Yufei Wang, Vivek Sampathkumar, Uri Shaham, Bibo Xu, Yasemin Altun, Mingqiu Wang, Takaaki Saeki, Guanjie Chen, Emanuel Taropa, Shanthal Vasanth, Sophia Austin, Lu Huang, Goran Petrovic, Qingyun Dou, Daniel Golovin, Grigory Rozhdestvenskiy, Allie Culp, Will Wu, Motoki Sano, Divya Jain, Julia Proskurnia, Sébastien Cevey, Alejandro Cruzado Ruiz, Piyush Patil, Mahdi Mirzazadeh, Eric Ni, Javier Snaider, Lijie Fan, Alexandre Fréchette, AJ Pierigiovanni, Shariq Iqbal, Kenton Lee, Claudio Fantacci, Jinwei Xing, Lisa Wang, Alex Irpan, David Raposo, Yi Luan, Zhuoyuan Chen, Harish Ganapathy, Kevin Hui, Jiazhong Nie, Isabelle Guyon, Heming Ge, Roopali Vij, Hui Zheng, Dayeong Lee, Alfonso Castaño, Khuslen Baatarsukh, Gabriel Ibagon, Alexandra Chronopoulou, Nicholas FitzGerald, Shashank Viswanadha, Safeen Huda, Rivka Moroshko, Georgi Stoyanov, Prateek Kolhar, Alain Vaucher, Ishaan Watts, Adhi Kuncoro, Henryk Michalewski, Satish Kambala, Bat-Orgil Batsaikhan, Alek Andreev, Irina Jurenka, Maigo Le, Qihang Chen, Wael Al Jishi, Sarah Chakera, Zhe Chen, Aditya Kini, Vikas Yadav, Aditya Siddhant, Ilia Labzovsky, Balaji Lakshminarayanan, Carrie Grimes Bostock, Pankil Botadra, Ankesh Anand, Colton Bishop, Sam Conway-Rahman, Mohit Agarwal, Yani Donchev, Achintya Singhal, Félix de Chaumont Quitry, Natalia Ponomareva, Nishant Agrawal, Bin Ni, Kalpesh Krishna, Masha Samsikova, John Karro, Yilun Du, Tamara von Glehn, Caden Lu, Christopher A. Choquette-Choo, Zhen Qin, Tingnan Zhang, Sicheng Li, Divya Tyam, Swaroop Mishra, Wing Lowe, Colin Ji, Weiyi Wang, Manaal Faruqui, Ambrose Slone, Valentin Dalibard, Arunachalam Narayanaswamy, John Lambert, Pierre-Antoine Manzagol, Dan Karliner, Andrew Bolt, Ivan Lobov, Aditya Kusupati, Chang Ye, Xuan Yang, Heiga Zen, Nelson George, Mukul Bhutani, Olivier Lacombe, Robert Riachi, Gagan Bansal, Rachel Soh, Yue Gao, Yang Yu, Adams Yu, Emily Nottage, Tania Rojas-Esponda, James Noraky, Manish Gupta, Ragha Kotikalapudi, Jichuan Chang, Sanja Deur, Dan Graur, Alex Mossin, Erin Farnese, Ricardo Figueira, Alexandre Moufarek, Austin Huang, Patrik Zochbauer, Ben Ingram, Tongzhou Chen, Zelin Wu, Adrià Puigdomènech, Leland Rechis, Da Yu, Sri Gayatri Sundara Padmanabhan, Rui Zhu, Chu ling Ko, Andrea Banino, Samira Daruki, Aarush Selvan, Dhruva Bhaswar, Daniel Hernandez Diaz, Chen Su, Salvatore Scellato, Jennifer Brennan, Woohyun Han, Grace Chung, Priyanka Agrawal, Urvashi Khandelwal, Khe Chai Sim, Morgane Lustman, Sam Ritter, Kelvin Guu, Jiawei Xia, Prateek Jain, Emma Wang, Tyrone Hill, Mirko Rossini, Marija Kostelac, Tautvydas Misiunas, Amit Sabne, Kyuyeun Kim, Ahmet Iscen, Congchao Wang, José Leal, Ashwin Sreevatsa, Utku Evci, Manfred Warmuth, Saket Joshi, Daniel Suo, James Lottes, Garrett Honke, Brendan Jou, Stefani Karp, Jieru Hu, Himanshu Sahni, Adrien Ali Taïga, William Kong, Samrat Ghosh, Renshen Wang, Jay Pavagadhi, Natalie Axelsson, Nikolai Grigorev, Patrick Siegler, Rebecca Lin, Guohui Wang, Emilio Parisotto, Sharath Maddineni, Krishan Subudhi, Eyal Ben-David, Elena Pochernina, Orgad Keller, Thi Avrahami, Zhe Yuan, Pulkit Mehta, Jialu Liu, Sherry Yang, Wendy Kan, Katherine Lee, Tom Funkhouser, Derek Cheng, Hongzhi Shi, Archit Sharma, Joe Kelley, Matan Eyal, Yury Malkov, Corentin Tal-

lec, Yuval Bahat, Shen Yan, Xintian, Wu, David Lindner, Chengda Wu, Avi Caciularu, Xiyang Luo, Rodolphe Jenatton, Tim Zaman, Yingying Bi, Ilya Kornakov, Ganesh Mallya, Daisuke Ikeda, Itay Karo, Anima Singh, Colin Evans, Praneeth Netrapalli, Vincent Nallatamby, Isaac Tian, Yannis Assael, Vikas Raunak, Victor Carbune, Ioana Bica, Lior Madmoni, Dee Cattle, Snchit Grover, Krishna Somandepalli, Sid Lall, Amelio Vázquez-Reina, Riccardo Patana, Jiaqi Mu, Pranav Talluri, Maggie Tran, Rajeev Aggarwal, RJ Skerry-Ryan, Jun Xu, Mike Burrows, Xiaoyue Pan, Edouard Yvinec, Di Lu, Zhiying Zhang, Duc Dung Nguyen, Hairong Mu, Gabriel Barcik, Helen Ran, Lauren Beltrone, Krzysztof Choromanski, Dia Kharrat, Samuel Albanie, Sean Purser-haskell, David Bieber, Carrie Zhang, Jing Wang, Tom Hudson, Zhiyuan Zhang, Han Fu, Johannes Mauerer, Mohammad Hossein Bateni, AJ Maschinot, Bing Wang, Muye Zhu, Arjun Pillai, Tobias Weyand, Shuang Liu, Oscar Akerlund, Fred Bertsch, Vittal Premachandran, Alicia Jin, Vincent Roulet, Peter de Boursac, Shubham Mittal, Ndaba Ndebele, Georgi Karadzhov, Sahra Ghalebikesabi, Ricky Liang, Allen Wu, Yale Cong, Nimesh Ghelani, Sumeet Singh, Bahar Fatemi, Warren, Chen, Charles Kwong, Alexey Kolganov, Steve Li, Richard Song, Chenkai Kuang, Sobhan Miryoosefi, Dale Webster, James Wendt, Arkadiusz Socala, Guolong Su, Artur Mendonça, Abhinav Gupta, Xiaowei Li, Tomy Tsai, Qiong, Hu, Kai Kang, Angie Chen, Sertan Girgin, Yongqin Xian, Andrew Lee, Nolan Ramsden, Leslie Baker, Madeleine Clare Elish, Varvara Krayvanova, Rishabh Joshi, Jiri Simsa, Yao-Yuan Yang, Piotr Ambroszczyk, Dipankar Ghosh, Arjun Kar, Yuan Shangguan, Yumeya Yamamori, Yaroslav Akulov, Andy Brock, Haotian Tang, Siddharth Vashishtha, Rich Munoz, Andreas Steiner, Kalyan Andra, Daniel Eppens, Qixuan Feng, Hayato Kobayashi, Sasha Goldshtein, Mona El Mahdy, Xin Wang, Jilei, Wang, Richard Killam, Tom Kwiatkowski, Kavya Kopparapu, Serena Zhan, Chao Jia, Alexei Bendebury, Sheryl Luo, Adrià Recasens, Timothy Knight, Jing Chen, Mohak Patel, YaGuang Li, Ben Withbroe, Dean Weesner, Kush Bhatia, Jie Ren, Danielle Eisenbud, Ebrahim Songhori, Yanhua Sun, Travis Choma, Tasos Kementsietsidis, Lucas Manning, Brian Roark, Wael Farhan, Jie Feng, Susheel Tatineni, James Cobon-Kerr, Yunjie Li, Lisa Anne Hendricks, Isaac Noble, Chris Breaux, Nate Kushman, Liqian Peng, Fuzhao Xue, Taylor Tobin, Jamie Rogers, Josh Lipschultz, Chris Alberti, Alexey Vlaskin, Mostafa Dehghani, Roshan Sharma, Tris Warkentin, Chen-Yu Lee, Benigno Uria, Da-Cheng Juan, Angad Chandorkar, Hila Sheftel, Ruibo Liu, Elnaz Davoodi, Borja De Balle Pigem, Kedar Dhamdhere, David Ross, Jonathan Hoech, Mahdis Mahdieh, Li Liu, Qiujia Li, Liam McCafferty, Chenxi Liu, Markus Mircea, Yunting Song, Omkar Savant, Alaa Saade, Colin Cherry, Vincent Hellendoorn, Siddharth Goyal, Paul Pucciarelli, David Vilar Torres, Zohar Yahav, Hyo Lee, Lars Lowe Sjoesund, Christo Kirov, Bo Chang, Deepanway Ghoshal, Lu Li, Gilles Baechler, Sébastien Pereira, Tara Sainath, Anudhyan Boral, Dominik Grewe, Afief Halumi, Nguyet Minh Phu, Tianxiao Shen, Marco Tulio Ribeiro, Dhriti Varma, Alex Kaskasoli, Vlad Feinberg, Navneet Potti, Jarrod Kahn, Matheus Wisniewski, Shakir Mohamed, Arnar Mar Hrafnkelsson, Bobak Shahriari, Jean-Baptiste Lespiau, Lisa Patel, Legg Yeung, Tom Paine, Lantao Mei, Alex Ramirez, Rakesh Shivanna, Li Zhong, Josh Woodward, Guilherme Tubone, Samira Khan, Heng Chen, Elizabeth Nielsen, Catalin Ionescu, Utsav Prabhu, Mingcen Gao, Qingze Wang, Sean Augenstein, Neesha Subramaniam, Jason Chang, Fotis Iliopoulos, Jiaming Luo, Myriam Khan, Weicheng Kuo, Denis Teplyashin, Florence Perot, Logan Kilpatrick, Amir Globerson, Hongkun Yu, Anfal Siddiqui, Nick Sukhanov, Arun Kandoor, Umang Gupta, Marco Andreetto, Moran Ambar, Donnie Kim, Paweł Wesołowski, Sarah Perrin, Ben Limonchik, Wei Fan, Jim Stephan, Ian Stewart-Binks, Ryan Kappedal, Tong He, Sarah Cogan, Romina Datta, Tong Zhou, Jiayu Ye, Leandro Kieliger, Ana Ramalho, Kyle Kastner, Fabian Mentzer, Wei-Jen Ko, Arun Suggala, Tianhao Zhou, Shiraz Butt, Hana Strejček, Lior Belenki, Subhashini Venugopalan, Mingyang Ling, Evgenii Eltyshev, Yunxiao Deng, Geza Kovacs, Mukund Raghavachari, Hanjun Dai, Tal Schuster, Steven Schwarcz, Richard Nguyen, Arthur Nguyen, Gavin Buttimore, Shrestha Basu Mallick, Sudeep Gandhe, Seth Benjamin, Michal Jastrzebski, Le Yan, Sugato Basu, Chris Apps, Isabel Edkins, James Allingham, Immanuel Odisho, Tomas Kocisky, Jewel Zhao, Linting Xue, Apoorv Reddy, Chrysovalantis Anastasiou, Aviel Atias, Sam Redmond, Kieran Milan, Nicolas Heess, Herman Schmit, Allan Dafoe, Daniel Andor, Tynan Gangwani, Anca Dragan, Sheng Zhang, Ashyana Kachra, Gang Wu, Siyang Xue, Kevin Aydin, Siqi Liu, Yuxiang Zhou, Mahan Malihi, Austin Wu, Siddharth Gopal, Candice Schumann, Peter Stys, Alek Wang, Mirek Olšák, Dangyi Liu, Christian Schallhart, Yiran Mao, Demetra Brady, Hao Xu, Tomas Mery, Chawin Sitawarin, Siva Velusamy, Tom Cobley, Alex Zhai, Christian Walder, Nitzan Katz, Ganesh Jawahar, Chinmay Kulkarni, Antoine Yang, Adam Paszke, Yinan Wang, Bogdan Damoc, Zalán Borsos, Ray Smith, Jinning Li, Mansi Gupta, Andrei Kapishnikov, Sushant Prakash, Florian Luisier, Rishabh Agarwal, Will Grathwohl, Kuangyuan Chen, Kehang Han, Nikhil Mehta, Andrew Over,

