

GUI-WORLD: A GUI-ORIENTED DATASET FOR MULTIMODAL LLM-BASED AGENTS

Anonymous authors

Paper under double-blind review



Figure 1: GUI-WORLD: A comprehensive dataset for GUI-oriented capabilities encompasses six scenarios and diverse tasks, offering significant potential for real-world applications. All screenshots presented are sampled from our dataset.

ABSTRACT

Recently, Multimodal Large Language Models (MLLMs) have been used as agents to control keyboard and mouse inputs by directly perceiving the Graphical User Interface (GUI) and generating corresponding commands. However, current agents primarily demonstrate strong understanding capabilities in static environments and are mainly applied to relatively simple domains, such as Web or mobile interfaces. We argue that a robust GUI agent should be capable of perceiving temporal information on the GUI, including dynamic Web content and multi-step tasks. Additionally, it should possess a comprehensive understanding of various GUI scenarios, including desktop software and multi-window interactions. To this end, this paper introduces a new dataset, termed GUI-WORLD, which features meticulously crafted Human-MLLM annotations, extensively covering six GUI scenarios and eight types of GUI-oriented questions in three formats. We evaluate the capabilities of current state-of-the-art MLLMs, including vision LLMs and video LLMs, in understanding various types of GUI content, especially dynamic and sequential content. Our findings reveal that vision LLMs struggle with dynamic GUI content without manually annotated keyframes or operation history. On the other hand, video LLMs fall short in all GUI-oriented tasks given the sparse GUI video dataset. Based on GUI-WORLD, we take the initial step of leveraging a fine-tuned video LLM as a GUI agent, demonstrating an improved understanding of various GUI tasks. However, due to the limitations in the performance of base LLMs, we conclude that using video LLMs as GUI agents remains a significant challenge. We believe our work provides valuable insights for future research in dynamic GUI content understanding.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs), such as GPT-4V(ision) (OpenAI, 2023) and LLaVA (Liu et al., 2023a), have significantly contributed to the development of both vision and

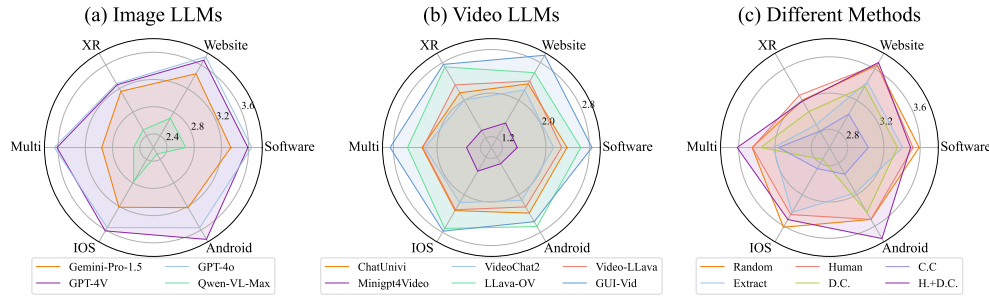


Figure 2: Comparative performance of different MLLMs in six scenarios of GUI-WORLD. (a) Performance of four mainstream Image LLMs. (b) Performance of three video LLMs and our GUI-Vid. (c) Performance among six methods. See subsection 4.2 for more details.

language domains (Yin et al., 2024). These models bring forth innovative solutions and paradigms for traditional visual tasks, including visual reasoning (Yang et al., 2023), medical image interpretation (Li et al., 2023a; Zhang et al., 2024a), and applications in embodied agents (Huang et al., 2024a). One particularly promising area is Graphical User Interface (GUI) understanding, which holds significant potential for real-world applications, such as webpage comprehension (Hong et al., 2023; Lai et al., 2024) and navigation by GUI agents (Zhang et al., 2023a; Niu et al., 2024; Wang et al., 2024a). The key challenges of GUI understanding are twofold: effective GUI agents are expected to (1) possess a deep understanding of GUI elements, including webpage icons, text identified through Optical Character Recognition (OCR), and page layouts, and (2) exhibit an exceptional ability to follow instructions within GUI contexts, such as conducting searches through search engines.

Despite significant progress, as illustrated in Table 1, existing works suffer from the following limitations: (1) *Inability to Handle Dynamic Environments*. Most studies predominantly focus on the static features of GUI scenarios, neglecting the need for MLLMs to effectively process sequential information and dynamic operations. For instance, an agent’s task performance can be disrupted by unexpected elements such as pop-up advertisements, underscoring a gap in handling dynamic sequential tasks. (2) *Limited Scenarios*. Current research is typically restricted to Web-based environments, which limits the models’ generalization and robustness. For instance, GUI agents may need to operate across diverse platforms such as Windows, macOS, Linux, iOS, Android, and XR environments. Additionally, operations may sometimes involve multiple windows. Therefore, expanding the scope of research to encompass these varied environments will enhance the adaptability and effectiveness of GUI agents.

To mitigate these gaps, this paper introduces GUI-WORLD, a comprehensive dataset containing 12,379 GUI videos, specifically designed to evaluate and enhance the capabilities of GUI agents. This dataset encompasses a wide range of GUI scenarios, including popular websites, desktop and mobile applications across various operating systems, multi-window interactions, as well as XR environments. The data collection process involves sourcing GUI videos from screen recordings and instructional videos on YouTube. Subsequently, we utilize a Human-MLLM collaborative approach to generate a diverse set of questions and instructions and finally construct GUI-WORLD.

Likewise, we also establish a comprehensive benchmark for GUI understanding, which encompasses seven mainstream MLLMs, three keyframe selection strategies, six GUI scenarios, and a diverse array of queries in multiple-choice, free-form, and conversational formats, aiming to provide a thorough evaluation of the MLLMs’ GUI-oriented capabilities. As shown in Figure 2, the assessment results indicate that most MLLMs struggle with GUI-WORLD, highlighting their limited dynamic understanding of graphical interfaces and underscoring the need for further enhancement.

Leveraging this dataset, we take the first step to fine-tune a video GUI agent adept at handling dynamic and sequential GUI tasks, leading to substantial enhancements in the general capabilities of GUI agents and showcasing the utility and effectiveness of GUI-WORLD. Additionally, we delve into discussing various factors critical to GUI understanding, including the integration of textual information, the number of keyframes, image resolutions, and vision perception for a pioneer comprehensive study in the GUI domain.

Overall, the key contributions of this paper are three-fold:

Table 1: Comparison of GUI datasets and benchmarks. ‘Sem.’: semantic instruction level, ‘VL’: Vision-Language, ‘Seq.’: Tasks for sequential images, ‘Cro.’: Cross-app or multi-window tasks, ‘Dyn.’: Tasks for dynamic GUI content.

Dataset	Size	Sem.	VL	Video	Env Type				Task Coverage			Task
					Web.	Mob.	Desk.	XR	Seq.	Cro.	Dyn.	
Rico (Deka et al., 2017)	72,219	Low	✓	✓	✗	✓	✗	✗	✓	✓	✗	UI Code/Layout Generation Mobile Navigation
MetaGUI (Sun et al., 2022)	1,125	Low	✓	✗	✗	✓	✗	✗	✓	✗	✗	
UGIF (Venkatesh et al., 2022)	523	High	✓	✗	✗	✓	✗	✗	✓	✗	✗	Instruction Following GUI Understanding
AITW (Rawles et al., 2023)	715,142	High	✓	✗	✗	✓	✗	✗	✓	✗	✗	
Ferret-UI (You et al., 2024)	123,702	Low	✓	✗	✗	✓	✗	✗	✗	✗	✗	UI Grounding & Understanding Web Navigation
MiniWoB++ (Liu et al., 2018)	100	Low	✓	✗	✓	✗	✗	✗	✗	✗	✗	
WebArena (Zhou et al., 2023)	812	Low	✓	✗	✓	✗	✗	✗	✓	✗	✗	Web Navigation Web Navigation
Mind2Web (Deng et al., 2024)	2,350	Both	✓	✓	✓	✗	✗	✗	✓	✗	✗	
OmniAct (Kapoor et al., 2024)	9,802	Low	✓	✓	✓	✗	✓	✗	✓	✗	✗	Code Generation
MMINA (Zhang et al., 2024b)	1,050	Low	✓	✗	✓	✗	✗	✗	✓	✓	✗	Web Navigation
AgentStudio (Zheng et al., 2024a)	304	High	✓	✗	✓	✗	✓	✗	✓	✓	✗	General Control
OSWorld (Xie et al., 2024)	369	High	✓	✗	✓	✗	✓	✗	✓	✓	✗	General Control
GUI-WORLD (Ours)	12,379	Both	✓	✓	✓	✓	✓	✓	✓	✓	✓	GUI Understanding Instruction Following

- **A New Dataset.** We propose GUI-WORLD, a comprehensive GUI dataset comprising 12,379 videos specifically designed to assess and improve GUI-oriented capabilities of MLLMs, spanning a range of categories and scenarios, including desktop, mobile, and eXtended Reality (XR), and representing the first GUI-oriented instruction-tuning dataset in video domain.
- **A Novel Model.** Based on GUI-WORLD, we propose GUI-Vid, a GUI-oriented video LLM with enhanced capabilities to handle various and complex GUI tasks. GUI-Vid shows a significant improvement on the benchmark and achieves results comparable to the top-performing models.
- **Comprehensive Experiments and Valuable Insights.** Our experiments indicate that most existing MLLMs continue to face challenges with GUI-oriented tasks, particularly in sequential and dynamic GUI content. Empirical findings suggest that improvements in vision perception, along with an increase in the number of keyframes and higher resolution, can boost performance in GUI-oriented tasks, thereby paving ways for the future of GUI agents.

2 GUI-WORLD: A DATASET FOR GUI UNDERSTANDING

2.1 OVERVIEW

We introduce GUI-WORLD, a comprehensive dataset covering six GUI scenarios including video, human-annotated keyframes, as well as detailed captions and diverse types of QA produced by our data curation framework, aiming at benchmarking and enhancing the general GUI-oriented capabilities. These GUI scenarios encompass desktop operating systems (e.g., macOS, Windows) and mobile platforms (e.g., Android and iOS), websites, software, and even extended-range technologies (XR) (e.g., GUI in Apple Vision Pro (Apple, 2024)). We divide the dataset into a train-test split, each containing 10,702 and 1,677 samples. Discussion for each scenario is in subsection A.1.

As illustrated in Figure 3, the development of GUI-WORLD is structured around a two-stage process. Details regarding video and query statistics are provided in Table 2, which includes distributions of the number of keyframes, video lengths, and the lengths of queries and their corresponding golden answers shown in Figure 4 and Figure 5 for case study.

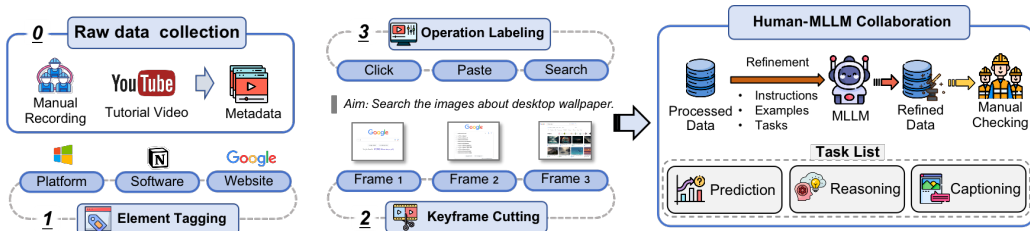


Figure 3: An overview construction pipeline of GUI-WORLD.

Table 2: The statistics of GUI-WORLD. For Android, we select videos from Rico (Deka et al., 2017) and randomly sample 10 frames. Avg. Frame refers to the average number of frames in each clip, and Avg. Anno. refers to the average number of manually annotated GUI actions.

Category	Total Videos	Free-form	MCQA	Conversation	Total Frame. (Avg.)	Avg. Anno.
Software	4,720	27,840	9,440	9,440	23,520 (4.983)	7.558
Website	2,499	14,994	4,998	4,998	15,371 (6.151)	6.862
IOS	492	2,952	984	984	2,194 (4.459)	7.067
Multi	475	2,850	950	950	2,507 (5.277)	7.197
XR	393	2,358	786	786	1,584 (4.030)	10.970
Android	3,800	15,199	7,600	7,600	38,000 (10.000)	-
Summary	12,379	76,673	24,758	24,758	83,176 (6.719)	7.463

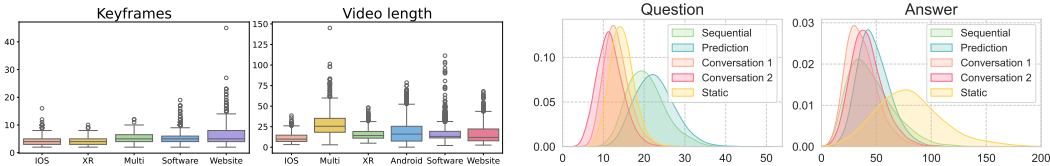


Figure 4: **Left:** Distribution of the number of keyframes and video lengths. **Right:** Length distribution for each type of question and its golden answer.

2.2 GUI VIDEO COLLECTION AND KEYFRAME ANNOTATION PROCESS

We describe the pipeline for collecting screen recordings from student workers and GUI-related instructional videos from YouTube for GUI-WORLD and the procedures followed to convert these videos into keyframe sequences.

A significant portion of our video data is derived from screen recordings executed by student workers, which can directly reflect real-life GUI usage scenarios. A typical video collection scenario involves assigning a student worker a specific software task. The student begins by familiarizing themselves with the software, followed by recording a series of operations in a short video clip, such as “Sign up”, “Sign in”, “Create a New Page”, and “Invite Other Collaborators” in the software “Notion¹”.

Despite the high fidelity of these manually recorded videos, we encounter several challenges: (1) Student workers often require substantial time to acquaint themselves with professional software (e.g., MATLAB, Adobe After Effects (Ae)), which can hinder the progress of data collection. (2) The videos may lack comprehensiveness, typically capturing only commonly used operations and overlooking rarer functions crucial for dataset completeness. To address these issues, we also source videos from social media platforms that host a diverse array of GUI videos. Specifically, we download tutorial videos from YouTube—given its prevalence as a video-sharing platform—because they richly detail various GUI operations. These videos are then segmented into shorter clips, each representing a distinct sequence of operations.

The subsequent step involves annotating these video clips with keyframes and textual descriptions of each keyframe using custom-designed annotation software. Although several algorithms exist for keyframe extraction (Zhu et al., 2016; Yan et al., 2018; Mahasseni et al., 2017; OpenCV), they typically underperform with GUI videos where changes between frames might be minimal (e.g., a slight movement in the mouse cursor). To ensure high-quality datasets, we therefore perform manual extraction of these keyframes. Each keyframe is meticulously annotated to include details such as the operation performed, the purpose between two keyframes, the software or website used, mouse actions (e.g., scroll, click), and keyboard inputs (e.g., copy (Ctrl + C), paste (Ctrl + V), specific input). We detail our annotation process in subsection A.3.