Shekoofeh Azizi, Lei Meng, Niccolò Dal Santo, Kelvin Zheng, Jane Shapiro, Igor Petrovski, Jeffrey Hui, Amin Ghafouri, Jasper Snoek, James Qin, Mandy Jordan, Caitlin Sikora, Jonathan Malmaud, Yuheng Kuang, Aga Świetlik, Ruoxin Sang, Chongyang Shi, Leon Li, Andrew Rosenberg, Shubin Zhao, Andy Crawford, Jan-Thorsten Peter, Yun Lei, Xavier Garcia, Long Le, Todd Wang, Julien Amelot, Dave Orr, Praneeth Kacham, Dana Alon, Gladys Tyen, Abhinav Arora, James Lyon, Alex Kurakin, Mimi Ly, Theo Guidroz, Zhipeng Yan, Rina Panigrahy, Pingmei Xu, Thais Kagohara, Yong Cheng, Eric Noland, Jinhyuk Lee, Jonathan Lee, Cathy Yip, Maria Wang, Efrat Nehoran, Alexander Bykovsky, Zhihao Shan, Ankit Bhagatwala, Chaochao Yan, Jie Tan, Guillermo Garrido, Dan Ethier, Nate Hurley, Grace Vesom, Xu Chen, Siyuan Qiao, Abhishek Nayyar, Julian Walker, Paramjit Sandhu, Mihaela Rosca, Danny Swisher, Mikhail Dektiarev, Josh Dillon, George-Cristian Muraru, Manuel Tragut, Artiom Myaskovsky, David Reid, Marko Velic, Owen Xiao, Jasmine George, Mark Brand, Jing Li, Wenhao Yu, Shane Gu, Xiang Deng, François-Xavier Aubet, Soheil Hassas Yeganeh, Fred Alcober, Celine Smith, Trevor Cohn, Kay McKinney, Michael Tschannen, Ramesh Sampath, Gowoon Cheon, Liangchen Luo, Luyang Liu, Jordi Orbay, Hui Peng, Gabriela Botea, Xiaofan Zhang, Charles Yoon, Cesar Magalhaes, Paweł Stradomski, Ian Mackinnon, Steven Hemingray, Kumaran Venkatesan, Rhys May, Jaeyoun Kim, Alex Druinsky, Jingchen Ye, Zheng Xu, Terry Huang, Jad Al Abdallah, Adil Dostmohamed, Rachana Fellinger, Tsendsuren Munkhdalai, Akanksha Maurya, Peter Garst, Yin Zhang, Maxim Krikun, Simon Bucher, Aditya Srikanth Veerubhotla, Yaxin Liu, Sheng Li, Nishesh Gupta, Jakub Adamek, Hanwen Chen, Bernett Orlando, Aleksandr Zaks, Joost van Amersfoort, Josh Camp, Hui Wan, HyunJeong Choe, Zhichun Wu, Kate Olszewska, Weiren Yu, Archita Vadali, Martin Scholz, Daniel De Freitas, Jason Lin, Amy Hua, Xin Liu, Frank Ding, Yichao Zhou, Boone Severson, Katerina Tsihlas, Samuel Yang, Tammo Spalink, Varun Yerram, Helena Pankov, Rory Blevins, Ben Vargas, Sarthak Jauhari, Matt Miecnikowski, Ming Zhang, Sandeep Kumar, Clement Farabet, Charline Le Lan, Sebastian Flennerhag, Yonatan Bitton, Ada Ma, Arthur Bražinskas, Eli Collins, Niharika Ahuja, Sneha Kudugunta, Anna Bortsova, Minh Giang, Wanzheng Zhu, Ed Chi, Scott Lundberg, Alexey Stern, Subha Puttagunta, Jing Xiong, Xiao Wu, Yash Pande, Amit Jhindal, Daniel Murphy, Jon Clark, Marc Brockschmidt, Maxine Deines, Kevin R. McKee, Dan Bahir, Jiajun Shen, Minh Truong, Daniel McDuff, Andrea Gesmundo, Edouard Rosseel, Bowen Liang, Ken Caluwaerts, Jessica Hamrick, Joseph Kready, Mary Cassin, Rishikesh Ingale, Li Lao, Scott Pollom, Yifan Ding, Wei He, Lizzetth Bellot, Joana Iljazi, Ramya Sree Boppana, Shan Han, Tara Thompson, Amr Khalifa, Anna Bulanova, Blagoj Mitrevski, Bo Pang, Emma Cooney, Tian Shi, Rey Coaguila, Tamar Yakar, Marc'aurelio Ranzato, Nikola Momchev, Chris Rawles, Zachary Charles, Young Maeng, Yuan Zhang, Rishabh Bansal, Xiaokai Zhao, Brian Albert, Yuan Yuan, Sudheendra Vijayanarasimhan, Roy Hirsch, Vinay Ramasesh, Kiran Vodrahalli, Xingyu Wang, Arushi Gupta, DJ Strouse, Jianmo Ni, Roma Patel, Gabe Taubman, Zhouyuan Huo, Dero Gharibian, Marianne Monteiro, Hoi Lam, Shobha Vasudevan, Aditi Chaudhary, Isabela Albuquerque, Kilol Gupta, Sebastian Riedel, Chaitra Hegde, Avraham Ruderman, András György, Marcus Wainwright, Ashwin Chaugule, Burcu Karagol Ayan, Tomer Levinboim, Sam Shleifer, Yogesh Kalley, Vahab Mirrokni, Abhishek Rao, Prabakar Radhakrishnan, Jay Hartford, Jialin Wu, Zhenhai Zhu, Francesco Bertolini, Hao Xiong, Nicolas Serrano, Hamish Tomlinson, Myle Ott, Yifan Chang, Mark Graham, Jian Li, Marco Liang, Xiangzhu Long, Sebastian Borgeaud, Yanif Ahmad, Alex Grills, Diana Mincu, Martin Izzard, Yuan Liu, Jinyu Xie, Louis O'Bryan, Sameera Ponda, Simon Tong, Michelle Liu, Dan Malkin, Khalid Salama, Yuankai Chen, Rohan Anil, Anand Rao, Rigel Swavely, Misha Bilenko, Nina Anderson, Tat Tan, Jing Xie, Xing Wu, Lijun Yu, Oriol Vinyals, Andrey Ryabtsev, Rumen Dangovski, Kate Baumli, Daniel Keysers, Christian Wright, Zoe Ashwood, Betty Chan, Artem Shtefan, Yaohui Guo, Ankur Bapna, Radu Soricut, Steven Pecht, Sabela Ramos, Rui Wang, Jiahao Cai, Trieu Trinh, Paul Barham, Linda Friso, Eli Stickgold, Xiangzhuo Ding, Siamak Shakeri, Diego Ardila, Eleftheria Briakou, Phil Culliton, Adam Raveret, Jingyu Cui, David Saxton, Subhrajit Roy, Javad Azizi, Pengcheng Yin, Lucia Loher, Andrew Bunner, Min Choi, Faruk Ahmed, Eric Li, Yin Li, Shengyang Dai, Michael Elabd, Sriram Ganapathy, Shivani Agrawal, Yiqing Hua, Paige Kunkle, Sujeevan Rajayogam, Arun Ahuja, Arthur Conmy, Alex Vasiloff, Parker Beak, Christopher Yew, Jayaram Mudigonda, Bartek Wydrowski, Jon Blanton, Zhengdong Wang, Yann Dauphin, Zhuo Xu, Martin Polacek, Xi Chen, Hexiang Hu, Pauline Sho, Markus Kunesch, Mehdi Hafezi Manshadi, Eliza Rutherford, Bo Li, Sissie Hsiao, Iain Barr, Alex Tudor, Matija Kecman, Arsha Nagrani, Vladimir Pchelin, Martin Sundermeyer, Aishwarya P S, Abhijit Karmarkar, Yi Gao, Grishma Chole, Olivier Bachem, Isabel Gao, Arturo BC, Matt Dibb, Mauro Verzetti, Felix Hernandez-Campos, Yana Lunts, Matthew Johnson, Julia Di Trapani, Raphael Koster, Idan Brusilovsky, Binbin Xiong, Megha Mohabey, Han Ke, Joe Zou, Tea Sabolić,