2.3 GUI TASKS GENERATION FROM HUMAN-MLLM COLLABORATION

Drawing insights from prior research (Li et al., 2024a; Dekoninck et al., 2024; Yu et al., 2023a), we develop a Human-MLLM collaboration pipeline to annotate captions and diverse types of QA

¹<https://www.notion.so/>

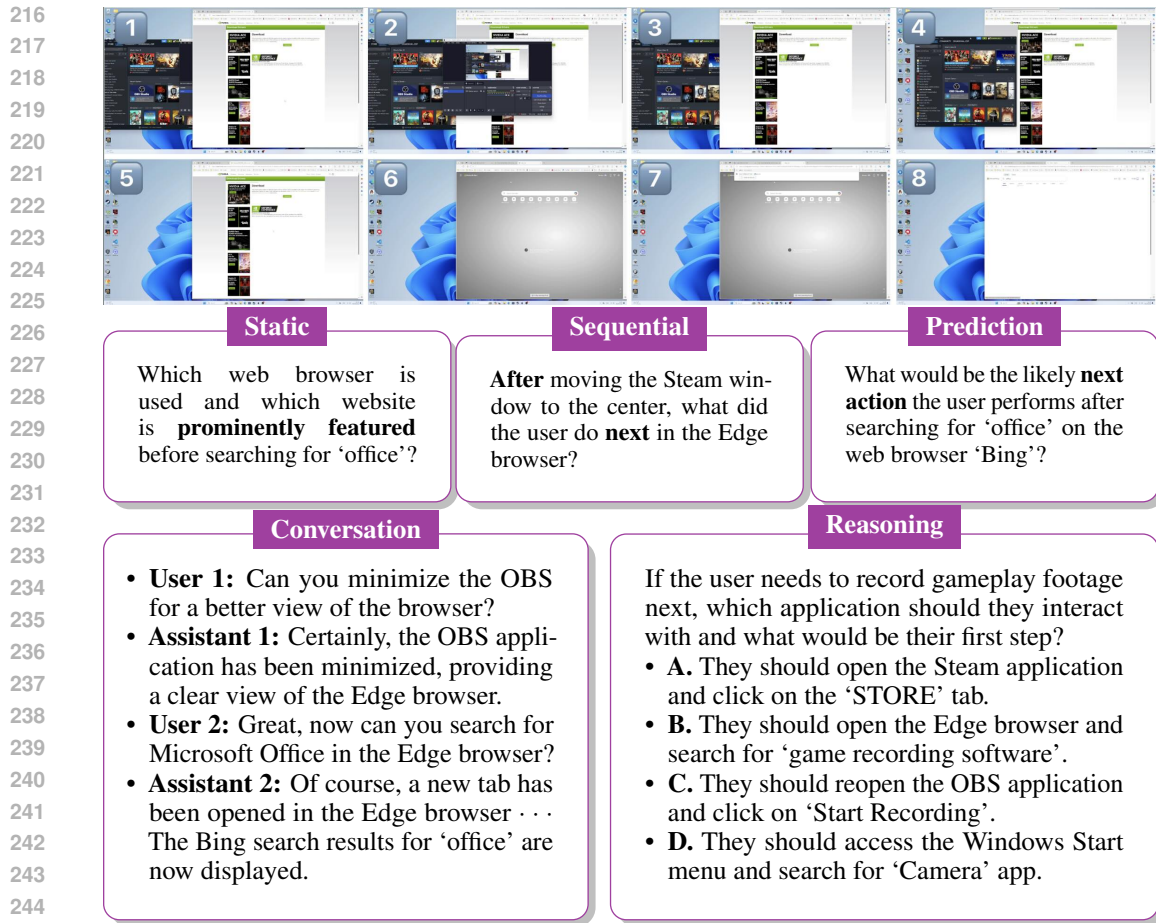


Figure 5: An example in multi-window GUI scene as a case study.

specifically tailored for GUI comprehension. The process involves inputting an instructional prompt, a comprehensive description, key information (*e.g.*, system or application), and a sequence of human-annotated keyframes into GPT-4V. As depicted in [Table 10](#), GUI-WORLD features various question types, detailed as follows:

▷ **Detailed and Summarized Captioning:** This task challenges basic GUI knowledge and multimodal perception, also addressing the deficiency of detailed GUI content in video-caption pairs. Initially, GPT-4V generates two distinct descriptions for each video: one concentrating on fine-grained details and the other on overall information. Furthermore, GPT-4V provides a succinct summary, highlighting core operations and overarching objectives in the video.

▷ **Static GUI Content:** This task challenges MLLM with textual, layout, and iconographic analysis of static GUI content. We instruct GPT-4V to generate free-form queries with a golden answer concerning static GUI elements or specific scenes that recur in more than two keyframes, ensuring their consistent presence in the video. Additionally, GPT-4V also crafts QA pairs that evaluate inferential skills in static content, focusing on interrelations among icons or textual information.

▷ **Dynamic and Sequential GUI Content:** This task concentrates on temporal content in GUI video, such as dynamically changing interfaces, and aims to elucidate the sequential information and reasoning chains within GUI content. We direct GPT-4V to identify consistently changing elements to create queries for dynamic content. Moreover, predictive tasks are formulated on order and temporal relation in provided sequential images, challenging agents to anticipate future events or states.

In the last stage, human annotators will follow the guideline in [subsection A.3](#) and carefully review the entire video and MLLM-generated QA pairs to correct inaccuracies and hallucinations, as well as supplement information for both questions and answers to **make these tasks more challenging**.

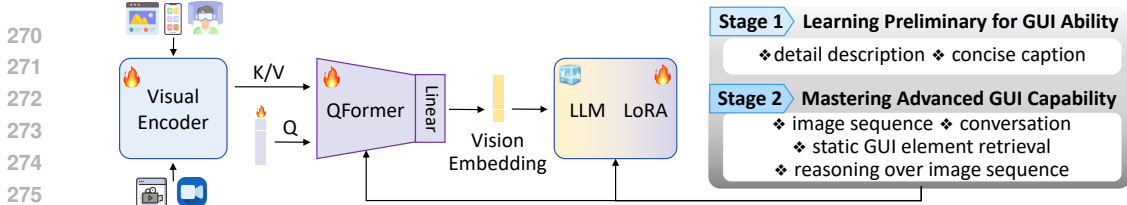


Figure 6: An overview of our fine-tuning architecture, focusing on 1) GUI content alignment and 2) GUI-oriented tasks instruction tuning.

3 PROGRESSIVE ENHANCEMENT ON GUI PERCEPTION ABILITY

We introduce our strategy to enhance the GUI-oriented capabilities of current MLLMs on both static and dynamic GUI content. Inspired by previous studies (Lai et al., 2024; Li et al., 2023b), we structure our methodology into two distinct fine-tuning stages, as illustrated in Figure 6. Initially, we fine-tune the MLLM on simpler tasks, such as description queries and captioning exercises, to instill a basic understanding of GUI elements. Subsequently, building on this foundation, the second stage aims to augment the MLLM’s proficiency with more complex and challenging tasks. Our fine-tuning is all based on the Supervised Fine-Tuning (SFT): $\mathcal{L}_{\text{SFT}}(\pi_{\theta}) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} [\log \pi_{\theta}(y | x)]$, where x is the input, y is LLMs’ output, and π_{θ} denotes the model parameters that need to be optimized.

Stage-1: Learning Preliminary for GUI Content. The initial phase focuses on aligning GUI content with a pre-trained vision encoder and a base LLM, utilizing GUI videos accompanied by detailed descriptions and captions. This phase aims to embed a robust understanding of fundamental GUI concepts and terminology within the MLLM. By engaging the model in basically captioning various GUI components, the model learns to recognize and articulate the functionalities and visual characteristics of these elements, thereby laying a solid groundwork for GUI knowledge.

Stage-2: Mastering Advanced GUI Capability. Building on the foundational knowledge established in Stage 1, the second stage focuses on advancing the MLLM’s proficiency in interacting with GUI elements through more complex tasks. These tasks are designed to simulate real-world scenarios that the MLLM might encounter in GUI environments, which include predicting based on image sequences, engaging in conversations, retrieving both static and dynamic GUI elements, and performing reasoning tasks.

As illustrated in Figure 6, We employ the two-stage training architecture utilizing VideoChat2 (Li et al., 2023b) as our foundational model. Initially, videos and images are encoded using the UMT-L visual encoder (Li et al., 2024b). Subsequently, a QFormer compresses visual tokens into a smaller set of query tokens. Drawing inspiration from (Dai et al., 2023), we enhance the QFormer (Zhang et al., 2023b) by integrating instructions to enable it to extract visual representations pertinent to the given instructions. Additionally, we apply low-rank adaptation (LoRA (Hu et al., 2021)) to base LLM. This model is concurrently fine-tuned with the visual encoder and QFormer using a Vision-grounded Text Generation (VTG) loss: $\mathcal{L}_{\text{VTG}}(\theta) = -\mathbb{E} [\log p(y|v; \theta)]$, where v represents the visual tokens derived from the QFormer, and y represents the text output grounded in the visual context. Training dataset and details can be found in Appendix C.

4 EXPERIMENTS AND ANALYSIS

4.1 EXPERIMENTAL SETUPS

Models.² We conduct evaluations on four of the most popular vision LLMs: GPT-4V(ision) (OpenAI, 2023), GPT-4o (OpenAI, 2024a), Qwen-VL-Max (Bai et al., 2023), and Gemini-Pro-1.5 (GeminiTeam, 2023). We benchmark on three keyframe selection settings: (1) *Random*, where frames are sampled at fixed time intervals within a video; (2) *Extracted*, with keyframes extracted using Katna (KeplerLab, 2023), and ablation study with UVD (Zhang et al., 2024c) in VIP (Ma et al., 2022) and R3M (Nair et al., 2022) settings; and (3) *Human*, where keyframes are selected by humans during the annotation process. **For the *Random* and *Extracted* settings, we input 10 frames into**

²Given that GPT-4V was announced to be deprecated, we used GPT-4o to conduct some ablation studies instead of GPT-4V, aiming to ensure our results provide longer-term reference value.

Table 3: The overall performance in six GUI scenarios for MCQA and Free-form queries. ‘R.’, ‘E.’, and ‘H.’ denote random-selected, programmatic-selected, and human-selected keyframes, respectively. ‘MC’ means Multiple-Choice QA and ‘Free’ represents the average score of all free-form and conversational queries.

Models	Setting	Software		Website		XR		Multi		IOS		Android		Avg.			
		MC	Free	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free		
Vision LLMs	Gemini-Pro-1.5	R.	81.7%	3.339	82.6%	3.452	81.2%	3.154	81.2%	2.959	82.0%	3.213	81.6%	3.220	81.7%	3.223	
		E.	78.5%	3.152	77.8%	3.215	80.8%	3.006	71.8%	2.777	79.3%	3.007	78.5%	3.168	77.8%	3.054	
	Qwen-VL-Max	R.	74.9%	2.676	76.9%	2.656	74.2%	2.469	68.8%	2.432	75.4%	2.779	73.7%	2.309	74.0%	2.553	
		E.	74.3%	2.624	75.8%	2.627	69.0%	2.499	64.8%	2.362	77.4%	2.659	65.8%	2.277	71.2%	2.508	
		H.	75.8%	2.651	75.5%	2.698	77.6%	2.373	66.9%	2.490	74.3%	2.633	74.2%	2.559	74.0%	2.568	
	GPT-4V	R.	81.5%	3.589	80.9%	3.648	80.6%	3.200	75.0%	3.452	82.5%	3.614	78.3%	3.515	79.8%	3.503	
		E.	85.1%	3.407	80.1%	3.433	81.8%	2.892	81.9%	3.219	86.4%	3.427	79.9%	3.176	82.6%	3.259	
		H.	86.0%	3.520	79.8%	3.655	83.4%	3.265	76.9%	3.449	79.9%	3.453	81.3%	3.466	81.2%	3.469	
	GPT-4o	H.	86.5%	3.644	83.3%	3.740	84.3%	3.285	81.1%	3.654	83.3%	3.558	90.0%	3.561	84.8%	3.573	
	Video LLMs	ChatUnivi	-	28.4%	2.389	22.2%	2.349	20.6%	2.161	17.5%	2.275	22.6%	2.337	23.0%	2.390	22.4%	2.317
		Minigt4Video	-	18.9%	1.475	15.3%	1.520	16.3%	1.362	15.4%	1.457	20.1%	1.501	14.6%	1.342	16.8%	1.443
		VideoChat2	-	45.5%	2.144	42.6%	2.221	44.0%	2.005	40.4%	2.222	40.2%	2.169	44.7%	2.119	42.9%	2.147
Video-LLaVA		-	52.9%	2.290	52.4%	2.410	44.2%	2.258	45.9%	2.329	49.7%	2.319	51.3%	2.259	49.4%	2.311	
LLaVA-One-Vision		-	56.9%	2.641	48.4%	2.588	59.6%	2.709	52.9%	2.306	58.1%	2.717	24.3%	2.675	50.0%	2.606	
GUI-Vid		-	59.9%	2.847	54.1%	2.957	55.6%	2.764	52.9%	2.861	51.8%	2.773	53.4%	2.572	54.6%	2.796	

Table 4: Overall performance in six GUI scenarios for MCQA and Free-form queries. ‘D.C.’ means detailed caption, and ‘C.C.’ means concise caption, and **X** means no vision input.

Models	Setting		Software		Website		XR		Multi		IOS		Android		Avg.	
	Vision	Text	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free
GPT-4V	X	D.C.	85.0%	3.350	83.1%	3.380	82.3%	3.056	84.2%	3.358	81.6%	2.751	81.7%	3.427	83.0%	3.316
	X	C.C.	80.7%	3.028	72.2%	3.025	82.8%	2.809	81.3%	3.160	76.5%	2.868	76.4%	2.939	78.3%	2.971
	✓	D.C.	82.5%	3.494	83.2%	3.682	85.9%	3.191	83.9%	3.617	80.9%	3.516	84.9%	3.758	83.5%	3.543

each MLLM, while the *Human* setting uses an average of 6.719 frames, as detailed in Table 2. Each model’s responses employ a three-step Chain-of-Thought (CoT) (Wei et al., 2023) process, i.e., ‘Describe-Analyze-Answer’, to evaluate their peak performance. Additionally, we assessed three advanced video LLMs—ChatUnivi (Jin et al., 2023), Minigt4-video (Ataallah et al., 2024), and Videochat2 (Li et al., 2023c)—for their performance on GUI content. Detailed experimental setups are referred to Appendix C. Comparison to some recently released video LLMs—VideoLLaVA (Lin et al., 2023a) and LLaVA-OneVision (Li et al., 2024c)—are referred to Appendix D.

Evaluation Metrics. To assess free-form questions and multiple-round conversations, we utilize the LLM-as-a-Judge methodology, which assigns a similarity score ranging from 1 to 5 between MLLM’s response and a predefined golden answer, already validated by previous studies (Zheng et al., 2023; Liu et al., 2023b; Chen et al., 2024a). For a comprehensive evaluation, we also provide BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) in Appendix D. For multiple-choice questions, we measure performance using accuracy as the primary evaluation metric.

Textual Information Integration. To investigate the effectiveness of integrating image-caption models for LLMs—typically employed in natural videos—and the helpfulness of textual GUI content in accomplishing GUI-oriented tasks, we implement three experimental settings: Detailed Caption, Concise Caption, and Vision + Detailed Caption. GPT-4V is utilized to provide captions of these keyframes, integrating human annotators’ operational intents to more accurately describe each frame, being validated in subsection A.3.

Quality and Quantity of Vision Input. To explore the upper bound of GUI-oriented capabilities, particularly in dynamic and sequential tasks, we conduct ablation studies focusing on the impact of the quality and quantity of vision input. We vary the number of keyframes (8, 16) fed into GUI-Vid. Additionally, we test the effect of different vision input on GPT-4o, using both low and high settings, as well as without providing images, to further assess how resolution influences performance.

4.2 EMPIRICAL RESULTS

Commercial Vision LLMs outperform Open-source Video LLMs in Zero-shot Settings. Commercial vision LLMs, notably GPT-4V and GPT-4o, consistently outperform open-source video

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

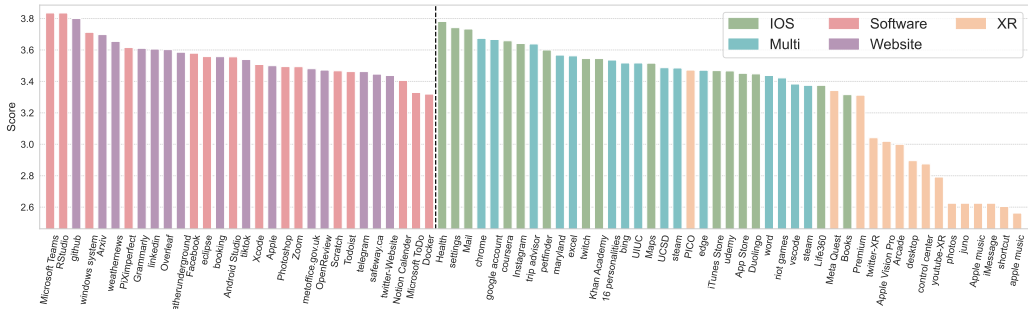


Figure 7: Fine-grained performance of GPT-4V in each GUI scenario (w.o. Android).

Table 5: Detailed scores for each tasks in the software-related scenarios. ‘Dyn.’ refers to queries on dynamic GUI content, and ‘Pred.’ indicates prediction tasks.

Models	Setting	Caption		Complex Tasks			Conversation		Average		
		Concise	Detailed	Static	Dyn.	Pred.	Round 1	Round 2			
Vision LLMs	Gemini-Pro-1.5	R.	3.659	2.837	2.969	2.822	3.450	3.608	3.845	3.339	
		E.	3.350	2.468	2.741	2.431	3.292	3.458	3.837	3.152	
	Qwen-VL-Max	R.	2.381	1.758	2.277	2.144	2.724	3.125	3.317	2.676	
		E.	2.459	1.693	2.143	1.954	2.742	3.174	3.298	2.624	
		H.	2.474	1.711	2.137	2.032	2.834	3.223	3.257	2.651	
	GPT-4V	R.	3.579	2.676	3.243	3.011	3.630	3.925	4.131	3.589	
		E.	3.141	2.301	2.927	2.627	3.541	3.844	4.103	3.407	
		H.	3.352	2.509	3.053	2.849	3.609	3.928	4.163	3.520	
	GPT-4o	H.	4.048	3.028	3.125	3.117	3.562	4.129	4.318	3.644	
	Video LLMs	ChatUnivi	-	1.587	1.240	1.705	1.656	2.524	2.698	3.366	2.389
		Minigt4Video	-	1.246	1.073	1.249	1.235	1.675	1.494	1.719	1.475
		VideoChat2	-	1.992	1.312	1.812	1.682	2.158	2.342	2.720	2.144
Video-LLaVA		-	1.519	1.241	1.657	2.367	1.550	2.587	3.293	2.290	
LLaVA-One-Vision		-	2.149	1.762	1.868	2.970	1.925	2.947	3.492	2.641	
GUI-Vid		-	3.562	2.058	2.376	2.090	3.435	3.080	3.260	2.847	

LLMs in zero-shot settings. As detailed in Table 3, GPT-4o exhibits superior performance across all GUI scenarios in complex tasks, reflected in its high scores in both multiple-choice and free-form queries, with an average of 84.8% and 3.573. Similarly, Gemini demonstrates strong capabilities in captioning and descriptive tasks within software and iOS environments, scoring 2.836 and 2.936, respectively, as shown in Table 21. Further analysis (Figure 7) reveals that GPT-4V excels in applications with minimal textual content and simple layouts, such as TikTok, health apps, and GitHub. In contrast, its performance drops in more intricate applications like Microsoft ToDo and XR software. As for video LLMs, their significantly poorer performance is attributed to two main factors: their inability to accurately interpret GUI content from user inputs and a lack of sufficient GUI-oriented pretraining, which is evident from their inadequate performance in basic captioning and description tasks. See Appendix D for other metrics and detailed fine-grained performance.

Performance Variate in Different GUI Scenarios and Applications. GPT-4V (Figure 7) and Gemini (Figure 16) excel in common scenarios such as mobile and website interfaces but show marked deficiencies in more complex GUI environments like XR and multi-window interactions, across both captioning and intricate tasks. This performance gap highlights a significant shortfall in understanding environments where GUI elements are scattered and demand sophisticated interpretation. It emphasizes the critical need for specialized benchmarks and datasets tailored to these complex GUI scenarios, which is essential for enhancing the GUI-oriented capabilities of MLLMs, paving the way for them to become truly reliable and high-performing general control agents.

Keyframe Selection is Important for GUI-oriented Tasks. Across both basic tasks such as captioning and more complex tasks like prediction and reasoning, significant variations are evident among keyframe selection methods. As shown in Table 22 and Table 24, GPT-4V and Gemini significantly benefit from using human-selected keyframes, scoring approximately 0.2-0.3 points higher in both captioning and free-form tasks than those using programmatic extraction. This suggests that traditional keyframe technologies, designed for natural videos, are less effective for detecting essential GUI operations, particularly when subtle movements like mouse clicks and dynamic changes

Table 6: The overall results for ablation study on GUI-Vid finetuning. F.K. and E.K. mean keyframes during the finetuning and evaluation process respectively. L.: Image, V.: Video.

Setting	F.K.	E.K.	Data		Software		Website		XR		Multi		IOS		Android		Avg.	
			I.	V.	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free	MC	Free
Baseline	-	8	-	-	45.5%	2.144	42.6%	2.221	44.0%	2.005	40.4%	2.222	40.2%	2.169	44.7%	2.119	42.9%	2.147
	-	16	-	-	45.1%	2.144	41.8%	2.240	41.0%	2.007	40.7%	2.238	39.9%	2.138	44.7%	2.147	42.2%	2.154
GUI-Vid	8	8	✗	✓	58.3%	2.709	53.6%	2.817	62.2%	2.626	54.2%	2.627	53.1%	2.708	54.9%	2.501	56.0%	2.665
			✓	✓	59.9%	2.856	54.1%	2.925	59.0%	2.751	52.1%	2.837	50.0%	2.756	54.0%	2.571	54.8%	2.782
	16	16	✗	✓	59.0%	2.709	55.1%	2.821	62.8%	2.645	53.3%	2.624	55.5%	2.727	55.7%	2.501	56.9%	2.671
			✓	✓	59.9%	2.847	54.1%	2.957	55.6%	2.764	52.9%	2.861	51.8%	2.772	53.4%	2.572	54.6%	2.796

Table 7: Average score of GPT-4o under w/o vision input, low and high resolution.

Setting	Desc.	Conv.	Dyn.	Static	Caption	Average
w/o Vision	1.872	3.915	2.979	2.486	2.187	2.688
Low Resolution	2.794	3.912	3.150	2.869	3.672	3.279
High Resolution	3.031	4.056	3.318	3.131	3.911	3.489

are involved. Therefore, we try model-based keyframe identifier mainly utilized in robotics, with UVD+VIP achieving the best results shown in Table 18, even comparative with human selected, further indicating that keyframe selection is crucial. However, small differences in Qwen-VL-Max demonstrate that different methods may exert less influence on less capable models.

Dynamic GUI Tasks Continue to Challenge MLLMs. In the fine-grained tasks depicted in Table 5, GPT-4V and GPT-4o excel with static GUI content and prediction tasks over image sequences but struggle with providing detailed descriptions for entire videos and dynamic content. This discrepancy is attributed to minor variations in GUI that significantly impact its semantic meaning. Enhancing the number of keyframes and the granularity of perception might mitigate these issues. Among video LLMs, ChatUnivi excels in conversational tasks by effectively leveraging contextual nuances, particularly in subsequent rounds, yet it underperforms in caption tasks. In contrast, GUI-Vid demonstrates proficiency in sequential tasks but falls short in both captioning and static content. This gap is linked to deficiencies in backbone pretraining, which lacked comprehensive GUI content crucial for effective vision-text alignment, as evidenced by its poor performance in simple caption task shown in Table 21 and an instruction tuning process failed to fully address these shortcomings.

Vision Perception is Important for Sequential GUI Tasks. As demonstrated in Table 5, integrating detailed textual information slightly outperforms purely vision-based inputs or detailed captions, akin to a Chain of Thought (CoT) (Wei et al., 2023) setting. Surprisingly, GPT-4V excels in caption and prediction tasks with just detailed captions, providing insights on enhancing specific GUI-oriented tasks through additional textual information. However, it still falls short in more challenging tasks, such as retrieving static or dynamic content. This underscores the critical role of visual perception in GUI environments, where even minor changes can significantly impact outcomes.

Supreme Enhancement of GUI-Vid on Graphic-based Interface After Fine-tuning on GUI-WORLD.

As a pioneering study in training video LLMs as screen agents, GUI-Vid significantly outperforms the baseline model, showing an average improvement of 30% across various tasks and GUI scenarios, even surpassing the commercial vision LLM, Qwen-VL-Max. This enhancement is particularly notable in captioning and prediction over image sequences, where GUI-Vid matches the performance of GPT-4V and Gemini-Pro. As shown in Figure 8, our two-stage progressive finetuning significantly enhances the performance in all GUI scenarios. Remarkably, GUI-Vid also perform well in qualitative experiments as shown in Table 8, where more than 80% of users consider GUI-Vid more helpful after fine-tuned on GUI-WORLD. However, in our explorative experiments, GUI-Vid still fails in GUI operating tasks via code generation like GPT-4o performing in Cradle (Tan et al., 2024), which is likely due to the baseline LLM’s weaker performance and the challenges of code generation instruction fine-tuning.

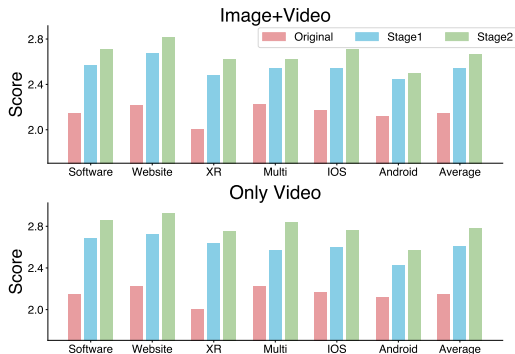


Figure 8: Two stages of progressive training enhance GUI ability.

Upper Bound of GUI-oriented Capability with More Keyframes and High Resolution. As depicted in Table 6, our two ablation studies during the fine-tuning phase demonstrate that utilizing GUI image-text captioning data significantly enhances the model’s preliminary understanding of GUI elements, outperforming training that relies solely on videos. Additionally, an increased number of keyframes correlates with improved performance across various scenarios, notably in environments featuring multiple windows and software applications. Further evidence from Table 7 reveals that higher image resolutions substantially boost task performance, both basic and complex, for GPT-4o. These findings underscore the potential for further developing a more robust GUI Agent.

Table 8: User Preference: GUI-Vid vs VideoChat2 (w. and w.o. Fine-tuning on GUI-World).

Scenarios	GUI-Vid	Tie	VideoChat2
Software	82.7%	13.3%	4.0%
Website	86.0%	12.0%	2.0%
XR	88.0%	8.7%	3.3%
Multi	85.3%	10.0%	8.7%
IOS	92.0%	6.0%	2.0%
Android	82.0%	16.0%	2.0%
Average	86.0%	11.0%	3.7%

5 RELATED WORK

MLLM-based Agents for GUI. Building upon the significant advancements in LLMs (Team, 2024; Meta, 2023a;b; OpenAI, 2024b) and advanced modality-mixing technologies (Li et al., 2023d; Alayrac et al., 2022), groundbreaking MLLMs such as GPT-4V (OpenAI, 2023) and Gemini-Pro (GeminiTeam, 2023), along with open-source MLLMs like the LLaVA-1.6 series (Liu et al., 2023a;c), CogVLM (Wang et al., 2024b), and Qwen-VL series (Bai et al., 2023), have shown outstanding performance across various tasks (Yu et al., 2023b; Liu et al., 2023d; Chen et al., 2024b; Wu et al., 2023; Wake et al., 2023; Huang et al., 2024b; Zhang et al., 2024d; Zhao et al., 2024; Gui et al., 2024). Venturing beyond text and single image, several studies are now exploring the integration of video modalities for tasks requiring dynamic or sequential visual content (Jin et al., 2023; Li et al., 2023b; Maaz et al., 2023; Lin et al., 2023a). In the GUI domain, leveraging the robust vision perception capabilities of MLLMs, applications such as WebAgents (Hong et al., 2023; Zhang et al., 2024e; Zheng et al., 2024a) and Mobile Agents (You et al., 2024; Wang et al., 2024a; Chu et al., 2023) have gained popularity for handling everyday tasks like navigation and VQA. Frontier research is also investigating the use of MLLMs as general control agents, such as in playing computer games (Tan et al., 2024; Lin et al., 2023b) and serving as OS co-pilots (Song et al., 2024; Xie et al., 2024), paving the way for more complex GUI operations.

GUI Benchmark & Dataset. Building upon the foundational work of Rico (Deka et al., 2017), the first mobile GUI video dataset, and AitW (Rawles et al., 2023), which features 715k episodes of sequential images, research has extensively covered mobile (Sun et al., 2022; Li et al., 2020; Zhang et al., 2023c) and web GUI environments (Lù et al., 2024; Zhou et al., 2023; Yao et al., preprint; Koh et al., 2024; Liu et al., 2024a). Mind2Web (Deng et al., 2024) stands out in web-based datasets with over 2,000 tasks from 137 websites across 31 domains. Advances continue into desktop GUIs with new toolkits (Zheng et al., 2024a), benchmarks (Kapoor et al., 2024; Mialon et al., 2023), and frameworks (Zheng et al., 2024b; Liu et al., 2023e; Niu et al., 2024). Research on GUI also transfers from comprehending single images in a static workspace (Hong et al., 2023) to sequential operations or multi-hop scenarios (Xie et al., 2024; Zhang et al., 2024b), challenging the understanding and operation capability of these powerful models.

6 CONCLUSION

In this paper, we have introduced GUI-WORLD, a comprehensive GUI-oriented video dataset designed to benchmark and enhance understanding of virtual interfaces, especially sequential and dynamic tasks. This dataset extensively covers six scenarios and various tasks, addressing the previous research gap in comprehensively evaluating models’ capabilities in graphic-based understanding. We conduct extensive benchmarks on leading MLLMs and the first video-LLM-based Agent ‘GUI-Vid’ fine-tuned on GUI-WORLD specifically for GUI-oriented content, achieving results comparable to top-performing models, providing detailed insights into enhancing GUI-related capabilities. We believe our work offers valuable insights for future research in dynamic GUI content understanding.

REFERENCES

- 540
541
542 OpenAI. Openai models - gpt-4-vision. [https://openai.com/research/
543 gpt-4v-system-card](https://openai.com/research/gpt-4v-system-card), 2023.
- 544 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*,
545 2023a.
- 546
547 Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on
548 multimodal large language models, 2024.
- 549
550 Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu,
551 Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning
552 and action, 2023.
- 553
554 Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan
555 Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision
assistant for biomedicine in one day, 2023a.
- 556
557 Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang
558 He, Brian Davison, Xiang Li, Hui Ren, Sunyang Fu, James Zou, Wei Liu, Jing Huang, Chen Chen,
559 Yuyin Zhou, Tianming Liu, Xun Chen, Yong Chen, Quanzheng Li, Hongfang Liu, and Lichao
560 Sun. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision,
561 language, and multimodal tasks, 2024a.
- 562
563 Jianguo Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li,
564 Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world,
2024a.
- 565
566 Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan
567 Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A
568 visual language model for gui agents, 2023.
- 569
570 Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen
571 Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. Autowebglm: Bootstrap and reinforce a large
language model-based web navigating agent, 2024.
- 572
573 Chi Zhang, Zhao Yang, Jiakuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu.
574 Appagent: Multimodal agents as smartphone users, 2023a.
- 575
576 Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and
Qi Wang. Screenagent: A vision language model-driven computer control agent, 2024.
- 577
578 Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao
579 Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception, 2024a.
- 580
581 Biplob Deka, Zifeng Huang, Chad Franzen, Joshua Hirschman, Daniel Afegan, Yang Li, Jeffrey
582 Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design
583 applications. In *Proceedings of the 30th annual ACM symposium on user interface software and
technology*, pages 845–854, 2017.
- 584
585 Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. Meta-gui: towards
586 multi-modal conversational agents on mobile gui. *arXiv preprint arXiv:2205.11029*, 2022.
- 587
588 Sagar Gubbi Venkatesh, Partha Talukdar, and Srini Narayanan. Ugif: Ui grounded instruction
following. *arXiv preprint arXiv:2211.07615*, 2022.
- 589
590 Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Android in the
591 wild: A large-scale dataset for android device control. *arXiv preprint arXiv:2307.10088*, 2023.
- 592
593 Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei
Yang, and Zhe Gan. Ferret-ui: Grounded mobile ui understanding with multimodal llms. *arXiv
preprint arXiv:2404.05719*, 2024.