Víctor Campos, John Palowitch, Alex Morris, Linhai Qiu, Pranavaraj Ponnuramu, Fangtao Li, Vivek Sharma, Kiranbir Sodhia, Kaan Tekelioglu, Aleksandr Chuklin, Madhavi Yenugula, Erika Gemzer, Theofilos Strinopoulos, Sam El-Husseini, Huiyu Wang, Yan Zhong, Edouard Leurent, Paul Natsev, Weijun Wang, Dre Mahaarachchi, Tao Zhu, Songyou Peng, Sami Alabed, Cheng-Chun Lee, Anthony Brohan, Arthur Szlam, GS Oh, Anton Kovsharov, Jenny Lee, Renee Wong, Megan Barnes, Gregory Thornton, Felix Gimeno, Omer Levy, Martin Sevenich, Melvin Johnson, Jonathan Mallinson, Robert Dadashi, Ziyue Wang, Qingchun Ren, Preethi Lahoti, Arka Dhar, Josh Feldman, Dan Zheng, Thatcher Ulrich, Liviu Panait, Michiel Blokzijl, Cip Baetu, Josip Matak, Jitendra Harlalka, Maulik Shah, Tal Marian, Daniel von Dincklage, Cosmo Du, Ruy Ley-Wild, Bethanie Brownfield, Max Schumacher, Yury Stuken, Shadi Noghabi, Sonal Gupta, Xiaoqi Ren, Eric Malmi, Felix Weissenberger, Blanca Huergo, Maria Bauza, Thomas Lampe, Arthur Douillard, Mojtaba Seyedhosseini, Roy Frostig, Zoubin Ghahramani, Kelvin Nguyen, Kashyap Krishnakumar, Chengxi Ye, Rahul Gupta, Alireza Nazari, Robert Geirhos, Pete Shaw, Ahmed Eleryan, Dima Damen, Jennimaria Palomaki, Ted Xiao, Qiyin Wu, Quan Yuan, Phoenix Meadowlark, Matthew Bilotti, Raymond Lin, Mukund Sridhar, Yannick Schroecker, Da-Woon Chung, Jincheng Luo, Trevor Strohman, Tianlin Liu, Anne Zheng, Jesse Emond, Wei Wang, Andrew Lampinen, Toshiyuki Fukuzawa, Folawiyo Campbell-Ajala, Monica Roy, James Lee-Thorp, Lily Wang, Iftekhar Naim, Tony, Nguy ên, Guy Bensky, Aditya Gupta, Dominika Rogozińska, Justin Fu, Thanumalayan Sankaranarayana Pillai, Petar Veličković, Shahar Drath, Philipp Neubeck, Vaibhav Tulsyan, Arseniy Klimovskiy, Don Metzler, Sage Stevens, Angel Yeh, Junwei Yuan, Tianhe Yu, Kelvin Zhang, Alec Go, Vincent Tsang, Ying Xu, Andy Wan, Isaac Galatzer-Levy, Sam Sobell, Abodunrinwa Toki, Elizabeth Salesky, Wenlei Zhou, Diego Antognini, Sholto Douglas, Shimu Wu, Adam Lelkes, Frank Kim, Paul Cavallaro, Ana Salazar, Yuchi Liu, James Besley, Tiziana Refice, Yiling Jia, Zhang Li, Michal Sokolik, Arvind Kannan, Jon Simon, Jo Chick, Avia Aharon, Meet Gandhi, Mayank Daswani, Keyvan Amiri, Vighnesh Birodkar, Abe Ittycheriah, Peter Grabowski, Oscar Chang, Charles Sutton, Zhixin, Lai, Umesh Telang, Susie Sargsyan, Tao Jiang, Raphael Hoffmann, Nicole Brichtova, Matteo Hessel, Jonathan Halcrow, Sammy Jerome, Geoff Brown, Alex Tomala, Elena Buchatskaya, Dian Yu, Sachit Menon, Pol Moreno, Yuguo Liao, Vicky Zayats, Luming Tang, SQ Mah, Ashish Shenoy, Alex Siegman, Majid Hadian, Okwan Kwon, Tao Tu, Nima Khajehnouri, Ryan Foley, Parisa Haghani, Zhongru Wu, Vaishakh Keshava, Khyatti Gupta, Tony Bruguier, Rui Yao, Danny Karmon, Luisa Zintgraf, Zhicheng Wang, Enrique Piqueras, Junehyuk Jung, Jenny Brennan, Diego Machado, Marissa Giustina, MH Tessler, Kamyu Lee, Qiao Zhang, Joss Moore, Kaspar Daugaard, Alexander Frömmgen, Jennifer Beattie, Fred Zhang, Daniel Kasenberg, Ty Geri, Danfeng Qin, Gaurav Singh Tomar, Tom Ouyang, Tianli Yu, Luowei Zhou, Rajiv Mathews, Andy Davis, Yaoyiran Li, Jai Gupta, Damion Yates, Linda Deng, Elizabeth Kemp, Ga-Young Joung, Sergei Vassilvitskii, Mandy Guo, Pallavi LV, Dave Dopson, Sami Lachgar, Lara McConnaughey, Himadri Choudhury, Dragos Dena, Aaron Cohen, Joshua Ainslie, Sergey Levi, Parthasarathy Gopavarapu, Polina Zablotskaia, Hugo Vallet, Sanaz Bahargam, Xiaodan Tang, Nenad Tomasev, Ethan Dyer, Daniel Balle, Hongrae Lee, William Bono, Jorge Gonzalez Mendez, Vadim Zubov, Shentao Yang, Ivor Rendulic, Yanyan Zheng, Andrew Hogue, Golan Pundak, Ralph Leith, Avishkar Bhoopchand, Michael Han, Mislav Žanić, Tom Schaul, Manolis Delakis, Tejas Iyer, Guanyu Wang, Harman Singh, Abdelrahman Abdelhamed, Tara Thomas, Siddhartha Brahma, Hilal Dib, Naveen Kumar, Wenxuan Zhou, Liang Bai, Pushkar Mishra, Jiao Sun, Valentin Anklin, Roykrong Sukkerd, Lauren Agubuzu, Anton Briukhov, Anmol Gulati, Maximilian Sieb, Fabio Pardo, Sara Nasso, Junquan Chen, Kexin Zhu, Tiberiu Sosea, Alex Goldin, Keith Rush, Spurthi Amba Hombaiah, Andreas Noever, Allan Zhou, Sam Haves, Mary Phuong, Jake Ades, Yi ting Chen, Lin Yang, Joseph Pagadora, Stan Bileschi, Victor Cotruta, Rachel Saputro, Arijit Pramanik, Sean Ammirati, Dan Garrette, Kevin Villela, Tim Blyth, Canfer Akbulut, Neha Jha, Alban Rrustemi, Arissa Wongpanich, Chirag Nagpal, Yonghui Wu, Morgane Rivière, Sergey Kishchenko, Pranesh Srinivasan, Alice Chen, Animesh Sinha, Trang Pham, Bill Jia, Tom Hennigan, Anton Bakalov, Nithya Attaluri, Drew Garmon, Daniel Rodriguez, Dawid Wegner, Wenhao Jia, Evan Senter, Noah Fiedel, Denis Petek, Yuchuan Liu, Cassidy Hardin, Harshal Tushar Lehri, Joao Carreira, Sara Smoot, Marcel Prasetya, Nami Akazawa, Anca Stefanoiu, Chia-Hua Ho, Anelia Angelova, Kate Lin, Min Kim, Charles Chen, Marcin Sieniek, Alice Li, Tongfei Guo, Sorin Baltateanu, Pouya Tafti, Michael Wunder, Nadav Olmert, Divyansh Shukla, Jingwei Shen, Neel Kovelamudi, Balaji Venkatraman, Seth Neel, Romal Thoppilan, Jerome Connor, Frederik Benzing, Axel Stjerngren, Golnaz Ghiasi, Alex Polozov, Joshua Howland, Theophane Weber, Justin Chiu, Ganesh Poomal Girirajan, Andreas Terzis, Pidong Wang, Fangda Li, Yoav Ben Shalom, Dinesh Tewari, Matthew Denton,

Roee Aharoni, Norbert Kalb, Heri Zhao, Junlin Zhang, Angelos Filos, Matthew Rahtz, Lalit Jain, Connie Fan, Vitor Rodrigues, Ruth Wang, Richard Shin, Jacob Austin, Roman Ring, Mariella Sanchez-Vargas, Mehadi Hassen, Ido Kessler, Uri Alon, Gufeng Zhang, Wenhu Chen, Yenai Ma, Xiance Si, Le Hou, Azalia Mirhoseini, Marc Wilson, Geoff Bacon, Becca Roelofs, Lei Shu, Gautam Vasudevan, Jonas Adler, Artur Dwornik, Tayfun Terzi, Matt Lawlor, Harry Askham, Mike Bernico, Xuanyi Dong, Chris Hidey, Kevin Kilgour, Gaël Liu, Surya Bhupatiraju, Luke Leonhard, Siqi Zuo, Partha Talukdar, Qing Wei, Aliaksei Severyn, Vít Listík, Jong Lee, Aditya Tripathi, SK Park, Yossi Matias, Hao Liu, Alex Ruiz, Rajesh Jayaram, Jackson Tolins, Pierre Marcenac, Yiming Wang, Bryan Seybold, Henry Prior, Deepak Sharma, Jack Weber, Mikhail Sirotenko, Yunhsuan Sung, Dayou Du, Ellie Pavlick, Stefan Zinke, Markus Freitag, Max Dylla, Montse Gonzalez Arenas, Natan Potikha, Omer Goldman, Connie Tao, Rachita Chhaparia, Maria Voitovich, Pawan Dogra, Andrija Ražnatović, Zak Tsai, Chong You, Oleaser Johnson, George Tucker, Chenjie Gu, Jae Yoo, Maryam Majzoubi, Valentin Gabeur, Bahram Raad, Rocky Rhodes, Kashyap Kolipaka, Heidi Howard, Geta Sampemane, Benny Li, Chulayuth Asawaroengchai, Duy Nguyen, Chiyuan Zhang, Timothee Cour, Xinxin Yu, Zhao Fu, Joe Jiang, Po-Sen Huang, Gabriela Surita, Iñaki Iturrate, Yael Karov, Michael Collins, Martin Baeuml, Fabian Fuchs, Shilpa Shetty, Swaroop Ramaswamy, Sayna Ebrahimi, Qiuchen Guo, Jeremy Shar, Gabe Barth-Maron, Sravanti Addepalli, Bryan Richter, Chin-Yi Cheng, Eugénie Rives, Fei Zheng, Johannes Griesser, Nishanth Dikkala, Yoel Zeldes, Ilkin Safarli, Dipanjan Das, Himanshu Srivastava, Sadh MNM Khan, Xin Li, Aditya Pandey, Larisa Markeeva, Dan Belov, Qiqi Yan, Mikołaj Rybiński, Tao Chen, Megha Nawhal, Michael Quinn, Vineetha Govindaraj, Sarah York, Reed Roberts, Roopal Garg, Namrata Godbole, Jake Abernethy, Anil Das, Lam Nguyen Thiet, Jonathan Tompson, John Nham, Neera Vats, Ben Caine, Wesley Helmholz, Francesco Pongetti, Yeongil Ko, James An, Clara Huiyi Hu, Yu-Cheng Ling, Julia Pawar, Robert Leland, Keisuke Kinoshita, Waleed Khawaja, Marco Selvi, Eugene Ie, Danila Sinopalnikov, Lev Proleev, Nilesh Tripuraneni, Michele Bevilacqua, Seungji Lee, Clayton Sanford, Dan Suh, Dustin Tran, Jeff Dean, Simon Baumgartner, Jens Heitkaemper, Sagar Gubbi, Kristina Toutanova, Yichong Xu, Chandu Thekkath, Keran Rong, Palak Jain, Annie Xie, Yan Virin, Yang Li, Lubo Litchev, Richard Powell, Tarun Bharti, Adam Kraft, Nan Hua, Marissa Ikonomidis, Ayal Hitron, Sanjiv Kumar, Loic Matthey, Sophie Bridgers, Lauren Lax, Ishaan Malhi, Ondrej Skopek, Ashish Gupta, Jiawei Cao, Mitchelle Rasquinha, Siim Põder, Wojciech Stokowiec, Nicholas Roth, Guowang Li, Michaël Sander, Joshua Kessinger, Vihan Jain, Edward Loper, Wonpyo Park, Michal Yarom, Liqun Cheng, Guru Guruganesh, Kanishka Rao, Yan Li, Catarina Barros, Mikhail Sushkov, Chun-Sung Ferng, Rohin Shah, Ophir Aharoni, Ravin Kumar, Tim McConnell, Peiran Li, Chen Wang, Fernando Pereira, Craig Swanson, Fayaz Jamil, Yan Xiong, Anitha Vijayakumar, Prakash Shroff, Kedar Soparkar, Jindong Gu, Livio Baldini Soares, Eric Wang, Kushal Majmundar, Aurora Wei, Kai Bailey, Nora Kassner, Chizu Kawamoto, Goran Žužić, Victor Gomes, Abhirut Gupta, Michael Guzman, Ishita Dasgupta, Xinyi Bai, Zhufeng Pan, Francesco Piccinno, Hadas Natalie Vogel, Octavio Ponce, Adrian Hutter, Paul Chang, Pan-Pan Jiang, Ionel Gog, Vlad Ionescu, James Manyika, Fabian Pedregosa, Harry Ragan, Zach Behrman, Ryan Mullins, Coline Devin, Aroonalok Pyne, Swapnil Gawde, Martin Chadwick, Yiming Gu, Sasan Tavakkol, Andy Twigg, Naman Goyal, Ndidi Elue, Anna Goldie, Srinivasan Venkatachary, Hongliang Fei, Ziqiang Feng, Marvin Ritter, Isabel Leal, Sudeep Dasari, Pei Sun, Alif Raditya Rochman, Brendan O'Donoghue, Yuchen Liu, Jim Sproch, Kai Chen, Natalie Clay, Slav Petrov, Sailesh Sidhwani, Ioana Mihailescu, Alex Panagopoulos, AJ Piergiovanni, Yunfei Bai, George Powell, Deep Karkhanis, Trevor Yacovone, Petr Mitrichev, Joe Kovac, Dave Uthus, Amir Yazdanbakhsh, David Amos, Steven Zheng, Bing Zhang, Jin Miao, Bhuvana Ramabhadran, Soroush Radpour, Shantanu Thakoor, Josh Newlan, Oran Lang, Orion Jankowski, Shikhar Bharadwaj, Jean-Michel Sarr, Shereen Ashraf, Sneha Mondal, Jun Yan, Ankit Singh Rawat, Sarmishta Velury, Greg Kochanski, Tom Eccles, Franz Och, Abhanshu Sharma, Ethan Mahintorabi, Alex Gurney, Carrie Muir, Vered Cohen, Saksham Thakur, Adam Bloniarz, Asier Mujika, Alexander Pritzel, Paul Caron, Altaf Rahman, Fiona Lang, Yasumasa Onoe, Petar Sirkovic, Jay Hoover, Ying Jian, Pablo Duque, Arun Narayanan, David Soergel, Alex Haig, Loren Maggiore, Shyamal Buch, Josef Dean, Ilya Figotin, Igor Karpov, Shaleen Gupta, Denny Zhou, Muhuan Huang, Ashwin Vaswani, Christopher Semturs, Kaushik Shivakumar, Yu Watanabe, Vinodh Kumar Rajendran, Eva Lu, Yanhan Hou, Wenting Ye, Shikhar Vashishth, Nana Nti, Vytenis Sakenas, Darren Ni, Doug DeCarlo, Michael Bendersky, Sumit Bagri, Nacho Cano, Elijah Peake, Simon Tokumine, Varun Godbole, Carlos Guía, Tanya Lando, Vittorio Selo, Seher Ellis, Danny Tarlow, Daniel Gillick, Alessandro Epasto, Siddhartha Reddy Jonnalagadda, Meng Wei, Meiyan Xie, Ankur Taly, Michela Paganini, Mukund Sundararajan, Daniel Toyama, Ting Yu, Dessie Petrova, Aneesh