- 594 Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement
595 learning on web interfaces using workflow-guided exploration. In *International Conference on*
596 *Learning Representations (ICLR)*, 2018. URL <https://arxiv.org/abs/1802.08802>.
597
- 598 Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng,
599 Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building
600 autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.
- 601 Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su.
602 Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing*
603 *Systems*, 36, 2024.
- 604 Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem Alshikh,
605 and Ruslan Salakhutdinov. Omniact: A dataset and benchmark for enabling multimodal generalist
606 autonomous agents for desktop and web. *arXiv preprint arXiv:2402.17553*, 2024.
607
- 608 Ziniu Zhang, Shulin Tian, Liangyu Chen, and Ziwei Liu. Mmina: Benchmarking multihop multimodal
609 internet agents. *arXiv preprint arXiv:2404.09992*, 2024b.
- 610 Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng Yan. Agentstu-
611 dio: A toolkit for building general virtual agents. *arXiv preprint arXiv:2403.17918*, 2024a.
612
- 613 Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing
614 Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal
615 agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*,
616 2024.
- 617 Apple. Apple vision pro. <https://www.apple.com/apple-vision-pro/>, 2024.
618
- 619 Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. A key volume mining deep framework
620 for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern*
621 *recognition*, pages 1991–1999, 2016.
- 622 Xiang Yan, Syed Zulqarnain Gilani, Hanlin Qin, Mingtao Feng, Liang Zhang, and Ajmal Mian. Deep
623 keyframe detection in human action videos. *arXiv preprint arXiv:1804.10021*, 2018.
- 624 Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with
625 adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern*
626 *Recognition*, pages 202–211, 2017.
627
- 628 OpenCV. Opencv. <https://opencv.org/>.
- 629 Yuan Li, Yue Huang, Yuli Lin, Siyuan Wu, Yao Wan, and Lichao Sun. I think, therefore i am:
630 Benchmarking awareness of large language models using awarebench, 2024a.
631
- 632 Jasper Dekoninck, Marc Fischer, Luca Beurer-Kellner, and Martin Vechev. Understanding large
633 language models through the lens of dataset generation, 2024. URL [https://openreview.](https://openreview.net/forum?id=miGpIhquyB)
634 [net/forum?id=miGpIhquyB](https://openreview.net/forum?id=miGpIhquyB).
- 635 Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen,
636 and Chao Zhang. Large language model as attributed training data generator: A tale of diversity
637 and bias, 2023a.
- 638 KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and
639 Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.
640
- 641 Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked
642 teacher: Towards training-efficient video foundation models, 2024b.
- 643 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,
644 Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language
645 models with instruction tuning, 2023.
646
- 647 Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle
attention, 2023b.

- 648 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
649 and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
650
- 651 OpenAI. Hello gpt-4o, May 2024a. URL <https://openai.com/index/hello-gpt-4o/>.
652 Accessed: 2024-06-06.
- 653 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,
654 and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization,
655 text reading, and beyond, 2023.
- 656 GeminiTeam. Gemini: A family of highly capable multimodal models, 2023.
657
- 658 KeplerLab. Katna - video to keyframes and video to image crops utility. [https://github.com/
659 keplerlab/katna](https://github.com/keplerlab/katna), 2023. Accessed: 2024-09-07.
- 660 Zichen Zhang, Yunshuang Li, Osbert Bastani, Abhishek Gupta, Dinesh Jayaraman, Yecheng Jason
661 Ma, and Luca Weihs. Universal visual decomposer: Long-horizon manipulation made easy. In
662 *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6973–6980. IEEE,
663 2024c.
- 664 Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy
665 Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training.
666 *arXiv preprint arXiv:2210.00030*, 2022.
667
- 668 Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal
669 visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.
670
- 671 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le,
672 and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- 673 Peng Jin, Ryuichi Takano, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified
674 visual representation empowers large language models with image and video understanding. *arXiv
675 preprint arXiv:2311.08046*, 2023.
- 676 Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and
677 Mohamed Elhoseiny. Minigt4-video: Advancing multimodal llms for video understanding with
678 interleaved visual-textual tokens, 2024.
- 679 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen,
680 Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv
681 preprint arXiv:2311.17005*, 2023c.
682
- 683 Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual
684 representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a.
685
- 686 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei
687 Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint
688 arXiv:2408.03326*, 2024c.
- 689 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
690 Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.
691 Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- 692 Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke,
693 Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Hongning Wang, Jing Zhang, Minlie
694 Huang, Yuxiao Dong, and Jie Tang. Alignbench: Benchmarking chinese alignment of large
695 language models, 2023b.
- 696 Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang,
697 Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with
698 vision-language benchmark, 2024a.
699
- 700 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic
701 evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association
for Computational Linguistics*, pages 311–318, 2002.

- 702 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating
703 text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
704
- 705 Weihao Tan, Ziluo Ding, Wentao Zhang, Boyu Li, Bohan Zhou, Junpeng Yue, Haochong Xia, Jiechuan
706 Jiang, Longtao Zheng, Xinrun Xu, et al. Towards general computer control: A multimodal agent
707 for red dead redemption ii as a case study. *arXiv preprint arXiv:2403.03186*, 2024.
708
- 709 OpenAI Team. Gpt-4 technical report, 2024.
- 710 Meta. Llama 2, 2023a. <https://llama.meta.com/llama2>.
711
- 712 Meta. Llama 3, 2023b. <https://llama.meta.com/llama3>.
713
- 714 OpenAI. Mistral ai, 2024b. <https://mistral.ai/company/>.
715
- 716 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image
717 pre-training with frozen image encoders and large language models, 2023d.
- 718 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel
719 Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan
720 Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian
721 Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo
722 Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language
723 model for few-shot learning, 2022.
724
- 725 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction
726 tuning, 2023c.
- 727
- 728 Weihao Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
729 Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang.
730 Cogvlm: Visual expert for pretrained language models, 2024b.
- 731 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and
732 Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023b.
733
- 734 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi
735 Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player?
736 *arXiv preprint arXiv:2307.06281*, 2023d.
- 737
- 738 Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi
739 Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large
740 vision-language models?, 2024b.
- 741 Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou,
742 Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case
743 studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023.
744
- 745 Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-
746 4v (ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint*
747 *arXiv:2311.12015*, 2023.
- 748
- 749 Sili Huang, Jifeng Hu, Zhejian Yang, Liwei Yang, Tao Luo, Hechang Chen, Lichao Sun, and Bo Yang.
750 Decision mamba: Reinforcement learning via hybrid selective sequence modeling, 2024b.
- 751 Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang,
752 Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can mixed human-
753 written and machine-generated text be detected? In Kevin Duh, Helena Gomez, and Steven
754 Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages
755 409–436, Mexico City, Mexico, June 2024d. Association for Computational Linguistics. URL
<https://aclanthology.org/2024.findings-naacl.29>.

- 756 Wei Zhao, Zhitao Hou, Siyuan Wu, Yan Gao, Haoyu Dong, Yao Wan, Hongyu Zhang, Yulei Sui, and
757 Haidong Zhang. NL2Formula: Generating spreadsheet formulas from natural language queries. In
758 Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Lin-*
759 *guistics: EACL 2024*, pages 2377–2388, St. Julian’s, Malta, March 2024. Association for Computa-
760 tional Linguistics. URL <https://aclanthology.org/2024.findings-eacl.158>.
- 761 Yi Gui, Zhen Li, Yao Wan, Yemin Shi, Hongyu Zhang, Yi Su, Shaoling Dong, Xing Zhou, and
762 Wenbin Jiang. Vision2ui: A real-world dataset with layout for code generation from ui designs,
763 2024.
- 764 Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt:
765 Towards detailed video understanding via large vision and language models. *arXiv preprint*
766 *arXiv:2306.05424*, 2023.
- 767 Chaoyun Zhang, Liqun Li, Shilin He, Xu Zhang, Bo Qiao, Si Qin, Minghua Ma, Yu Kang, Qingwei
768 Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Ufo: A ui-focused agent for windows os
769 interaction, 2024e.
- 770 Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu
771 Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. Mobilevlm : A fast, strong and open vision
772 language assistant for mobile devices, 2023.
- 773 Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng
774 Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with
775 gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023b.
- 776 Zirui Song, Yaohang Li, Meng Fang, Zhenhao Chen, Zecheng Shi, and Yuan Huang. Mmac-copilot:
777 Multi-modal agent collaboration operating system copilot. *arXiv preprint arXiv:2404.18074*, 2024.
- 778 Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge. Mapping natural language
779 instructions to mobile ui action sequences, 2020.
- 780 Danyang Zhang, Hongshen Xu, Zihan Zhao, Lu Chen, Ruisheng Cao, and Kai Yu. Mobile-Env: An
781 evaluation platform and benchmark for llm-gui interaction. *CoRR*, abs/2305.08144, 2023c. URL
782 <https://arxiv.org/abs/2305.08144>.
- 783 Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with
784 multi-turn dialogue, 2024.
- 785 Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable
786 real-world web interaction with grounded language agents. In *ArXiv*, preprint.
- 787 Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham
788 Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating
789 multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.
- 790 Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang
791 Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and
792 grounding?, 2024a.
- 793 Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom.
794 Gaia: a benchmark for general ai assistants, 2023.
- 795 Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web
796 agent, if grounded, 2024b.
- 797 Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong,
798 and Jie Tang. Webglm: Towards an efficient web-enhanced question answering system with human
799 preferences, 2023e.
- 800 Ambreen Kousar, Saif Ur Rehman Khan, Shahid Hussain, M. Abdul Basit Ur Rahim, Wen-Li Wang,
801 and Naseem Ibrahim. A systematic review on pattern-based gui testing of android and web apps:
802 State-of-the-art, taxonomy, challenges and future directions. In *2023 25th International Multitopic*
803 *Conference (INMIC)*, pages 1–7, 2023. doi: 10.1109/INMIC60434.2023.10465949.

- 810 Rodrigo Funabashi Jorge, Márcio Eduardo Delamaro, Celso Gonçalves Camilo-Junior, and Auri
811 Marcelo Rizzo Vincenzi. Test data generation based on gui: A systematic mapping. In
812 *International Conference on Software Engineering Advances*, 2014. URL [https://api.
813 semanticscholar.org/CorpusID:64041598](https://api.semanticscholar.org/CorpusID:64041598).
- 814 Ivans Kulesovs. ios applications testing, 2015. URL [https://api.semanticscholar.org/
815 CorpusID:59015994](https://api.semanticscholar.org/CorpusID:59015994).
- 816
817 Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu.
818 Seeclck: Harnessing gui grounding for advanced visual gui agents, 2024.
- 819 Han Hu, Haolan Zhan, Yujin Huang, and Di Liu. Pairwise gui dataset construction between android
820 phones and tablets, 2023.
- 821
822 Tony Beltramelli. pix2code: Generating code from a graphical user interface screenshot, 2017.
- 823
824 An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Qinghong Lin, Linjie Li, Jianfeng Wang, Jian-
825 wei Yang, Yiwu Zhong, Julian J. McAuley, Jianfeng Gao, Zicheng Liu, and Lijuan Wang.
826 Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation.
827 *ArXiv*, abs/2311.07562, 2023. URL [https://api.semanticscholar.org/CorpusID:
828 265149992](https://api.semanticscholar.org/CorpusID:265149992).
- 829 Hajime Nakajima, Takeshi Masuda, and Ikuya Takahashi. Gui ferret: Gui test tool to analyze complex
830 behavior of multi-window applications. *2013 18th International Conference on Engineering of
831 Complex Computer Systems*, pages 163–166, 2013. URL [https://api.semanticscholar.
832 org/CorpusID:837553](https://api.semanticscholar.org/CorpusID:837553).
- 833 Philipp A. Rauschnabel, Reto Felix, Christian Hinsch, Hamza Shahab, and Florain Alt. What is xr?
834 towards a framework for augmented and virtual reality. *Comput. Hum. Behav.*, 133:107289, 2022.
835 URL <https://api.semanticscholar.org/CorpusID:247861674>.
- 836
837 Meta quest 3: New mixed reality vr headset. <https://www.meta.com/quest/quest-3>.
- 838 Brian K. Sanders, Yuzhong Shen, and Dennis A. Vincenzi. Understanding user interface pref-
839 erences for xr environments when exploring physics and engineering principles. In *Interna-
840 tional Conference on Applied Human Factors and Ergonomics*, 2019. URL [https://api.
841 semanticscholar.org/CorpusID:197940610](https://api.semanticscholar.org/CorpusID:197940610).
- 842
843 OpenAI. Chatgpt, 2023. <https://openai.com/product/chatgpt>.
- 844 Kevin Qinghong Lin, Linjie Li, Difei Gao, Qinchen WU, Mingyi Yan, Zhengyuan Yang, Lijuan
845 Wang, and Mike Zheng Shou. Videogui: A benchmark for gui automation from instructional
846 videos. *arXiv e-prints*, pages arXiv–2406, 2024.
- 847
848 Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai,
849 Xinyi Liu, Hanlin Zhao, et al. Visualagentbench: Towards large multimodal models as visual
850 foundation agents. *arXiv preprint arXiv:2408.06327*, 2024b.
- 851 Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee.
852 Llava-next: Improved reasoning, ocr, and world knowledge, January 2024c. URL [https:
853 //llava-vl.github.io/blog/2024-01-30-llava-next/](https://llava-vl.github.io/blog/2024-01-30-llava-next/).
- 854
855
856
857
858
859
860
861
862
863

864 **Part I**

865

866 **Appendix**

867

868

869 **Table of Contents**

870

871	A Details of Dataset Construction	18
872	A.1 Six Main GUI Categories	18
873	A.2 Selected Website/Software	18
874	A.3 Human Keyframes Annotation Process	21
875		
876	B Dataset Analysis	21
877		
878	C Details of Experiments Setups	22
879	C.1 Fine-tune dataset construction	22
880	C.2 Hyperparameter Settings	24
881	C.3 Evaluation.	25
882		
883	D Additional Experiments Results	26
884		
885	E Prompts	27
886		
887	F Case Study	41

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

A DETAILS OF DATASET CONSTRUCTION

A.1 SIX MAIN GUI CATEGORIES

In earlier endeavors pertaining to GUI, such as those involving GUI testing (Kousar et al., 2023; Jorge et al., 2014; Kulesovs, 2015), the focus was segmented into GUIs for Website, Software, IOS and Android platforms. However, as a comprehensive GUI dataset, we include all potential GUI scenarios in our dataset to ensure that our data is the most comprehensive knowledge that the GUI Agent needs to learn; we divide these scenarios into six categories:

- **Android.** This category focuses on the GUI scenarios that occur within the Android operating system, which is predominantly used on smartphones. Android’s ubiquity in the mobile market has led to a wide variety of GUI designs and interaction patterns, making it a rich field for study. This category has been the subject of extensive scrutiny in scholarly works such as (Deka et al., 2017; Li et al., 2020; Rawles et al., 2023; Cheng et al., 2024).
- **Software.** This category encapsulates the GUI scenarios arising within software applications, whether they are standalone programs or components of a larger suite. The diversity of software applications, from productivity tools to creative suites, offers a wide range of GUI scenarios for exploration. The literature is rich with research in this area, such as (Hu et al., 2023).
- **Website.** This category is concerned with the GUI scenarios that manifest within a web browser. Given the ubiquity of web browsing in modern digital life, this category holds significant relevance. It holds a substantial representation in academic literature, with pioneering papers such as (Deng et al., 2024; Kapoor et al., 2024) proposing excellent GUI datasets for websites.
- **IOS.** This category zeroes in on the GUI scenarios that transpire within the iOS operating system, the proprietary system for Apple devices like the iPhone and iPad. The iOS platform is known for its distinct design aesthetics and interaction patterns, providing a unique context for GUI research. A number of studies, such as (Beltramelli, 2017; Yan et al., 2023) make use of GUI information in IOS.
- **Multi Windows.** This category is dedicated to GUI scenarios that necessitate simultaneous interaction with multiple windows, a common occurrence in desktop environments where users often juggle between several applications or documents. Despite the common use of multi-window interaction in everyday GUI usage, there has been relatively little research into this area (Nakajima et al., 2013). The need for efficient multitasking in such scenarios presents unique challenges and opportunities for GUI design and interaction research. As of our knowledge, there are no specific datasets catering to these multi-window GUI scenarios.
- **XR.** XR encompasses Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) (Rauschnabel et al., 2022). Given the advancements in XR technology and the growing accessibility of commercial-grade head-mounted displays (Apple, 2024; Met), XR has emerged as a novel medium for human-computer interaction. This necessitates the exploration of GUI within XR environments. In these scenarios, the GUI takes on a 3D, immersive form (Sanders et al., 2019), demanding the agent to comprehend and navigate a 3D space. The emerging field of XR presents a new frontier for GUI research, with unique challenges and opportunities due to its immersive and interactive nature. To date, as far as we are aware, there are no datasets that specifically address GUI in the realm of XR.

A.2 SELECTED WEBSITE/SOFTWARE

In our study, we select a diverse range of websites and software to comprehensively evaluate GUI understanding capabilities across various user scenarios. These selections cover essential categories such as social media, productivity tools, online shopping, and educational platforms, providing a broad spectrum of GUI environments.

The chosen websites, as shown in Figure 9, include popular social media platforms like Instagram, Twitter, and LinkedIn, which are integral to understanding dynamic and interactive GUI elements. We also include widely-used productivity tools such as Microsoft Teams, Notion, and Slack to evaluate GUI tasks in professional and collaborative settings.

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025







<p>Productivity </p> <ul style="list-style-type: none"> Asana Dropbox EndNote Evernote Google Drive Google Meet Mendeley Microsoft OneDrive Microsoft Teams Notion OneNote Slack Todolist Trello Zoom Zotero 	<p>Education </p> <ul style="list-style-type: none"> Adobe Digital Editions Amazon Kindle Blackboard Coursera edX Google Classroom Kahoot! Khan Academy MATLAB Microsoft Teams for Education Moodle PhET Interactive Simulations Quizlet Scratch Stellarium Turnitin Udemy
<p>Entertainment </p> <ul style="list-style-type: none"> Amazon Prime Video Apple Music Disney+ HBO Max Hulu Netflix Pandora Spotify Twitch YouTube 	<p>Social Media </p> <ul style="list-style-type: none"> Discord Facebook Instagram LinkedIn Messenger (Facebook) Pinterest Snapchat TikTok Twitter WeChat WhatsApp
<p>Windows System Software </p> <ul style="list-style-type: none"> Alarm & Clock Calculator Calendar Control Panel Cortana File Explorer Mail Microsoft Edge Microsoft Store Paint Photos Settings Snipping Tool Sticky Notes Task Manager Windows Media Player WordPad 	<p>Mac System Software </p> <ul style="list-style-type: none"> Calendar Contacts FaceTime Finder Mail Maps Messages Music Notes Photos Podcasts Preview Reminders Safari Siri TV

Figure 9: List of desktop softwares in GUI-WORLD.

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

Social Media

- <https://instagram.com/>
- <https://twitter.com/>
- <https://whatsapp.com/>
- <https://pinterest.com/>
- <https://linkedin.com/>
- <https://tiktok.com/>
- <https://discord.com/>
- <https://reddit.com/>
- <https://telegram.org/>

Search Engines

- <https://google.com/>
- <https://yandex.com/>
- <https://bing.com/>
- <https://baidu.com/>
- <https://search.aol.com/>

Online Shopping

- <https://etsy.com/>
- <https://alibaba.com/>
- <https://ebay.com/>

Education and Learning

- <https://quora.com/>
- <https://byjus.com/>
- <https://cambridge.org/>
- <https://udemy.com/>
- <https://coursera.org/>
- <https://khanacademy.org/>
- <https://edx.org/>
- <https://academia.edu/>

Technology and Software

- <https://microsoft.com/>
- <https://apple.com/>
- <https://adobe.com/>
- <https://github.com/>
- <https://openai.com/>
- <https://oracle.com/>
- <https://vmware.com/>

Travel and Hospitality

- <https://booking.com/>
- <https://tripadvisor.com/>
- <https://yelp.com/>
- <https://airbnb.com/>
- <https://expedia.com/>
- <https://hotels.com/>
- <https://trivago.com/>
- <https://homeaway.com/>

Finance

- <https://moneycontrol.com/>
- <https://bloomberg.com/>
- <https://investing.com/>
- <https://marketwatch.com/>
- <https://seekingalpha.com/>
- <https://zacks.com/>

Food and Cooking

- <https://cookpad.com/>
- <https://allrecipes.com/>
- <https://foodnetwork.com/>
- <https://yummly.com/>
- <https://epicurious.com/>

Figure 10: List of some websites in GUI-WORLD.

For software shown in [Figure 10](#), we incorporate key applications like Adobe Photoshop and MATLAB to assess GUI operations in specialized and technical environments. Additionally, video conferencing tools like Zoom and cloud storage services like Google Drive are included to represent common remote work and file management scenarios.

These selections ensure that our study encompasses a wide array of user interactions and GUI complexities, thereby providing a robust evaluation of the current state-of-the-art methods in GUI understanding by MLLMs and comprehensively constructing a high-quality dataset.

A.3 HUMAN KEYFRAMES ANNOTATION PROCESS

Annotator’s Information The annotation is conducted by 16 authors of this paper and 8 volunteers independently. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have knowledge in the GUI domain, with different genders, ages, and educational backgrounds. The education backgrounds of annotators are above undergraduate. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to use software to record videos or edit video clips. We also provide them with detailed criteria and task requirements in each annotation process.

Recording Video. For self-recording videos, we employ OBS³ on the Windows system for screen capturing and the official screen recording toolkit on the Mac/iOS system. This process necessitates human annotators to execute a series of targeted actions within specific websites or applications, which are subsequently captured as raw video footage. These actions, commonplace in everyday usage, enhance the reliability of our dataset. Subsequently, the raw videos are segmented into sub-videos, each encapsulating multiple actions (e.g., clicking a button) to achieve a specific objective (e.g., image search). The videos are then processed to extract keyframes annotated with detailed descriptions.

Edition Based on YouTube Videos. For sourcing videos from YouTube, we utilize a search protocol formatted as “[website name/application name] + tutorial” to compile relevant video lists. Human annotators first review these videos to understand the primary operations they depict. These videos are then divided into sub-videos, each containing several actions directed towards a single goal (e.g., image search). Like the self-recorded footage, these segments are processed to isolate keyframes and furnish them with descriptive annotations.

Keyframes Annotation. After obtaining the GUI video clips, human annotators will filter out the keyframes of the operations based on the video content and the mouse and keyboard actions at that time. They will also label the sub-operations or targets between the two keyframes. Once the annotation is complete, the annotators will provide an overall description of the entire video, summarizing the main goal of the human operations in the video. After all the information is annotated, we will use an LLM to refine the text content, reducing any errors made by human annotators and adjusting the sentence structure. The prompt we use for the LLM to polish the human annotations is shown in [Figure 11](#) and [Figure 12](#).

Human-LLM Cooperated Instruction Generation. To curate and refine the golden answer of each video-instruction pair generated by GPT-4V, given that the raw response from GPT-4V may contain harmful content or hallucinations. The role of humans in the golden answer generation process is to enhance the difficulty of the questions and remove harmful and incorrect content, as shown in [Table 9](#).

Human verifying GPT-4V annotated captions. We evaluate the quality of annotations from GPT-4V by selecting 1,000 detailed descriptions and captions generated by GPT-4V, which are then assessed by human annotators. The high satisfaction rate of 98% underscores the quality and relevance of the GPT-4V annotations.

B DATASET ANALYSIS

In this section, we provide an analysis of the length distribution of QA in each GUI scenario, as illustrated in [Figure 13](#) and [Figure 14](#). Questions focused on sequential and predictional tasks are

³<https://obsproject.com/>

Table 9: Comparison of raw answer from GPT-4V and golden answer from humans.

Raw response from GPT-4V	Golden answer from humans
<p>”Q”: ”What subcategories are visible under the ’Lifestyle’ section in the static GUI content of the video?”</p> <p>”A”: ”In the ’Lifestyle’ section of the telegram.com website, the visible subcategories are ’Worcester Living’, ’Faith’, and ’Food’. These are clearly displayed in a horizontal list just below the main ’Lifestyle’ category header, indicating different areas of content that users can explore within the lifestyle section.”</p>	<p>”Q”: ”What subcategories are visible under the ’Lifestyle’ section in the static GUI content of the video?”</p> <p>”A”: ”The ’Lifestyle’ section of the telegram.com website features visible subcategories such as ’Worcester Living’, ’Faith’, and ’Food’. These are prominently shown in a horizontal list beneath the main ’Lifestyle’ header, offering users various content areas to explore within the section.”</p>

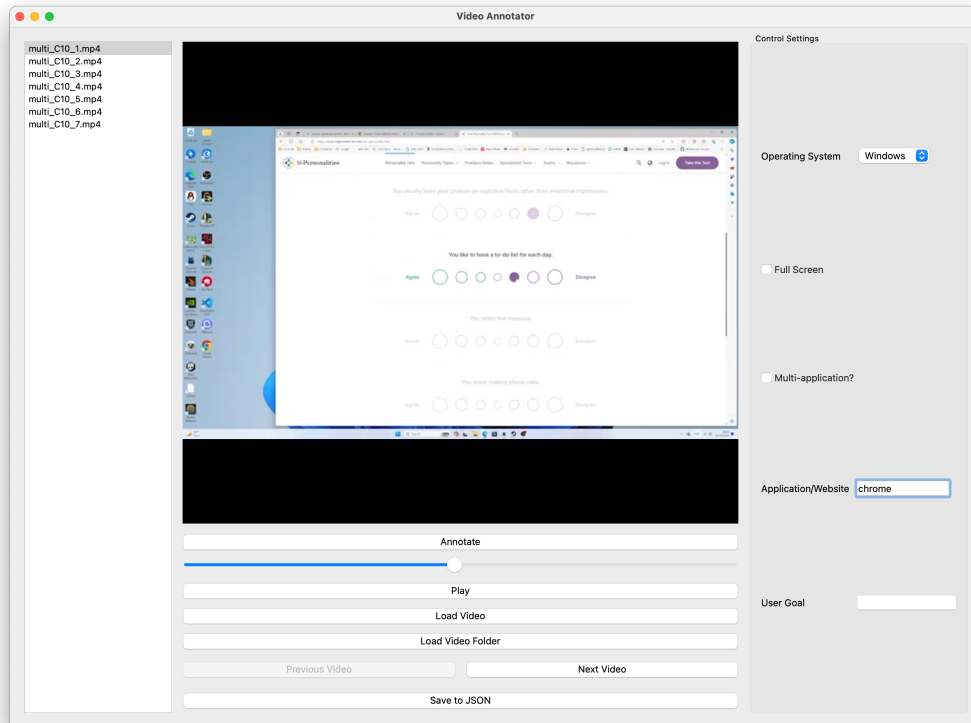


Figure 11: The overall preview of our annotating software.

slightly longer than other types, while the golden answer of static tasks tends to be longer. Length of Question-answer pair in various GUI scenarios is similarly distributed, with questions in Android environment being slightly shorter, and answers in XR environment being longer.

C DETAILS OF EXPERIMENTS SETUPS

C.1 FINE-TUNE DATASET CONSTRUCTION

We use two settings to fine-tune GUI-Vid, one with video-text pairs only, and the other with video-text and image-text pairs, which are all GUI content:

- **Video Only.** In this setting, we only train GUI-Vid with video-text pairs in GUI-WORLD, as shown in Table 11.

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

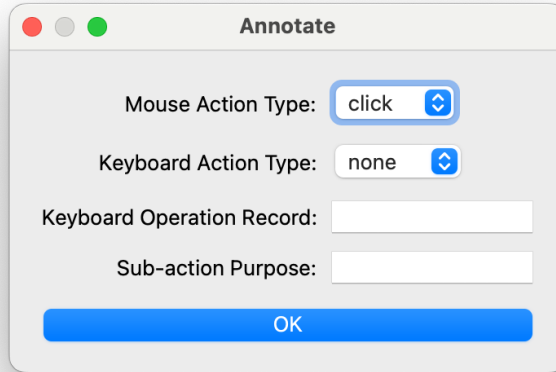


Figure 12: The interface for annotating a keyframe, consists of mouse action, keyboard action, and a short sub-action purpose.

Table 10: Examples of diverse question types in GUI-WORLD.

T.	Question	Examples
Caption	Detailed Description	Q: <i>Please provide a detailed description of what occurs throughout these sequential GUI images.</i> A: The video shows a user taking the 16 Personalities test on a Windows desktop using the Edge browser...
	Summarized Caption	Q: <i>Write a clear description of the video, make sure the key features are well covered.</i> A: Creating a new IT team in Todoist by selecting industry, job function, role, team size, and inviting members.
Static	Layout, Icon Retrieval	Q: <i>What related searches are suggested on the right side of the Bing results for 'emnlp 2024'?</i> A: The suggested related searches shown include 'emnlp 2024 miami', 'eac1 2024 call for papers'...
	Textual Retrieval	Q: <i>What is the estimated time to complete the content for Week 2 of the course?</i> A: The estimated time to complete the content for Week 2 of the course is 1 hour...
	Interrelations in GUI Content	Q: <i>What is the name of the browser and the tab where the user performs the product search?</i> A: The browser is Microsoft Edge, and the user performs the product search in the eBay tab.
Dynamic	Content Retrieval	Q: <i>What specific action does the user take after turning their head to the left to view the left side of the page?</i> A: After turning their head to the left to view the left side of the page, the user performs...
	Prediction	Q: <i>Given the mouse is over 'Add NeurIPS 2024 DB Track Submission,' what's the likely next step?</i> A: It would be to click on the 'Add NeurIPS 2024 Datasets and Benchmarks Track Submission' button...
	Sequential Reasoning	Q: <i>Scrolls down from the 'Moon Gravity', which of the following cheats? A. Change Weather B. Skyfall ...</i> A: [[B]]

- **Video-Image.** Inspired by the pre-trained process of Videochat2, we include image-text pairs to help the visual encoder align GUI knowledge. These images are selected from our GUI-WORLD, MetaGUI (Sun et al., 2022), and OmniAct (Kapoor et al., 2024) for high-quality GUI content. Subsequently, we use GPT-4V to generate a detailed description and a concise caption for each image. Finally, we construct a dataset consisting of video-text and image-text pairs for gaining comprehensive GUI-oriented capabilities.

Table 11: Video-only fine-tune dataset.

Stage	Data types	Amount
1	Detailed Description	14,276
	Concise Caption	7,138
2	GUI VQA	21,414
	Multiple-Choice QA	14,276
	Conversation	7,138

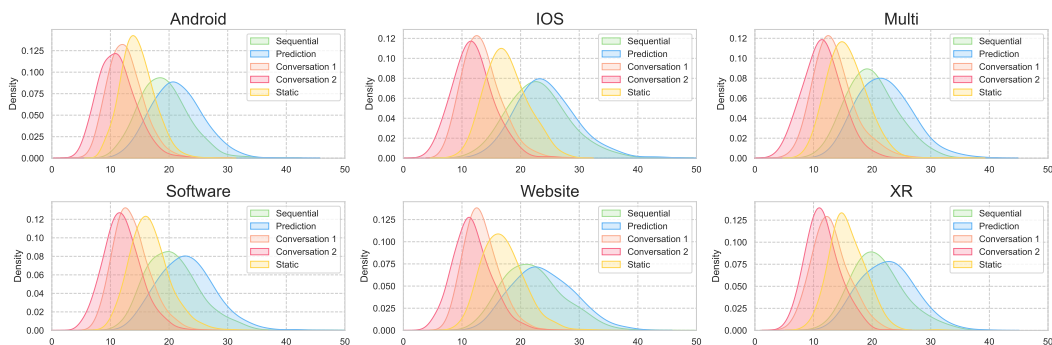


Figure 13: Length distribution of free-form questions.