Pappu, Rohan Agrawal, Senaka Buthpitiya, Justin Frye, Thomas Buschmann, Remi Crocker, Marco Tagliasacchi, Mengchao Wang, Da Huang, Sagi Perel, Brian Wieder, Hideto Kazawa, Weiyue Wang, Jeremy Cole, Himanshu Gupta, Ben Golan, Seojin Bang, Nitish Kulkarni, Ken Franko, Casper Liu, Doug Reid, Sid Dalmia, Jay Whang, Kevin Cen, Prasha Sundaram, Johan Ferret, Berivan Isik, Lucian Ionita, Guan Sun, Anna Shekhawat, Muqthar Mohammad, Philip Pham, Ronny Huang, Karthik Raman, Xingyi Zhou, Ross Mcilroy, Austin Myers, Sheng Peng, Jacob Scott, Paul Covington, Sofia Erell, Pratik Joshi, João Gabriel Oliveira, Natasha Noy, Tajwar Nasir, Jake Walker, Vera Axelrod, Tim Dozat, Pu Han, Chun-Te Chu, Eugene Weinstein, Anand Shukla, Shreyas Chandrakaladharan, Petra Poklukar, Bonnie Li, Ye Jin, Prem Eruvbetine, Steven Hansen, Avigail Dabush, Alon Jacovi, Samrat Phatale, Chen Zhu, Steven Baker, Mo Shomrat, Yang Xiao, Jean Pouget-Abadie, Mingyang Zhang, Fanny Wei, Yang Song, Helen King, Yiling Huang, Yun Zhu, Ruoxi Sun, Juliana Vicente Franco, Chu-Cheng Lin, Sho Arora, Hui, Li, Vivian Xia, Luke Vilnis, Mariano Schain, Kaiz Alarakyia, Laurel Prince, Aaron Phillips, Caleb Habtegebriel, Luyao Xu, Huan Gui, Santiago Ontanon, Lora Aroyo, Karan Gill, Peggy Lu, Yash Katariya, Dhruv Madeka, Shankar Krishnan, Shubha Srinivas Raghvendra, James Freedman, Yi Tay, Gaurav Menghani, Peter Choy, Nishita Shetty, Dan Abolafia, Doron Kukliansky, Edward Chou, Jared Lichtarge, Ken Burke, Ben Coleman, Dee Guo, Larry Jin, Indro Bhattacharya, Victoria Langston, Yiming Li, Suyog Kotecha, Alex Yakubovich, Xinyun Chen, Petre Petrov, Tolly Powell, Yanzhang He, Corbin Quick, Kanav Garg, Dawsen Hwang, Yang Lu, Srinadh Bhojanapalli, Kristian Kjems, Ramin Mehran, Aaron Archer, Hado van Hasselt, Ashwin Balakrishna, JK Kearns, Meiqi Guo, Jason Riesa, Mikita Sazanovich, Xu Gao, Chris Sauer, Chengrun Yang, XiangHai Sheng, Thomas Jimma, Wouter Van Gansbeke, Vitaly Nikolaev, Wei Wei, Katie Millican, Ruizhe Zhao, Justin Snyder, Levent Bolelli, Maura O'Brien, Shawn Xu, Fei Xia, Wentao Yuan, Arvind Neelakantan, David Barker, Sachin Yadav, Hannah Kirkwood, Farooq Ahmad, Joel Wee, Jordan Grimstad, Boyu Wang, Matthew Wiethoff, Shane Settle, Miaosen Wang, Charles Blundell, Jingjing Chen, Chris Duvarney, Grace Hu, Olaf Ronneberger, Alex Lee, Yuanzhen Li, Abhishek Chakladar, Alena Butryna, Georgios Evangelopoulos, Guillaume Desjardins, Jonni Kanerva, Henry Wang, Averi Nowak, Nick Li, Alyssa Loo, Art Khurshudov, Laurent El Shafey, Nagabhushan Baddi, Karel Lenc, Yasaman Razeghi, Tom Lieber, Amer Sinha, Xiao Ma, Yao Su, James Huang, Asahi Ushio, Hanna Klimczak-Plucińska, Kareem Mohamed, JD Chen, Simon Osindero, Stav Ginzburg, Lampros Lamprou, Vasilisa Bashlovkina, Duc-Hieu Tran, Ali Khodaei, Ankit Anand, Yixian Di, Ramy Eskander, Manish Reddy Vuyyuru, Jasmine Liu, Aishwarya Kamath, Roman Goldenberg, Mathias Bellaiche, Juliette Pluto, Bill Rosgen, Hassan Mansoor, William Wong, Suhas Ganesh, Eric Bailey, Scott Baird, Dan Deutsch, Jinoo Baek, Xuhui Jia, Chansoo Lee, Abe Friesen, Nathaniel Braun, Kate Lee, Amayika Panda, Steven M. Hernandez, Duncan Williams, Jianqiao Liu, Ethan Liang, Arnaud Autef, Emily Pitler, Deepali Jain, Phoebe Kirk, Oskar Bunyan, Jaume Sanchez Elias, Tongxin Yin, Machel Reid, Aedan Pope, Nikita Putikhin, Bidisha Samanta, Sergio Guadarrama, Dahun Kim, Simon Rowe, Marcella Valentine, Geng Yan, Alex Salcianu, David Silver, Gan Song, Richa Singh, Shuai Ye, Hannah DeBalsi, Majd Al Merey, Eran Ofek, Albert Webson, Shibl Mourad, Ashwin Kakarla, Silvio Lattanzi, Nick Roy, Evgeny Sluzhaev, Christina Butterfield, Alessio Tonioni, Nathan Waters, Sudhindra Kopalle, Jason Chase, James Cohan, Girish Ramchandra Rao, Robert Berry, Michael Voznesensky, Shuguang Hu, Kristen Chiafullo, Sharat Chikkerur, George Scrivener, Ivy Zheng, Jeremy Wiesner, Wolfgang Macherey, Timothy Lillicrap, Fei Liu, Brian Walker, David Welling, Elinor Davies, Yangsibo Huang, Lijie Ren, Nir Shabat, Alessandro Agostini, Mariko Iinuma, Dustin Zelle, Rohit Sathyanarayana, Andrea D'olimpio, Morgan Redshaw, Matt Ginsberg, Ashwin Murthy, Mark Geller, Tatiana Matejovicova, Ayan Chakrabarti, Ryan Julian, Christine Chan, Qiong Hu, Daniel Jarrett, Manu Agarwal, Jeshwanth Challagundla, Tao Li, Sandeep Tata, Wen Ding, Maya Meng, Zhuyun Dai, Giulia Vezzani, Shefali Garg, Jannis Bulian, Mary Jasarevic, Honglong Cai, Harish Rajamani, Adam Santoro, Florian Hartmann, Chen Liang, Bartek Perz, Apoorv Jindal, Fan Bu, Sungyong Seo, Ryan Poplin, Adrian Goedeckemeyer, Badih Ghazi, Nikhil Khadke, Leon Liu, Kevin Mather, Mingda Zhang, Ali Shah, Alex Chen, Jinliang Wei, Keshav Shivam, Yuan Cao, Donghyun Cho, Angelo Scorza Scarpati, Michael Moffitt, Clara Barbu, Ivan Jurin, Ming-Wei Chang, Hongbin Liu, Hao Zheng, Shachi Dave, Christine Kaeser-Chen, Xiaobin Yu, Alvin Abdagic, Lucas Gonzalez, Yanping Huang, Peilin Zhong, Cordelia Schmid, Bryce Petrini, Alex Wertheim, Jifan Zhu, Hoang Nguyen, Kaiyang Ji, Yanqi Zhou, Tao Zhou, Fangxiaoyu Feng, Regev Cohen, David Rim, Shubham Milind Phal, Petko Georgiev, Ariel Brand, Yue Ma, Wei Li, Somit Gupta, Chao Wang, Pavel Dubov, Jean Tarbouriech, Kingshuk Majumder, Huijian Li, Norman Rink, Apurv Suman, Yang Guo, Yinghao Sun, Arun Nair, Xiaowei Xu, Mo-

hamed Elhawaty, Rodrigo Cabrera, Guangxing Han, Julian Eisenschlos, Junwen Bai, Yuqi Li, Yamini Bansal, Thibault Sellam, Mina Khan, Hung Nguyen, Justin Mao-Jones, Nikos Parotsidis, Jake Marcus, Cindy Fan, Roland Zimmermann, Yony Kochinski, Laura Graesser, Feryal Behbahani, Alvaro Caceres, Michael Riley, Patrick Kane, Sandra Lefdal, Rob Willoughby, Paul Vicol, Lun Wang, Shujian Zhang, Ashleah Gill, Yu Liang, Gautam Prasad, Soroosh Mariooryad, Mehran Kazemi, Zifeng Wang, Kritika Muralidharan, Paul Voigtlaender, Jeffrey Zhao, Huanjie Zhou, Nina D'Souza, Aditi Mavalankar, Séb Arnold, Nick Young, Obaid Sarvana, Chace Lee, Milad Nasr, Tingting Zou, Seokhwan Kim, Lukas Haas, Kaushal Patel, Neslihan Bulut, David Parkinson, Courtney Biles, Dmitry Kalashnikov, Chi Ming To, Aviral Kumar, Jessica Austin, Alex Greve, Lei Zhang, Megha Goel, Yeqing Li, Sergey Yaroshenko, Max Chang, Abhishek Jindal, Geoff Clark, Hagai Taitelbaum, Dale Johnson, Ofir Roval, Jeongwoo Ko, Anhad Mohananey, Christian Schuler, Shenil Dodhia, Ruichao Li, Kazuki Osawa, Claire Cui, Peng Xu, Rushin Shah, Tao Huang, Ela Gruzewska, Nathan Clement, Mudit Verma, Olcan Sercinoglu, Hai Qian, Viral Shah, Masa Yamaguchi, Abhinit Modi, Takahiro Kosakai, Thomas Strohmann, Junhao Zeng, Beliz Gunel, Jun Qian, Austin Tarango, Krzysztof Jastrzębski, Robert David, Jyn Shan, Parker Schuh, Kunal Lad, Willi Gierke, Mukundan Madhavan, Xinyi Chen, Mark Kurzeja, Rebeca Santamaria-Fernandez, Dawn Chen, Alexandra Cordell, Yuri Chervonyi, Frankie Garcia, Nithish Kannen, Vincent Perot, Nan Ding, Shlomi Cohen-Ganor, Victor Lavrenko, Junru Wu, Georgie Evans, Cicero Nogueira dos Santos, Madhavi Sewak, Ashley Brown, Andrew Hard, Joan Puigcerver, Zeyu Zheng, Yizhong Liang, Evgeny Gladchenko, Reeve Ingle, Uri First, Pierre Sermanet, Charlotte Magister, Mihajlo Velimirović, Sashank Reddi, Susanna Ricco, Eirikur Agustsson, Hartwig Adam, Nir Levine, David Gaddy, Dan Holtmann-Rice, Xuanhui Wang, Ashutosh Sathe, Abhijit Guha Roy, Blaž Bratanič, Alen Carin, Harsh Mehta, Silvano Bonacina, Nicola De Cao, Mara Finkelstein, Verena Rieser, Xinyi Wu, Florent Altché, Dylan Scandinaro, Li Li, Nino Vieillard, Nikhil Sethi, Garrett Tanzer, Zhi Xing, Shibo Wang, Parul Bhatia, Gui Citovsky, Thomas Anthony, Sharon Lin, Tianze Shi, Shoshana Jakobovits, Gena Gibson, Raj Apte, Lisa Lee, Mingqing Chen, Arunkumar Byravan, Petros Maniatis, Kellie Webster, Andrew Dai, Pu-Chin Chen, Jiaqi Pan, Asya Fadeeva, Zach Gleicher, Thang Luong, and Niket Kumar Bhumihar. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li, Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaze Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Joya Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, Ke Shen, Ke Wang, Keyu Pan, Kun Zhang, Kunchang Li, Lanxin Li, Lei Li, Lei Shi, Li Han, Liang Xiang, Liangqiang Chen, Lin Chen, Lin Li, Lin Yan, Liying Chi, Longxiang Liu, Mengfei Du, Mingxuan Wang, Ningxin Pan, Peibin Chen, Pengfei Chen, Pengfei Wu, Qingqing Yuan, Qingyao Shuai, Qiuyan Tao, Renjie Zheng, Renrui Zhang, Ru Zhang, Rui Wang, Rui Yang, Rui Zhao, Shaoqiang Xu, Shihao Liang, Shipeng Yan, Shu Zhong, Shuaishuai Cao, Shuangzhi Wu, Shufan Liu, Shuhan Chang, Songhua Cai, Tenglong Ao, Tianhao Yang, Tingting Zhang, Wanjun Zhong, Wei Jia, Wei Weng, Weihao Yu, Wenhao Huang, Wenjia Zhu, Wenli Yang, Wenzhi Wang, Xiang Long, XiangRui Yin, Xiao Li, Xiaolei Zhu, Xiaoying Jia, Xijin Zhang, Xin Liu, Xinchen Zhang, Xinyu Yang, Xiongcai Luo, Xiuli Chen, Xuantong Zhong, Xuefeng Xiao, Xujing Li, Yan Wu, Yawei Wen, Yifan Du, Yihao Zhang, Yining Ye, Yonghui Wu, Yu Liu, Yu Yue, Yufeng Zhou, Yufeng Yuan, Yuhang Xu, Yuhong Yang, Yun Zhang, Yunhao Fang, Yuntao Li, Yurui Ren, Yuwen Xiong, Zehua Hong, Zehua Wang, Zewei Sun, Zeyu Wang, Zhao Cai, Zhaoyue Zha, Zhecheng An, Zhehui Zhao, Zhengzhuo Xu, Zhipeng Chen, Zhiyong Wu, Zhuofan Zheng, Zihao Wang, Zilong Huang, Ziyu Zhu, and Zuquan Song. Seed1.5-vl technical report, 2025a. URL https://arxiv.org/abs/2505.07062.