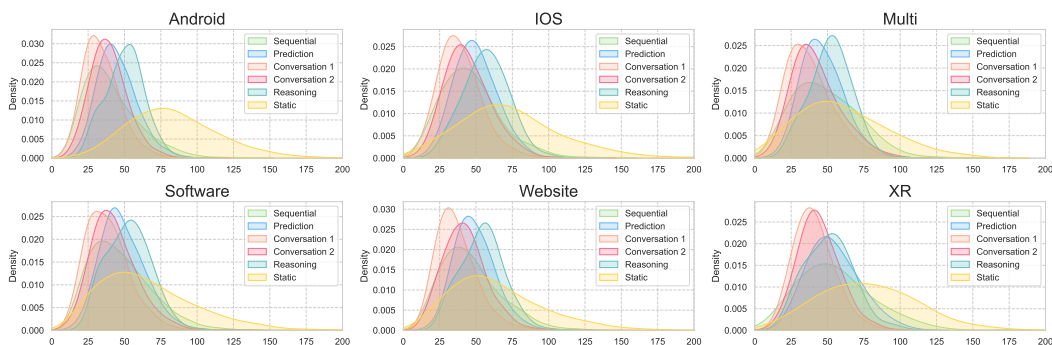


Figure 14: Length distribution of answers to free-form questions.

C.2 HYPERPARAMETER SETTINGS

In this section, we will introduce the hyperparameters of MLLMs to facilitate experiment reproducibility and transparency. We divide them into three parts: the inference phase during benchmark and dataset construction, the LLM-as-a-Judge phase, and the fine-tuning phase. All our experiments were conducted on a server equipped with dual A800 and dual 4090 GPUs.

Inference. We empirically study 7 MLLMs, involving 4 Image-LLMs and 3 Video-LLMs, with their hyperparameters detailed as follows:

- **GPT-4V (OpenAI, 2023) & GPT-4o (OpenAI, 2024a):** We set the temperature and top-p as 0.9, max-token as 2048, and both all images input are set as high quality in *Instruction Dataset Construction* and benchmarking.
- **Gemini-Pro-1.5 (GeminiTeam, 2023):** We use the default settings, which set temperature as 0.4, top-p as 1, and max-token as 2048. It should be noted that during our project, Gemini-Pro-1.5 is still under the user request limit, which only provides 100 requests per day, making our benchmark difficult. Given that Gemini hasn't launched Pay-as-you-go⁴, we will include benchmark results on 'Human' setting as soon as possible.
- **Qwen-VL-Max (Bai et al., 2023):** We use the default settings for Qwen-VL-Max, with top-p as 0.8 and max-token as 2048. Given that the input context window is merely 6,000 for Qwen, we scale the resolution for all images to 0.3.
- **ChatUnivi (Jin et al., 2023):** We use ChatUnivi-7B built upon Vicuna-v0-7B and set the max frame as 100, temperature as 0.2, and max-token as 1024.
- **Minigt4video (Atallah et al., 2024):** We use the suggested settings⁵ for this model and the max-frame are set as 45, with only the max-token being modified to 1024.

⁴<https://ai.google.dev/pricing>

⁵<https://github.com/Vision-CAIR/MiniGPT4-video>

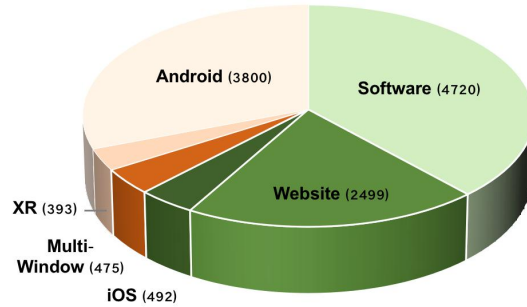


Figure 15: Statistic of different GUI scenarios in GUI-WORLD.

Table 12: Video-image fine-tune dataset.

Stage	Data types	Source	Type	Amount
1	GUI-WORLD	Video	Detailed Description	14,276
			Concise Caption	7,138
		Image	Detailed Description	5,555
			Concise Caption	5,555
	METAGUI	Image	Detailed Description	19,626
			Concise Caption	19,626
OmniAct	Image	Detailed Description	260	
		Concise Caption	260	
2	GUI-WORLD	Video	GUI VQA	21,414
			Multiple-Choice QA	14,276
			Conversation	7,138

- **VideoChat2 & GUI-Vid (Li et al., 2023c):** For a fair comparison, we set the same hyperparameters for VideoChat2 & GUI-Vid. We set the max-token as 1024, top-p as 0.9, temperature as 1.0, max-frame as 8/16, repetition penalty as 1.2, and length penalty as 1.2.

LLM-as-a-Judge. We investigate four LLM-as-a-Judge in giving a similarity score for the MLLM’s response and ground truth, namely GPT-4 (Team, 2024), ChatGPT (OpenAI, 2023), LLaMA-3-70b-instruct (Meta, 2023b), and Mixtral-8x22b-instruct-v0.1 (OpenAI, 2024b). Hyperparameter settings are detailed as follows:

- **GPT-4 & ChatGPT.** We set the temperature as 0.6 and others as default.
- **LLaMA-3-70b-instruct.** We set the temperature as 0.6, top-p as 0.9, top-k as 50.
- **Mixtral-8x22b-instruct-v0.1.** We set top-p as 0.7, top-k as 50, and temperature as 0.7.

Fine-tune. We include several hyperparameter settings in experiment settings and ablation studies, as shown in Table 13.

C.3 EVALUATION.

Given the complexity of free-form answers in GUI scenarios, the evaluation includes specific positions of GUI elements, textual content, and comparing the response to the golden answer. LLM-as-a-Judge has been widely used in previous studies for complex evaluation tasks (Zheng et al., 2023; Liu et al., 2023b). Therefore, we leverage LLM-as-a-Judge (Zheng et al., 2023) in a similar setting to MM-vet (Yu et al., 2023b), which compares the MLLM’s response to the golden answer. We carefully evaluate the accessibility of leveraging LLM-as-a-Judge, selecting 1,000 samples covering 6 free-form questions mentioned in our dataset. As shown in Table 14, GPT-4 outperforms other LLMs, exhibiting a better human alignment on providing a similarity score for the response compared to the golden answer, although it is approximately 10 times more expensive than other models.

Table 13: Configuration settings for fine-tuning.

Config	Setting
input frame	8
input resolution	224
max text length	512
input modal	I. + V.
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
weight decay	0.02
learning rate schedule	cosine decay
learning rate	$2e-5$
batch size	4
warmup epochs	0.6
total epochs	3
backbone drop path	0
QFormer drop path	0.1
QFormer dropout	0.1
QFormer token	96
flip augmentation	yes
augmentation	MultiScaleCrop [0.5, 1]

Table 14: Evaluating LLM-as-a-Judge as a replacement for human judging in the scoring setting.

Models	Pearson(\uparrow)	Spearman(\uparrow)	Kendall(\uparrow)	\$ per Benchmark(\downarrow)
GPT-4	0.856	0.853	0.793	120\$
ChatGPT	0.706	0.714	0.627	12\$
Llama-3-70b-instruct	0.774	0.772	0.684	12\$
Mixtral-8x22b-instruct-v0.1	0.759	0.760	0.670	15\$

D ADDITIONAL EXPERIMENTS RESULTS

In this section, we first provide an ablation study on keyframe selection methods. Then, we conduct statistics and human preference experiments on correlations of GUI understanding capability to other mainstream GUI-related tasks. Furthermore, we provide detailed, performance on newly released models after our submission of the first version, followed by very detailed results on each task in each GUI scenario.

Ablation study on keyframe identify methods. Firstly, we show performance on model-based keyframe identify methods in Table 18, with details of UVD+VIP and UVD+R3M in Table 19 and Table 20.

Correlation between GUI understanding and other mainstream GUI tasks. Furthermore, we conduct additional analysis and experiments to show how GUI understanding capability helps mainstream GUI-related tasks, including generating code to operate GUI (Cheng et al., 2024) and assist people through chat (Hong et al., 2023). Both demonstrate the strong correlation between GUI understanding capability and specific tasks for GUI agents.

- We compare the benchmark results on GUI-world with existing benchmarks (Xie et al., 2024; Qinghong Lin et al., 2024; Liu et al., 2024b) for operating on GUI as shown in Table 15, and find that the results generally match, i.e., the stronger the understanding ability, the stronger the agent performance.
- For the definition of chat helping humans, we select 180 videos from the benchmark, choosing 30 videos for each scenario. We ask 5 human annotators to pose the question they most wanted to ask after watching each video. We then use GUI-Vid, both before and after fine-tuning, to answer these questions. The human annotators who ask the questions are then asked to indicate which

Table 15: Strong Correlation Between Our Benchmark (GUI Understanding) and Other GUI Agent Benchmarks.

Model	GUI-World	VisualAgentBench	VideoGUI	OS-World
GPT-4o	1	1	1	2
GPT-4V	2	2	2	1
Gemini-1.5-Pro	3	3	3	3
Qwen-VL-Max	4	4	4	/

Table 16: The Performance of Video-LLaVA.

Scene	MCQA	Description	Conversation	Dynamic	Static	Caption	Average
XR	0.442	1.100	2.686	2.055	1.808	1.654	2.258
Android	0.513	1.162	2.952	1.858	1.673	1.763	2.259
IOS	0.497	1.143	2.966	1.992	1.680	1.654	2.319
Multi	0.459	1.106	2.863	2.069	1.781	1.772	2.329
Website	0.524	1.183	3.059	2.102	1.736	1.371	2.410
Software	0.529	1.241	2.942	1.958	1.657	1.519	2.290
Average	0.494	1.156	2.911	2.005	1.722	1.622	2.311

answer is more helpful. The results are shown in [Table 8](#), demonstrating that models trained in GUI understanding are more favored by people when acting as GUI agents.

Performance of newly released models in GUI-WORLD test set. We evaluate two latest models, LLaVA-Next-Video-7B-DPO ([Liu et al., 2024c](#)) and Video-LLaVA ([Lin et al., 2023a](#)). We show their performance in [Table 17](#) and [Table 16](#). Our model outperforms these in most tasks, except conversation, likely due to their use of DPO during training.

For captioning tasks, [Table 21](#) shows comprehensive experimental results among six scenarios. For scores of LLM-as-a-Judge in a specific task, see [Table 22](#), [Table 23](#), [Table 24](#), [Table 25](#), and [Table 26](#). For BLEU ([Papineni et al., 2002](#)) and BERTScore ([Zhang et al., 2019](#)) in validating free-form and conversational questions, see [Table 27](#), [Table 28](#), [Table 29](#), [Table 32](#), [Table 30](#), and [Table 31](#). For performance in fine-grain (application level), see [Figure 16](#) for Gemini-Pro and [Figure 17](#) for Qwen-VL-Max.

E PROMPTS

In this section, we provide detailed prompts for models and human annotators. [Figure 19](#) shows the guideline of human annotation, [Figure 18](#) shows the prompt for leveraging LLMs to refine grammarly mistakes and polish sentence for human annotations. [Figure 20](#), [Figure 21](#), and [Figure 22](#) present the prompt for Human-MLLM collaboration method to generate GUI-oriented tasks. [Figure 23](#) illustrate the prompt for benchmarking MLLMs, different GUI scenarios and different QA type has different prompt. [Figure 24](#) and [Figure 25](#) show prompt for LLM-as-a-Judge for free-form as well as conversational tasks and multiple-choice QA respectively.

Table 17: LLaVA-Next-Video-7B-DPO Performance

File	MCQA	Description	Conversation	Dynamic	Static	Caption	Average
XR	0.596	1.867	3.123	2.580	2.147	1.987	2.709
Android	0.243	1.675	3.338	2.360	1.980	2.189	2.675
IOS	0.581	1.762	3.229	2.536	2.051	2.017	2.717
Multi	0.355	1.069	2.982	2.437	1.870	2.541	2.541
Website	0.484	1.729	3.123	2.422	1.854	2.004	2.588
Software	0.569	1.762	3.220	2.448	1.868	2.149	2.641
Average	0.471	1.644	3.169	2.464	1.961	2.148	2.645

Table 18: Average Performance of GPT-4o among 6 Scenarios: Two Automated Keyframe Identification Methods vs. Human-Selected Keyframes

Setting	MCQA	Description	Conversation	Dynamic	Static	Caption	Average
Human	84.8%	3.031	4.056	3.318	3.131	3.911	3.573
UVD+vip	83.5%	3.150	4.044	3.265	3.346	3.923	3.581
UVD+r3m	84.5%	3.136	4.058	3.292	3.363	3.940	3.612

Table 19: Detailed Performance of GPT-4o using UVD+ViP Keyframe Identification Method.

Scenario	MCQA	Description	Conversation	Dynamic	Static	Caption	Average
Software	86.2%	3.297	4.282	3.354	3.478	4.112	3.749
Website	82.0%	3.248	4.155	3.415	3.567	4.074	3.744
XR	84.2%	2.980	3.775	3.034	3.122	3.587	3.347
Multi	82.1%	3.391	4.165	3.466	3.404	3.868	3.659
IOS	86.0%	3.157	4.017	3.353	3.492	4.050	3.648
Mobile	80.7%	2.827	3.871	2.970	3.014	3.844	3.340
Average	83.5%	3.150	4.044	3.265	3.346	3.923	3.581

Table 20: Detailed Performance of GPT-4o using UVD+R3M Keyframe Identification Method.

Scenario	MCQA	Description	Conversation	Dynamic	Static	Caption	Average
Software	85.8%	3.290	4.273	3.352	3.458	4.134	3.741
Website	82.7%	3.282	4.114	3.460	3.591	4.065	3.746
XR	87.7%	3.010	3.861	3.142	3.161	3.600	3.433
Multi	83.6%	3.237	4.129	3.503	3.417	3.897	3.737
IOS	86.4%	3.165	4.094	3.328	3.480	4.078	3.663
Android	80.6%	2.835	3.876	2.968	3.072	3.865	3.353
Average	84.5%	3.136	4.058	3.292	3.363	3.940	3.612

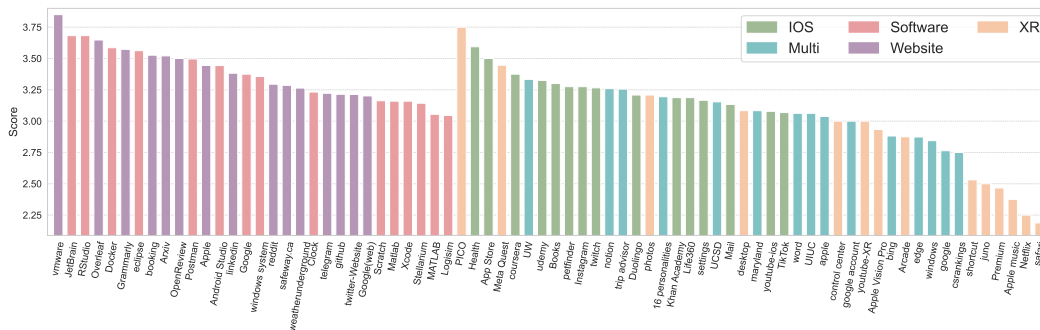


Figure 16: Fine-grained performance of Gemini-Pro-1.5 in each software and website.

Table 21: Scores of Caption (Cap.) and Description (Des.) tasks in six GUI scenarios.

Models	Setting	Software		Website		XR		Multi		IOS		Android		Avg.	
		Cap.	Des.	Cap.	Des.	Cap.	Des.	Cap.	Des.	Cap.	Des.	Cap.	Des.	Cap.	Des.
Gemini-Pro-1.5	R.	3.659	2.837	3.613	2.860	2.995	2.590	3.276	2.470	3.678	2.936	-	-	3.444	2.739
	E.	3.350	2.468	3.159	2.422	2.837	2.279	2.824	2.109	3.394	2.519	3.185	2.312	3.125	2.351
Qwen-VL-Max	R.	2.381	1.758	2.326	1.681	2.172	1.772	2.035	1.463	2.513	1.662	2.141	1.565	2.261	1.650
	E.	2.459	1.693	2.317	1.599	2.167	1.638	2.190	1.438	2.189	1.615	2.002	1.429	2.221	1.569
	H.	2.474	1.711	2.457	1.698	2.383	1.777	1.910	1.346	2.577	1.795	2.474	1.711	2.360	1.665
GPT-4V	R.	3.579	2.676	3.612	2.699	2.975	2.525	3.281	2.661	3.757	2.775	3.655	2.755	3.479	2.682
	E.	3.141	2.301	3.293	2.380	2.471	2.085	3.063	2.324	3.624	2.611	3.201	2.312	3.132	2.335
	H.	3.352	2.509	3.702	2.750	3.050	3.556	3.524	2.673	3.670	2.588	-	-	3.460	2.614
GPT-4o	H.	4.048	3.028	4.067	3.233	3.398	2.729	3.869	3.111	4.014	2.993	4.071	3.095	3.911	3.869
ChatUnivi	-	1.587	1.240	1.569	1.254	1.417	1.148	1.575	1.267	1.480	1.146	1.778	1.249	1.568	1.217
Minigpt4Video	-	1.246	1.073	1.200	1.057	1.320	1.106	1.130	1.034	1.190	1.076	1.184	1.061	1.212	1.068
VideoChat2	-	1.992	1.312	1.817	1.307	1.838	1.426	2.222	1.433	2.169	1.270	2.119	1.294	1.900	1.340
GUI-Vid	-	3.562	2.085	3.655	2.167	3.747	2.153	3.370	1.742	3.566	2.071	2.662	1.248	3.427	1.911

Table 22: Detailed scores for each tasks in **Website** scenarios.

Models	Setting	Static	Sequential	Prediction	Conversation1	Conversation2	Average
Gemini-Pro-1.5	R.	3.279	3.050	3.560	3.579	3.796	3.452
	E.	2.983	2.491	3.432	3.405	3.760	3.215
Qwen-VL-Max	R.	2.317	2.271	2.802	2.995	3.069	2.656
	E.	2.256	2.198	2.821	2.861	3.144	2.627
	H.	2.308	2.078	2.832	3.061	3.358	2.698
GPT-4V	R.	3.461	3.214	3.754	3.778	4.029	3.648
	E.	3.197	2.808	3.487	3.717	3.954	3.433
	H.	3.498	3.255	3.727	3.731	4.061	3.655
	C.C.	1.746	2.738	3.645	3.363	3.632	3.025
	D.C.	2.704	2.917	3.686	3.680	3.901	3.380
	H.+D.C.	3.313	3.221	3.852	3.850	4.171	3.682
GPT-4o	H.	3.443	3.373	3.672	4.086	4.122	3.740
ChatUnivi	-	1.701	1.668	2.524	2.514	3.338	2.349
Minigpt4Video	-	1.309	1.233	1.766	1.439	1.854	1.520
VideoChat2	-	1.771	1.777	2.288	2.461	2.812	2.221
GUI-Vid	-	2.406	2.341	3.544	3.135	3.355	2.957

Table 23: Detailed scores for each tasks in **XR** scenarios.

Models	Setting	Static	Sequential	Prediction	Conversation1	Conversation2	Average
Gemini-Pro-1.5	R.	2.892	2.505	3.543	3.222	3.611	3.154
	E.	2.814	2.163	3.510	3.108	3.455	3.006
Qwen-VL-Max	R.	2.047	1.968	2.712	2.879	3.132	2.469
	E.	2.125	1.973	2.658	2.760	3.029	2.499
	H.	1.886	1.920	2.656	2.727	3.012	2.373
GPT-4V	R.	2.934	2.668	3.392	3.291	3.714	3.200
	E.	2.222	2.153	3.310	3.151	3.618	2.892
	H.	2.893	2.778	3.538	3.364	3.747	3.265
	C.C.	1.744	2.412	3.327	3.080	3.485	2.809
	D.C.	2.427	2.409	3.518	3.176	3.749	3.056
	H.+D.C.	2.775	2.635	3.580	3.235	3.734	3.191
GPT-4o	H.	2.871	2.745	3.370	3.596	3.836	3.285
ChatUnivi	-	1.660	1.420	2.205	2.250	3.270	2.161
Minigpt4Video	-	1.225	1.161	1.610	1.347	1.465	1.362
VideoChat2	-	1.654	1.547	2.192	2.099	2.529	2.005
GUI-Vid	-	2.444	2.147	3.347	2.836	3.036	2.764

Table 24: Detailed scores for each tasks in **Multi-windows** scenarios.

Models	Setting	Static	Sequential	Prediction	Conversation1	Conversation2	Average
Gemini-Pro-1.5	R.	2.538	2.410	3.296	3.152	3.402	2.959
	E.	2.545	2.049	2.972	2.930	3.389	2.777
Qwen-VL-Max	R.	1.793	1.872	2.770	2.897	3.122	2.432
	E.	1.866	1.780	2.730	2.627	3.105	2.362
	H.	1.884	1.969	2.913	2.689	3.104	2.490
GPT-4V	R.	3.185	2.655	3.745	3.699	3.973	3.452
	E.	2.902	2.406	3.636	3.420	3.729	3.219
	H.	3.000	2.952	3.801	3.597	3.889	3.449
	C.C.	2.097	2.973	3.774	3.331	3.621	3.160
	D.C.	2.671	2.979	3.849	3.466	3.822	3.358
	H.+D.C.	3.037	3.162	4.079	3.748	4.036	3.617
GPT-4o	H.	3.108	3.106	3.829	4.043	4.188	3.654
ChatUnivi	-	1.658	1.623	2.514	2.384	3.199	2.275
Minigpt4Video	-	1.205	1.186	1.690	1.400	1.801	1.457
VideoChat2	-	1.754	1.774	2.479	2.420	2.699	2.222
GUI-Vid	-	2.485	2.067	3.537	2.954	3.247	2.861

Table 25: Detailed scores for each tasks in **IOS** scenarios.

Models	Setting	Static	Sequential	Prediction	Conversation1	Conversation2	Average
Gemini-Pro-1.5	R.	3.076	2.637	3.370	3.366	3.615	3.213
	E.	2.852	2.356	3.137	3.126	3.566	3.007
Qwen-VL-Max	R.	2.438	2.244	2.923	3.102	3.273	2.779
	E.	2.303	2.150	2.614	3.145	3.264	2.659
	H.	1.884	1.969	2.913	2.689	3.104	2.490
GPT-4V	R.	3.364	3.080	3.684	3.766	4.184	3.614
	E.	3.209	2.774	3.545	3.611	4.006	3.427
	H.	3.107	2.830	3.631	3.680	4.011	3.453
	C.C.	1.788	2.291	3.511	3.212	3.542	2.868
	D.C.	2.751	2.732	3.654	3.642	3.842	3.324
	H.+D.C.	3.090	2.965	3.740	3.786	3.994	3.516
GPT-4o	H.	3.183	2.993	3.460	4.050	4.141	3.558
ChatUnivi	-	1.771	1.642	2.408	2.559	3.307	2.337
Minigpt4Video	-	1.291	1.219	1.698	1.556	1.737	1.501
VideoChat2	-	1.955	1.803	2.145	2.315	2.626	2.169
GUI-Vid	-	2.262	2.133	3.401	2.843	3.224	2.773

Table 26: Detailed scores for each tasks in **Android** scenarios.

Models	Setting	Static	Sequential	Prediction	Conversation1	Conversation2	Average
Gemini-Pro-1.5	E.	2.703	2.460	3.157	3.642	3.881	3.168
Qwen-VL-Max	R.	1.887	1.804	2.398	2.823	3.056	2.309
	E.	1.785	1.630	2.311	2.605	3.233	2.277
GPT-4V	R.	3.116	3.047	3.477	3.924	4.008	3.515
	E.	2.705	2.470	3.175	3.647	3.885	3.176
	C.C.	2.092	2.243	3.139	3.443	3.782	2.939
	D.C.	3.015	2.890	3.357	3.883	3.990	3.427
GPT-4o	H.	3.057	3.220	3.373	3.981	4.186	3.561
ChatUnivi	-	1.835	1.654	2.317	2.712	3.433	2.390
Minigpt4Video	-	1.183	1.159	1.507	1.342	1.521	1.342
VideoChat2	-	1.732	1.754	2.125	2.340	2.645	2.119
GUI-Vid	-	2.010	1.928	3.053	2.755	3.105	2.572

Table 27: Detailed BLEU and BERTScore (B.S.) in **Software** scenarios.

Models	Setting	Static		Sequential		Prediction		Description		Caption		Conversation		Avg.	
		BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.
Gemini-Pro-1.5	R.	0.109	0.789	0.150	0.720	0.078	0.680	0.056	0.716	0.016	0.605	0.122	0.761	0.089	0.712
	E.	0.093	0.758	0.134	0.699	0.072	0.659	0.046	0.682	0.011	0.558	0.106	0.747	0.077	0.684
Qwen-VL-Max	R.	0.085	0.698	0.101	0.649	0.064	0.576	0.010	0.521	0.008	0.443	0.121	0.749	0.065	0.606
	E.	0.094	0.704	0.103	0.633	0.062	0.595	0.009	0.524	0.006	0.437	0.113	0.739	0.065	0.605
	H.	0.081	0.676	0.098	0.620	0.067	0.596	0.009	0.504	0.004	0.429	0.117	0.743	0.063	0.595
GPT-4V	R.	0.162	0.814	0.206	0.753	0.190	0.739	0.041	0.676	0.033	0.581	0.181	0.793	0.136	0.726
	E.	0.161	0.792	0.191	0.726	0.175	0.724	0.030	0.609	0.017	0.486	0.165	0.786	0.123	0.687
	H.	0.153	0.805	0.194	0.737	0.183	0.731	0.037	0.639	0.025	0.537	0.179	0.791	0.129	0.707
GPT-4o	H.	0.131	0.806	0.212	0.776	0.147	0.728	0.041	0.711	0.018	0.575	0.159	0.803	0.118	0.733
ChatUnivi	-	0.097	0.697	0.074	0.581	0.101	0.619	0.005	0.409	0.000	0.195	0.084	0.723	0.060	0.537
Minigt4Video	-	0.019	0.516	0.022	0.470	0.029	0.516	0.000	0.399	0.000	0.249	0.013	0.510	0.014	0.443
VideoChat2	-	0.095	0.698	0.080	0.595	0.076	0.574	0.004	0.341	0.000	0.193	0.100	0.733	0.059	0.523
GUI-Vid	-	0.142	0.758	0.145	0.681	0.114	0.698	0.049	0.658	0.004	0.519	0.093	0.717	0.091	0.672

Table 28: Detailed BLEU and BERTScore (B.S.) in **Website** scenarios.

Models	Setting	Static		Sequential		Prediction		Description		Caption		Conversation		Avg.	
		BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.
Gemini-Pro-1.5	R.	0.113	0.793	0.145	0.727	0.083	0.676	0.054	0.720	0.016	0.664	0.098	0.736	0.085	0.719
	E.	0.095	0.754	0.121	0.681	0.079	0.661	0.041	0.676	0.011	0.602	0.092	0.725	0.073	0.683
Qwen-VL-Max	R.	0.099	0.728	0.099	0.634	0.080	0.610	0.008	0.519	0.005	0.471	0.085	0.694	0.063	0.609
	E.	0.083	0.710	0.101	0.631	0.093	0.611	0.011	0.503	0.004	0.469	0.099	0.709	0.065	0.605
	H.	0.079	0.693	0.089	0.597	0.093	0.606	0.009	0.488	0.007	0.449	0.103	0.705	0.063	0.590
GPT-4V	R.	0.173	0.830	0.241	0.765	0.205	0.751	0.040	0.694	0.032	0.645	0.164	0.763	0.142	0.741
	E.	0.159	0.802	0.204	0.727	0.202	0.727	0.033	0.648	0.031	0.590	0.149	0.757	0.130	0.708
	H.	0.182	0.823	0.234	0.771	0.213	0.758	0.043	0.696	0.041	0.660	0.165	0.768	0.147	0.746
GPT-4o	H.	0.141	0.813	0.219	0.768	0.199	0.731	0.054	0.700	0.026	0.602	0.146	0.755	0.131	0.728
ChatUnivi	-	0.078	0.645	0.068	0.581	0.102	0.607	0.008	0.399	0.000	0.192	0.061	0.661	0.053	0.514
Minigt4Video	-	0.022	0.527	0.016	0.448	0.027	0.501	0.000	0.344	0.000	0.186	0.011	0.522	0.013	0.421
VideoChat2	-	0.073	0.619	0.075	0.579	0.049	0.511	0.004	0.328	0.000	0.167	0.067	0.678	0.045	0.480
GUI-Vid	-	0.114	0.731	0.158	0.674	0.129	0.694	0.049	0.667	0.002	0.553	0.075	0.681	0.088	0.667

Table 29: Detailed BLEU and BERTScore (B.S.) in **XR** scenarios.

Models	Setting	Static		Sequential		Prediction		Description		Caption		Conversation		Avg.	
		BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.
Gemini-Pro-1.5	R.	0.088	0.772	0.101	0.678	0.070	0.678	0.026	0.650	0.002	0.463	0.082	0.733	0.062	0.662
	E.	0.073	0.760	0.090	0.651	0.062	0.666	0.015	0.618	0.002	0.449	0.084	0.720	0.054	0.644
Qwen-VL-Max	R.	0.069	0.703	0.075	0.602	0.049	0.601	0.006	0.486	0.000	0.338	0.117	0.738	0.053	0.578
	E.	0.048	0.689	0.079	0.657	0.058	0.605	0.005	0.498	0.000	0.359	0.112	0.739	0.050	0.591
	H.	0.051	0.651	0.073	0.593	0.044	0.591	0.004	0.493	0.001	0.357	0.101	0.726	0.046	0.569
GPT-4V	R.	0.093	0.794	0.169	0.715	0.165	0.736	0.028	0.625	0.006	0.457	0.147	0.768	0.101	0.683
	E.	0.085	0.726	0.131	0.665	0.162	0.724	0.020	0.541	0.003	0.382	0.141	0.760	0.090	0.633
	H.	0.091	0.797	0.181	0.732	0.180	0.744	0.027	0.630	0.006	0.471	0.154	0.773	0.106	0.691
GPT-4o	H.	0.077	0.800	0.154	0.717	0.153	0.718	0.020	0.615	0.006	0.468	0.138	0.759	0.091	0.680
ChatUnivi	-	0.083	0.686	0.061	0.538	0.091	0.575	0.006	0.475	0.000	0.282	0.086	0.693	0.054	0.541
Minigt4Video	-	0.014	0.545	0.016	0.466	0.027	0.502	0.001	0.453	0.000	0.262	0.013	0.474	0.012	0.450
VideoChat2	-	0.077	0.679	0.079	0.595	0.073	0.577	0.004	0.378	0.000	0.211	0.101	0.721	0.056	0.527
GUI-Vid	-	0.096	0.754	0.149	0.689	0.131	0.700	0.051	0.637	0.003	0.460	0.082	0.705	0.085	0.657

Table 30: Detailed BLEU and BERTScore (B.S.) in **IOS** scenarios.