Yuhan Guo, Cong Guo, Aiwen Sun, Hongliang He, Xinyu Yang, Yue Lu, Yingji Zhang, Xuntao Guo, Dong Zhang, Jianzhuang Liu, et al. Web-cogreasoner: Towards knowledge-induced cognitive reasoning for web agents. *arXiv preprint arXiv:2508.01858*, 2025b.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14281–14290, 2024.

Marko Jurmu, Sebastian Boring, and Jukka Riekki. Screenspot: Multidimensional resource discovery for distributed applications in smart spaces. In *Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 1–9, 2008.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025.

Shuquan Lian, Yuhang Wu, Jia Ma, Zihan Song, Bingqi Chen, Xiawu Zheng, and Hui Li. Ui-agile: Advancing gui agents with effective reinforcement learning and precise inference-time grounding. *arXiv preprint arXiv:2507.22025*, 2025.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Qinchen Wu, Mingyi Yan, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Videogui: A benchmark for gui automation from instructional videos. *arXiv preprint arXiv:2406.10227*, 4, 2024.

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

Yuhang Liu, Pengxiang Li, Zishu Wei, Congkai Xie, Xueyu Hu, Xinchen Xu, Shengyu Zhang, Xiaotian Han, Hongxia Yang, and Fei Wu. Infiguiagent: A multimodal generalist gui agent with native reasoning and reflection. *arXiv preprint arXiv:2501.04575*, 2025.

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

OpenAI. Gpt-5 system card, Aug 2025a. URL https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: 2025-09-25.

OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025b. URL https://openai.com/index/o3-o4-mini-system-card/. Accessed: 2025-09-25.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms, 2024. URL https://arxiv.org/abs/2312.05934.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL https://arxiv.org/abs/2501.12326.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, et al. Os-genesis: Automating gui agent trajectory construction via reverse task synthesis. *arXiv preprint arXiv:2412.19723*, 2024.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Xuehui Wang, Zhenyu Wu, JingJing Xie, Zichen Ding, Bowen Yang, Zehao Li, Zhaoyang Liu, Qingyun Li, Xuan Dong, Zhe Chen, et al. Mmbench-gui: Hierarchical multi-platform evaluation framework for gui agents. *arXiv preprint arXiv:2507.19478*, 2025.

Yuyang Wanyan, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Jiabo Ye, Yutong Kou, Ming Yan, Fei Huang, Xiaoshan Yang, et al. Look before you leap: A gui-critic-r1 model for pre-operative error diagnosis in gui automation. *arXiv preprint arXiv:2506.04614*, 2025.

Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. Os-copilot: Towards generalist computer agents with self-improvement, 2024a. URL https://arxiv.org/abs/2402.07456.

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

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, et al. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*, 2025a.

Tianbao Xie, Mengqi Yuan, Danyang Zhang, Xinzhuang Xiong, Zhennan Shen, Zilong Zhou, Xinyuan Wang, Yanxu Chen, Jiaqi Deng, Junda Chen, Bowen Wang, Haoyuan Wu, Jixuan Chen, Junli Wang, Dunjie Lu, Hao Hu, and Tao Yu. Introducing osworld-verified. *xlang.ai*, July 2025b. URL https://xlang.ai/blog/osworld-verified.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Jing Hua Toh, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2025c.

Yifan Xu, Xiao Liu, Xueqiao Sun, Siyi Cheng, Hao Yu, Hanyu Lai, Shudan Zhang, Dan Zhang, Jie Tang, and Yuxiao Dong. Androidlab: Training and systematic benchmarking of android autonomous agents, 2024a. URL https://arxiv.org/abs/2410.24024.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024b.

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

A Appendix

A.1 Question Generation Prompt Template for Interface Perception

Prompt for widget function understanding.

Widget Function Prompt

System Prompt:

[Role]
You will be provided with a single screenshot of a system interface (desktop app, web UI, or mobile app). Generate exactly one challenging GUI reasoning question about that screenshot that requires inspecting the image to answer.
[Knowledge Scope of the question]
Ask about the intended function of a specific UI widget (button, toggle, slider, icon, etc.) inferred from the widget's iconography and surrounding context. Avoid universally trivial icons unless combined with contextual clues.
[Generation Guidelines]

1. Question length: one concise sentence only. No hints, no steps, no extra context.
2. Position-only references: Do NOT use any visible text, icon names, or labels from the screenshot. Refer ONLY by position or coordinates (examples: "top-right corner", "third from left in the top toolbar", "second row, third column", "left sidebar, bottom icon", or "`<x, y>`" with origin top-left). The question must be unsolvable without the screenshot.
3. Question types and options:
   - If `multiple_choice`: produce exactly 4 options. The first option MUST be the correct answer.
   - If `yes_or_no`: produce exactly 3 options: {"yes", "no", "unknown"} and the correct one must be first.
   - If the correct answer is genuinely not deducible from the screenshot or you cannot answer the correct answer, then use:
     - multiple_choice: first option = "none of the other options are correct."
     - yes_or_no: first option = "unknown"
4. Option style: Options must describe actions or effects (not icon shapes). Keep options parallel in length and style ($\approx$ 6–16 words).
5. Distractors: The 3 incorrect options must be plausible and similar to the correct one.
6. Contextual reasoning: Prefer questions requiring reasoning across UI elements (e.g., highlighted rows, active tab, enabled/disabled states, adjacent panels).
7. Based on the provided screenshot, identify which application is currently being used and include this information in your output JSON under the field `app_type`.

[Output JSON schema — return exactly this JSON object (no extra text)]

```
{
  "question_type": "multiple_choice" or "yes_or_no",
  "question_text": "<one concise sentence using only positions>",
  "option_text": ["<first option correct>", "<distractor 1>",
  ↪  "<distractor 2>", "<distractor 3>"],
  "app_type": "<application type of the current screenshot>",
  "os_type": "Linux" | "Windows" | "Android" | "MacOS" | "IOS" | "Web"
}
```

[Example Output]

```
{
  "question_type": "yes_or_no",
  "question_text": "While cell B5 in the 'First Name' column shows
  ↪  'Walter' in the formula bar and the checkmark and 'X' icons are
  ↪  visible beside it, will clicking the 'X' icon clear formatting in
  ↪  the selected cell",
  "option_text": ["yes","no","unknown"],
  "app_type": "Excel",
  "os_type": "Linux"
}
{
  "question_type": "multiple_choice",
  "question_text": "Which of the following statement is correct
  ↪  according to the screenshots?",
  "option_text": [
    "The camera is not currently connected to WiFi",
    "The camera can not be controlled remotely from the phone",
    "Pressing the 'phone' mode icon in the top bar can lead to turning
    ↪  on the phone's airplane mode",
    "Pressing the 'clone' mode icon in the top bar can lead to signing
    ↪  out of the cloud gallery"
  ],
  "app_type": "Excel",
  "os_type": "Linux"
}
```

Prompt for layout semantics understanding.

Layout Semantics Prompt

System Prompt:

[Role]
You will be provided with a single screenshot of a system interface (desktop app, web UI, or mobile app). Generate exactly one challenging GUI reasoning question about that screenshot that requires inspecting the image to answer.
[Knowledge Scope of the question]
The questions should assess whether the model understands positional and grouping relationships between UI elements, inferring their roles from placement and hierarchy.
[Generation Guidelines]

1. Question length: one concise sentence only. No hints, no steps, no extra context.
2. Position-only references: Do NOT use any visible text, icon names, or labels from the screenshot. Refer ONLY by position or coordinates (examples: "top-right corner", "third from left in the top toolbar", "second row, third column", "left sidebar, bottom icon", or "<x,y>" with origin top-left). The question must be unsolvable without the screenshot.
3. Question types and options:
   - If `multiple_choice`: produce exactly 4 options. The first option MUST be the correct answer.
   - If `yes_or_no`: produce exactly 3 options: {"yes", "no", "unknown"} and the correct one must be first.
   - If the correct answer is genuinely not deducible from the screenshot or you cannot answer the correct answer, then use:
     - multiple_choice: first option = "none of the other options are correct."
     - yes_or_no: first option = "unknown"
4. Option style: Options must describe actions or effects (not icon shapes). Keep options parallel in length and style (≈ 6–16 words).
5. Distractors: The 3 incorrect options must be plausible and similar to the correct one.

6. Contextual reasoning: Prefer questions requiring reasoning across UI elements (e.g., highlighted rows, active tab, enabled/disabled states, adjacent panels).
7. Based on the provided screenshot, identify which application is currently being used and include this information in your output JSON under the field `app_type`.

[Output JSON schema — return exactly this JSON object (no extra text)]

```
{
  "question_type": "multiple_choice" or "yes_or_no",
  "question_text": "<one concise sentence using only positions>",
  "option_text": ["<first option correct>", "<distractor 1>",
  ↪  "<distractor 2>", "<distractor 3>"],
  "app_type": "<application type of the current screenshot>",
  "os_type": "Linux" | "Windows" | "Android" | "MacOS" | "IOS" | "Web"
}
```

[Example Output]

```
{
  "question_type": "multiple_choice",
  "question_text": "What is likely to be the departure city?",
  "option_text": ["Beijing", "Shanghai", "Guangzhou", "None of the
  ↪  other options."],
  "app_type": "website",
  "os_type": "Windows"
}

{
  "question_type": "yes_or_no",
  "question_text": "Is the folder in the second row under the
  ↪  'Documents' folder?",
  "option_text": ["yes", "no", "unknown"],
  "app_type": "Thunderbird",
  "os_type": "Windows"
}

{
  "question_type": "multiple_choice",
  "question_text": "Who sends this email. Please answer the email
  ↪  address.",
  "option_text": ["li@gmail.com", "zhang@gmail.com", "wang@gmail.com",
  ↪  "None of the other options."],
  "app_type": "Email",
  "os_type": "Windows"
}
```

Prompt for state information understanding.

State Information Prompt

System Prompt:

[Role]
You will be provided with a single screenshot of a system interface (desktop app, web UI, or mobile app). Generate exactly one challenging GUI reasoning question about that screenshot that requires inspecting the image to answer.
[Knowledge Scope of the question]
Ask about the current state information of the system, such as whether a control is enabled/disabled, a process is in-progress/completed, a request is pending, or the system is online/offline. Prefer reasoning that requires subtle visual cues or multi-element context.
[Generation Guidelines]

1. Question length: one concise sentence only. No hints, no steps, no extra context.

2. Position-only references: Do NOT use any visible text, icon names, or labels from the screenshot. Refer ONLY by position or coordinates (examples: "top-right corner", "third from left in the top toolbar", "second row, third column", "left sidebar, bottom icon", or "<x,y>" with origin top-left). The question must be unsolvable without the screenshot.
3. Question types and options:
   - If `multiple_choice`: produce exactly 4 options. The first option MUST be the correct answer.
   - If `yes_or_no`: produce exactly 3 options: {"yes", "no", "unknown"} and the correct one must be first.
   - If the correct answer is genuinely not deducible from the screenshot or you cannot answer the correct answer, then use:
     - multiple_choice: first option = "none of the other options are correct."
     - yes_or_no: first option = "unknown"
4. Option style: Options must describe actions or effects (not icon shapes). Keep options parallel in length and style (≈ 6–16 words).
5. Distractors: The 3 incorrect options must be plausible and similar to the correct one.
6. Contextual reasoning: Prefer questions requiring reasoning across UI elements (e.g., highlighted rows, active tab, enabled/disabled states, adjacent panels).
7. Based on the provided screenshot, identify which application is currently being used and include this information in your output JSON under the field `app_type`.

[Output JSON schema — return exactly this JSON object (no extra text)]

```
{
  "question_type": "multiple_choice" or "yes_or_no",
  "question_text": "<one concise sentence using only positions>",
  "option_text": ["<first option correct>", "<distractor 1>",
  ↪  "<distractor 2>", "<distractor 3>"],
  "app_type": "<application type of the current screenshot>",
  "os_type": "Linux" | "Windows" | "Android" | "MacOS" | "IOS" | "Web"
}
```

[Example Output]

```
{
  "question_type": "multiple_choice",
  "question_text": "The button in the lower toolbar is active, but the
  ↪  button next to it is greyed out. Which condition is most likely
  ↪  not met yet?",
  "option_text": [
    "All required fields are filled",
    "Network connection is active",
    "File format is supported",
    "None of the other options"
  ],
  "app_type": "Form Editor",
  "os_type": "Web"
}

{
  "question_type": "multiple_choice",
  "question_text": "How can the user enable more controls over the
  ↪  alignment of objects?",
  "option_text": [
    "Select more than one object",
    "Double click the alignment button",
    "None of the other options",
    "User is logged in"
  ],
  "app_type": "Graphics Editor",
  "os_type": "Windows"
```

```
}

{
  "question_type": "yes_or_no",
  "question_text": "Will the option in the toolbar become available
  ↪  immediately after selecting a file?",
  "option_text": ["yes","no","unknown"],
  "app_type": "Document Editor",
  "os_type": "MacOS"
}

{
  "question_type": "yes_or_no",
  "question_text": "Is the movie export function currently available?",
  "option_text": ["no","yes","unknown"],
  "app_type": "Video Editor",
  "os_type": "Linux"
}
```

A.2 Plan Generation Prompt Template for OSWorld Tasks.

User Instruction Prompt

User Prompt:

Analyze the given GUI task and break it down into essential, actionable steps. You will receive: - a task instruction: `{task_instruction}` - the app where the task occurs: `{app_name}` - the initial screenshot image
Your goal is to output a Python list of clear, concise steps in logical order to complete the task within the app. Each step should represent a key state, action, or milestone. Use simple, direct language. Avoid ambiguity or unnecessary complexity.
Output format:

- A valid Python list of strings, e.g.:
  `["First step.", "Second step.", "Third step."]`
- Each string must use double quotes ("), and the output must be directly parsable using `eval()` or `ast.literal_eval()`.
- Output only the list. No explanation, no extra text.

Constraints:

- Ensure each step is actionable and unambiguous,
- Ensure each step is necessary for task completion,
- Ensure each step is easy to follow by a user.

A.3 Evaluation Message Prompt Template

A.3.1 Interface Perception.

All evaluation questions in this knowledge category use the same prompt template as shown below.

GUI Agent Inference Prompt

System
You are a Graphical User Interface (GUI) agent. You will be given a screenshot, a question, and corresponding options. You need to choose one option as your answer.

User
`{question_images}`

{question_texts}
{question_options}

Response Rules

If question_type == 'yes_or_no':
Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<yes/no/unknown>"
}
```

If question_type == 'multiple_choice':
Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D>"
}
```

Interaction Prediction.

GUI Agent Task-Solving Prompt

System
You are a Graphical User Interface (GUI) agent. You will be given a task instruction, a screenshot, several GUI operations, and four options. Your goal is to select the best option that could solve the task.
{question_images}

User
{question_text}
Which of the above options are correct according to the screenshots? Think step by step. You must respond strictly in JSON format following this schema.

Response Schema

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D>"
}
```

A.3.2 Interaction Prediction

ActionEffect

GUI Agent Next-State Selection Prompt

System
You are a Graphical User Interface (GUI) agent. You will be given a screenshot, action descriptions, and multiple options, each containing an image. After performing one action on the screenshot, your goal is to select the option that correctly corresponds to the resulting screenshot after performing the action. Below is a short description of the action space:

```
if platform == Desktop:
        Action Space
        - click(point='x1 y1'): left click a position on the screen.
        - left_double(point='x1 y1'): left double click a position on
        ↪  the screen.
        - right_single(point='x1 y1'): right single click a position on
        ↪  the screen.
        - drag(start_point='x1 y1', end_point='x2 y2'): drag the mouse
        ↪  from one position to another.
```

```
        - hotkey(key='ctrl c'): keyboard shortcut, split keys with
        ↪  spaces
        - type(content='xxx'): type an answer, use escape characters
        ↪  (', ", \n) when needed. Add \n at the end if it is the
        ↪  final submission.
        - scroll(point='x1 y1', direction='down or up or right or
        ↪  left'): scroll to see more content

if platform == Mobile:
        Action Space
        - click(point='x1 y1')
        - long_press(point='x1 y1')
        - type(content='') #If you want to submit your input, use "\\n"
        ↪  at the end of `content`.
        - scroll(point='x1 y1', direction='down or up or right or
        ↪  left'): scroll to see more content
```

The size of the image is `{w}x{h}.\n`

User

`{question_image}`
Above is the current screenshot.
After I perform the described action `'action_type(action_parameter)'` (as drawn in the initial screenshot), which of the following options correctly corresponds to the resulting screenshot?
A. `{option_image_A}`
B. `{option_image_B}`
C. `{option_image_C}`
D. `{option_image_D}`

Response Schema

Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D>"
}
```

ActionPrediction - Parameter

GUI Agent Action-Parameter Selection Prompt

System

You are a Graphical User Interface (GUI) agent. You will be given two consecutive screenshots of the GUI, action descriptions, and multiple options. Your goal is to select which action was performed to transition from the first screenshot to the second. If the description specifies an action type, select the correct parameter value for the given action.

```
if platform == Desktop:
        Action Space
        - click(point='x1 y1'): left click a position on the screen.
        - left_double(point='x1 y1'): left double click a position on
        ↪  the screen.
        - right_single(point='x1 y1'): right single click a position on
        ↪  the screen.
        - drag(start_point='x1 y1', end_point='x2 y2'): drag the mouse
        ↪  from one position to another.
        - hotkey(key='ctrl c'): keyboard shortcut, split keys with
        ↪  spaces
        - type(content='xxx'): type an answer, use escape characters
        ↪  (', ", \n) when needed. Add \n at the end if it is the
        ↪  final submission.
        - scroll(point='x1 y1', direction='down or up or right or
        ↪  left'): scroll to see more content
```

```
if platform == Mobile:
        Action Space
        - click(point='x1 y1')
        - long_press(point='x1 y1')
        - type(content='') #If you want to submit your input, use "\\n"
        ↪  at the end of `content`.
        - scroll(point='x1 y1', direction='down or up or right or
        ↪  left'): scroll to see more content
```

The size of the image is `{w}x{h}.\n`
`{question_images}`

User

Above are two consecutive screenshots. Your task is to select the option containing the right parameter value of the given action ' `{action_type}` ' to transition from the first to the second screenshot.
As is drawn in the first screenshot. Which of the above options are correct according to the screenshots?
A. `{option_text}`
B. `{option_text}`
C. `{option_text}`
D. `{option_text}`

Response Schema

Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D>"
}
```

ActionPrediction - Type

GUI Agent Action Identification Prompt

System

You are a Graphical User Interface (GUI) agent. You will be given two consecutive screenshots of the GUI, action descriptions, and multiple options. Your goal is to select which action was performed to transition from the first screenshot to the second. If the description specifies an action type, select the correct parameter value for the given action.

```
if platform == Desktop:
        Action Space
        - click(point='x1 y1'): left click a position on the screen.
        - left_double(point='x1 y1'): left double click a position on
        ↪  the screen.
        - right_single(point='x1 y1'): right single click a position on
        ↪  the screen.
        - drag(start_point='x1 y1', end_point='x2 y2'): drag the mouse
        ↪  from one position to another.
        - hotkey(key='ctrl c'): keyboard shortcut, split keys with
        ↪  spaces
        - type(content='xxx'): type an answer, use escape characters
        ↪  (', ", \n) when needed. Add \n at the end if it is the
        ↪  final submission.
        - scroll(point='x1 y1', direction='down or up or right or
        ↪  left'): scroll to see more content

if platform == Mobile:
        Action Space
        - click(point='x1 y1')
        - long_press(point='x1 y1')
```

```
        - type(content='') #If you want to submit your input, use "\\n"
        ↪  at the end of `content`.
        - scroll(point='x1 y1', direction='down or up or right or
        ↪  left'): scroll to see more content
```

The size of the image is `{w}x{h}.\n`
`{question_images}`

User
Above are two consecutive screenshots. Your task is to select which action is performed in order to transition from the first screenshot to the second.

```
if platform == Desktop:
    {seven action types}
    Which of the above options are correct according to the
    ↪  screenshots?
    Think step by step. You must respond strictly in JSON format
    ↪  following this schema:
    {"thought": "<your reasoning>", "answer": "<A/B/C/D/E/F/G>" }

if platform == Mobile:
    {four action types}
    Which of the above options are correct according to the
    ↪  screenshots?
    Think step by step. You must respond strictly in JSON format
    ↪  following this schema:
    {"thought": "<your reasoning>", "answer": "<A/B/C/D>" }
```

Response Schema (Desktop)

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D/E/F/G>"
}
```

Response Schema (Mobile)

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D>"
}
```

A.3.3 INSTRUCTIONUNDERSTANDING

GoalInterpretation

Task Completion Verification Prompt

System
You are a Graphical User Interface (GUI) agent. You will be given a sequence of screenshots, a task instruction, and three possible answer options: `yes`, `no`, `unknown`. Your goal is to select the best option that indicates whether the task is completed.

- **yes** — The task is clearly completed.
- **no** — The task is not completed.
- **unknown** — The screenshots do not provide enough evidence to determine completion.

User
According to the screenshots below, has the task "`{task}`" been completed?
`{question_images}`

Response Schema
Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<yes/no/unknown>"
}
```

TaskPlanning

GUI Agent Conditional QA Prompt

System

If `question_type == 'yes_or_no'`:
You are a Graphical User Interface (GUI) agent. You will be given a screenshot, a question, and corresponding options. You need to choose one option as your answer.

If `question_type == 'multiple_choice'`:
You are a Graphical User Interface (GUI) agent. You will be given a task instruction, a screenshot, several GUI operations, and four options. Your goal is to select the best option that could solve the task.

`{question_images}`

User
`{question_text}`
`{option_texts}`
Which of the above options are correct according to the screenshot?

Response Rules

If `question_type == 'yes_or_no'`:
Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<yes/no/unknown>"
}
```

If `question_type == 'multiple_choice'`:
Think step by step. You must respond strictly in JSON format following this schema:

```
{
  "thought": "<your reasoning>",
  "answer": "<A/B/C/D>"
}
```

A.4 Full Application List

Here we include the full list of applications involved in our benchmark.

List of Applications

Office (30): Apple Notes, Apple Reminders, Calendar, Docs, Document Viewer, Evince, Gedit, Google Calendar, Google Docs, Google Keep, Keynote, Lark, Libreoffice, Notability, Note-taking App, Notepad, Notes, Notion, Numbers, Office, Overleaf, Pages, Powerpoint, Spreadsheet, Text Editor, VS Code, WPS Office, Microsoft Word, Xcode, Freeform.

Media (18): Amazon Music, Amazon Prime Video, Iheartradio, Likee, Music, Music Player, Pandora, Pocket FM, Podcast Player, Quicktime, Roku, Sofascore, Spotify, TikTok, Tubi, VLC media player, YouTube, YouTube Music.

Game (12): Arena_of_valor, CS2, Chess, Defense_of_the_ancients_2, Dream, Genshin_impact, Minecraft, Nintendo, Pubg, Red_dead_redemption_2, Steam, The Legend Of Zelda Breath Of The Wild.

Editing (20): 3dviewer, Adobe Acrobat, Adobe After Effects, Adobe Express, Adobe Photoshop, Adobe Photoshop Express, Adobe Premiere Pro, CapCut, Davinci Resolve, Draw.io, Gimp, Paint, PDF Editor, Photo Editing Tool, Photo Editor, Picsart, Procreate, Runway, Snapseed, Video Editing Software.

Social & Communication (28): Discord, Facebook, Flickr, Gmail, Google Meet, Google Messages, Imessage, Instagram, LinkedIn, Mail, Messenger, Outlook, Phone, Pinterest, Quora, Reddit, Signal, Slack, Teams Live, Telegram, Threads, Thunderbird, Tumblr, WeChat, Weibo, WhatsApp, X (Twitter), Zoom.

Shopping (25): 12306, Alibaba, Aliexpress, Amazon Shopping, Apartments.com, Applestore, Autoscout24, Autouncle, Booking.com, Car Marketplace, Cars.co.za, Ebay, Edmunds, Expedia, Magento, Offerup, Onestopmarket, Product Listing App, Realtor.com, Redfin, Shop, Taobao, Tripadvisor, Walmart, Wish.

AI & Tools (17): AI Art Generator, Align-anything-dev-omni, Amazon Alexa, Chatbot AI, Chatgpt, Chaton AI, DeepL Translate, Google Translate, Grammarly, Microsoft Copilot, Microsoft Translator, Remix AI Image Creator, Stable Diffusion, Translate, WOMBO Dream, Yandex Translate, Zhiyun Translate.

Browser & Search (10): Bing, DuckDuckGo, Firefox, Google App, Google Chrome, Google Search, Opera, Safari, Web Browser, Web.

Tools (60): Accerciser, Activities, Activity Monitor, App Lock, App Locker, Applock Pro, Automator, Baidu Netdisk, Bluetoothnotificationareaiconwindowclass, Calculator, Camera, Clean, ClevCalc - Calculator, Color Management Utility, Colorsync_utility, Contacts, Control Center, Cursor, Desktop, Dictionary, Digital Color Meter, Disk Utility, Drops, Electron, Email Client, File, File Explorer, File Manager, Files, Filezilla, Finder, Font Book, GPS, Image Viewer, Iphonelockscreen, Kid3, Launcher, Mi Mover, Microsoft Store, Preview, Recorder, Rosetta Stone, Scientific Calculator Plus 991, Script_editor, Search, Shortcuts, Spotlight, Stickies, System Information, System Search, System Settings, Task Manager, Terminal, Totem, ToDesk, Trash, Vim, Voicememos, Vottak, Wallpaper Picker.

Productivity (9): Any.do, Drive, Dropbox Paper, Google Drive, Onedrive, Paperflux, Things, TickTick, Todoist.

News & Reading (22): AP News, BBC News, BBC Sport, Bloomberg, Crimereads, Espn, Forbes, Goodreads, Google News, Google Play Books, Google Scholar, Kindle, Kobo Books, Metacritic, Microsoft News, Newsbreak, Wikidata, Wikipedia, Yahoo Sports, Apple News, Travel Guide App, Travel Review App.

Weather & Navigation (12): Accuweather, Apple Maps, Citymapper, Google Maps, Mapillary, Miuiweather, Msnweather, Navigation App, Openstreetmap, Waze, Weather, Windy.

Finance (8): Alipay, Budgeting App, Investing.com, Paymore, Stocks, Wallet For Your Business, Wallet: Budget Money Manager, Yahoo Finance.

Health & Fitness (4): Fitbit, Fiton, Mideaair, Mifitness.

Job Search (3): Indeed, Job Search By Ziprecruiter, Ziprecruiter.

Transportation (3): Didi, Ryanair, Uber.

System & Tools (15): Android, Android Home Screen, Android Launcher, Android Settings, Android Share Sheet, App Store, Apple, Applibrary, Gnome, Mobile Home Launcher, Mobile Launcher, Mobile Web Browser, OS, Ubuntu, Ubuntu Desktop.

A.5 Action Type Prediction Confusion Matrix

Figure 9 and Figure 10 show the confusion matrix of tested models on desktop and mobile. All of these models have a tendency for predicting click instead of the right actions.

Figure 9: Confusion matrix of action type prediction in desktop.

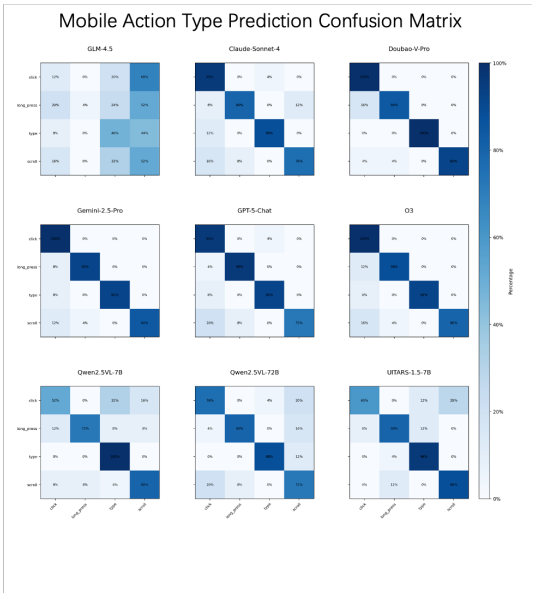

Figure 10: Confusion matrix of action type prediction in mobile.