Models	Setting	Static		Sequential		Prediction		Description		Caption		Conversation		Avg.	
		BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.
Gemini-Pro-1.5	R.	0.108	0.797	0.142	0.717	0.080	0.682	0.075	0.714	0.011	0.602	0.117	0.746	0.089	0.710
	E.	0.099	0.768	0.136	0.700	0.075	0.655	0.066	0.695	0.011	0.592	0.113	0.743	0.083	0.692
Qwen-VL-Max	R.	0.087	0.704	0.098	0.650	0.112	0.639	0.009	0.519	0.003	0.465	0.106	0.725	0.069	0.617
	E.	0.075	0.638	0.095	0.647	0.094	0.600	0.009	0.512	0.009	0.475	0.103	0.712	0.064	0.597
	H.	0.080	0.632	0.083	0.589	0.092	0.617	0.013	0.520	0.007	0.452	0.099	0.703	0.062	0.585
GPT-4V	R.	0.159	0.824	0.224	0.772	0.206	0.766	0.040	0.673	0.030	0.579	0.174	0.777	0.139	0.732
	E.	0.149	0.813	0.201	0.752	0.207	0.746	0.035	0.659	0.017	0.566	0.160	0.762	0.128	0.716
	H.	0.156	0.805	0.205	0.745	0.203	0.748	0.034	0.644	0.025	0.559	0.159	0.763	0.130	0.711
GPT-4o	H.	0.137	0.802	0.196	0.761	0.199	0.732	0.035	0.683	0.022	0.533	0.154	0.774	0.124	0.714
ChatUnivi	-	0.093	0.679	0.085	0.604	0.106	0.616	0.005	0.437	0.000	0.258	0.076	0.698	0.061	0.548
Minigt4Video	-	0.026	0.547	0.026	0.513	0.035	0.548	0.001	0.411	0.000	0.236	0.015	0.529	0.017	0.464
VideoChat2	-	0.089	0.683	0.078	0.605	0.061	0.555	0.002	0.355	0.000	0.190	0.086	0.710	0.053	0.516
GUI-Vid	-	0.114	0.725	0.144	0.693	0.123	0.700	0.048	0.641	0.002	0.518	0.083	0.686	0.085	0.661

Table 31: Detailed BLEU and BERTScore (B.S.) in **Android** scenarios.

Models	Setting	Static		Sequential		Prediction		Description		Caption		Conversation		Avg.	
		BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.
Gemini-Pro-1.5	E.	0.089	0.771	0.189	0.704	0.189	0.710	0.023	0.619	0.016	0.570	0.149	0.749	0.109	0.687
Qwen-VL-Max	R.	0.041	0.640	0.084	0.528	0.066	0.549	0.008	0.484	0.004	0.445	0.089	0.673	0.049	0.553
	E.	0.037	0.634	0.074	0.498	0.065	0.541	0.005	0.443	0.003	0.383	0.089	0.683	0.045	0.530
GPT-4V	R.	0.106	0.809	0.242	0.757	0.210	0.733	0.029	0.653	0.028	0.619	0.170	0.763	0.131	0.723
	E.	0.089	0.771	0.192	0.705	0.190	0.713	0.023	0.619	0.016	0.571	0.150	0.750	0.110	0.688
GPT-4o	H.	0.075	0.809	0.241	0.755	0.188	0.719	0.038	0.677	0.014	0.581	0.137	0.747	0.116	0.715
ChatUnivi	-	0.076	0.675	0.079	0.588	0.096	0.594	0.007	0.482	0.001	0.368	0.063	0.670	0.054	0.563
Minigt4Video	-	0.017	0.416	0.013	0.369	0.019	0.405	0.000	0.279	0.000	0.103	0.010	0.392	0.010	0.327
VideoChat2	-	0.057	0.641	0.077	0.560	0.063	0.523	0.004	0.402	0.000	0.272	0.075	0.654	0.046	0.509
GUI-Vid	-	0.083	0.682	0.130	0.628	0.126	0.644	0.023	0.500	0.001	0.393	0.071	0.659	0.072	0.584

Table 32: Detailed BLEU and BERTScore (B.S.) in **Multiple-windows** scenarios.

Models	Setting	Static		Sequential		Prediction		Description		Caption		Conversation		Avg.	
		BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.	BLEU	B.S.
Gemini-Pro-1.5	R.	0.113	0.739	0.126	0.693	0.086	0.658	0.061	0.685	0.012	0.586	0.090	0.674	0.081	0.673
	E.	0.106	0.728	0.131	0.680	0.072	0.622	0.055	0.655	0.015	0.550	0.084	0.679	0.077	0.652
Qwen-VL-Max	R.	0.079	0.599	0.076	0.591	0.080	0.595	0.002	0.444	0.006	0.370	0.072	0.666	0.053	0.544
	E.	0.064	0.609	0.087	0.567	0.089	0.608	0.003	0.445	0.004	0.398	0.073	0.647	0.053	0.546
	H.	0.089	0.634	0.078	0.580	0.093	0.612	0.003	0.409	0.005	0.344	0.080	0.656	0.058	0.539
GPT-4V	R.	0.172	0.800	0.186	0.737	0.212	0.745	0.040	0.671	0.021	0.592	0.145	0.728	0.129	0.712
	E.	0.160	0.763	0.169	0.703	0.198	0.759	0.034	0.621	0.012	0.527	0.116	0.709	0.115	0.680
	H.	0.173	0.781	0.196	0.748	0.220	0.775	0.046	0.672	0.021	0.577	0.133	0.724	0.132	0.713
GPT-4o	H.	0.156	0.792	0.185	0.754	0.213	0.769	0.040	0.683	0.019	0.588	0.121	0.717	0.122	0.717
ChatUnivi	-	0.076	0.628	0.063	0.573	0.103	0.605	0.009	0.413	0.000	0.191	0.057	0.643	0.051	0.509
Minigt4Video	-	0.015	0.504	0.024	0.473	0.023	0.527	0.001	0.326	0.000	0.155	0.009	0.469	0.012	0.409
VideoChat2	-	0.098	0.657	0.081	0.593	0.067	0.577	0.007	0.344	0.000	0.162	0.065	0.654	0.053	0.498
GUI-Vid	-	0.128	0.737	0.144	0.664	0.133	0.721	0.041	0.605	0.004	0.452	0.058	0.644	0.084	0.637

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

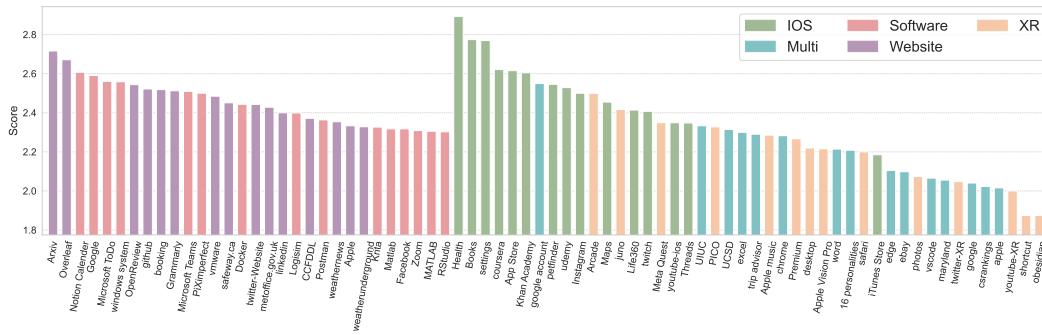


Figure 17: Fine-grained performance of Qwen-VL-Max in each software and website.

Refining Human Annotation on Goal and Sub-goal

As an expert in English, please refine the following English instructions (or objectives) into a polished phrase or a concise sentence.
Avoid including irrelevant content and provide the polished output directly.
Here is the English sentence: {string}

Figure 18: Refining Human Annotation on Goal and Sub-goal.

1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835

Guideline for Human Annotation

Main Interface

1. Video List Panel (Left Panel): Displays a list of loaded video files. Each video file is shown with its name for identification.
2. Video Display Area (Center Panel): Shows the currently selected video for playback and annotation.
3. Control Settings (Right Panel):
Operating System: Select the operating system of the machine where the video was recorded.
Full Screen: Toggle full screen mode for the video display.
Multi-application?: Indicate if multiple applications in the video.
Application/Website: Enter the name of the application or website being used in the video.
User Goal: Enter the goal of the user performing the annotation.
4. Playback and Annotation Controls (Bottom Panel)
Annotate: Open a annotation window to add a new keyframe annotation.
Play: Starts or pauses the video playback.
Load Video: Allows you to load a single video file.
Load Video Folder: Allows to load multiple video files from a folder.
Previous Video / Next Video: Navigate through the loaded video files.
Save to JSON: Save the annotations in a JSON format.

Annotation Window

1. Mouse Action: Select a type of mouse action (e.g. click, drag).
2. Keyboard Action: Select the type of keyboard action (e.g., typing, key press).
3. Keyboard Operation Record: Enter details of the keyboard operation, if any.
4. sub-action Purpose: Describe the purpose of the action being annotated.

How to Use

Loading Videos

1. Load Multiple Videos
Click on the Load Video Folder button.
Select the folder containing your video files.
All video files in the folder will be loaded and listed in the Video List Panel.

Playing Videos

Select a video from the Video List Panel. Click the Play button to start or pause the video.

Annotating Videos

1. Start Annotation
Pause the video at the desired frame.
Click the Annotate button to open the annotation window.
2. Annotation Window
Select the Mouse Action Type and Keyboard Action Type from the dropdown menus.
If there is a keyboard action, enter the details in the Keyboard Operation Record field.
Describe the action's purpose in the Sub-action Purpose field.
Click OK to save the annotation.

Saving Annotations

Once all annotations are completed, click the Save to JSON button.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

(Part 1) GPT-4V Generating GUI-oriented Tasks

You are an AI visual assistant. This is a video of a mobile GUI, which I've divided into multiple frames and sent to you. Please provide a detailed description of what occurs throughout the entire video, focusing on the changes in the GUI elements or scenes rather than static aspects of a single frame. The detailed description should be placed under the key 'Description'. Based on your description, please design the following tasks:

Generate a precise caption for the video. This caption should encapsulate the main activities or changes observed throughout the video sequence. Place this caption under the key 'Caption'.

Create a free-form QA question related to the video's static GUI content, along with its answer. The question should delve into the details or changes in the static GUI elements or scenes captured in the video. The QA task should be nested under the key 'static QA', with 'Question' and 'Answer' as subkeys.

Develop a multiple-choice QA question about the video, with four options: one correct answer and three incorrect or irrelevant options. This task should assess the understanding of specific elements retrieval or changes depicted in the video. Structure this task under the key 'MCQA', with 'Question' detailing the query, 'Options' listing the four choices including one correct answer, and 'Correct Answer' specifying the correct option, denoted, for example, as `{[[B]]}`.

Here are some key information of the video to help you understand the video comprehensively:

```
System: {item['system']}
Application: {item['app']}
Summary of the video: {item['goal']}
Key Operation/Sub goal in the video: {[i['sub_goal'] for i in item['keyframes']]}
```

Notice: Ensure that the questions you design for these tasks are answerable and the answers can be deduced from the GUI video content. The answerable question should be designed as difficult as possible. The tasks should be unambiguous and the answers must be definitively correct based on your understanding of the video content. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Each of these tasks should focus on the dynamic aspect of the GUI elements or scenes. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question.

For the free-form QA tasks, please ensure that the answers are as detailed and lengthy as possible, with no concern for length. You can include multiple paragraphs if necessary to provide a comprehensive and thorough response. Please structure your response using JSON format and specific keys mentioned in the task requirements.

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

(Part 2) GPT-4V Generating GUI-oriented Tasks.

You are an AI visual assistant. This is a video of a <Scene Name> GUI, which I've divided into multiple frames and sent to you. Please provide a detailed description of what occurs throughout the entire video, focusing on the changes in the GUI elements or scenes rather than static aspects of a single frame. The detailed description should be placed under the key 'Description'. Based on your description, please design the following tasks:

A Sequential QA task: Design a question that requires understanding the sequence of GUI element changes or scene transformations in the video. The question should be free-form and necessitate the use of temporal information from the sequential images. The task should be structured under the key 'Sequential-QA' with subkeys 'Question' and 'Answer'.

A Next Stage Prediction task: Formulate a question that asks about the subsequent state or event following a certain frame in the video. The question should be designed in a free-form manner and predict future GUI elements or scene changes, structured under the key 'Prediction' with subkeys 'Question' and 'Answer'.

A two-round dialogue task: Create a dialogue with two rounds of interaction. The first round includes a user instruction and an assistant response, and the second round's user instruction should be based on the response from the first round. Both rounds should be free-form and nested under the key 'Conversation', with subkeys 'User 1', 'Assistant 1', 'User 2', and 'Assistant 2'.

A reasoning task: Design a multi-choice QA task that requires reasoning to identify the correct answer from four options. This task should test the reasoning ability to infer or deduce information that is not explicitly provided. It should be structured under the key 'Reasoning', with subkeys 'Question', 'Options', and 'Correct Answer'.

Here are some key information of the video to help you understand the video comprehensively:

System: {item['system']}

Application: {item['app']}

Summary of the video: {item['goal']}

Key Operation/Sub goal in the video: {[i['sub_goal'] for i in item['keyframes']]}

Figure 21: (Part 2) GPT-4V Generating GUI-oriented Tasks.

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

(Part 3) GPT-4V Generating GUI-oriented Tasks.

Notice: Ensure that the questions you design for these tasks are answerable and the answers can be deduced from the GUI video content. The answerable question should be designed as difficult as possible. The tasks should be unambiguous and the answers must be definitively correct based on your understanding of the video content. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently. Each of these tasks should focus on the dynamic aspect of the GUI elements or scenes, with each answerable task as difficult as possible. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. For the free-form QA tasks, please ensure that the answers are as detailed and lengthy as possible, with no concern for length. You can include multiple paragraphs if necessary to provide a comprehensive and thorough response. Please structure your response using JSON format and specific keys mentioned in the task requirements.

Figure 22: (Part 3) GPT-4V Generating GUI-oriented Tasks.

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

Prompts for Benchmarking MLLMs

"XR": "You are an AI visual assistant. Here are sequential images of Mixed-Reality combining GUI interface and real world, which are selected from a GUI video.",
 "software": "You are an AI visual assistant. Here are sequential GUI interface images of a specific software, which are selected from a GUI video.",
 "website": "You are an AI visual assistant. Here are sequential GUI interface images of a desktop website, which are selected from a GUI video.",
 "mobile": "You are an AI visual assistant. Here are sequential GUI mobile interface images, which are selected from a GUI video.",
 "multi": "You are an AI visual assistant. Here are sequential GUI interface images of interaction among multiple softwares and websites, which are selected from a GUI video.",
 "IOS": "You are an AI visual assistant. Here are sequential GUI IOS interface images, which are selected from a GUI video.",

 "Sequential-QA": "This is a question about sequential information in sequential images.",
 "Prediction": "This is a question about predicting the next action base on the previous actions in the sequential images.",
 "Reasoning": "This is a multiple choice question with only one correct answer. This question may need multiple steps of reasoning according to the vision information in sequential images.",
 "Description1": "Please give me a detail description of these sequential images.",
 "Description2": "Offer a thorough analysis of these sequential images",
 "Caption": "Please give me a concise caption of these sequential images.",
 "static QA": "This is a question about static information such as text, icon, layout in these sequential images.",
 "MCQA": "This is a multiple choice question with only one correct answer. This question may require sequential analysis ability to the vision information in these sequential images.",
 "Conversation1": "Act as an assistant to answer the user's question in these sequential images.",
 "Conversation2": "This is a multi-turn conversation task. You will be provide the first round conversation and act as an assistant to answer the user's question in the second round according to these sequential images."
 Notice = "You can first provide an overall description of these sequential images, and then analyze the user's question according to the sequential images and description. Finally, give an answer based on this description and the image information. Please format your output in a Json format, with key 'Description' for the description of these sequential images, key 'Analysis' for your analysis on the user's question and key 'Answer' for your answer to the User's question."

Figure 23: Prompts for Benchmarking MLLMs.

2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105

Prompt for LLM-as-a-Judge: Judging Free-form and Conversational Tasks

You are an impartial judge. I will provide you with a question, a 'gold standard' answer, and a response that needs evaluation. Your task is to assess the quality of the response in comparison to the 'gold standard' answer. Please adhere to the following guidelines:

1. Start your evaluation by comparing the response to the 'gold standard' answer. Offer a brief explanation highlighting similarities and differences, focusing on relevance, accuracy, depth, and level of detail.
2. Conclude your evaluation with a score from 1 to 5, where 1 indicates the response is mostly irrelevant to the 'gold standard' answer, and 5 indicates it is very similar or equivalent.
3. Present your findings in JSON format, using 'Evaluation' for your textual analysis and 'Score' for the numerical assessment.
4. Ensure objectivity in your evaluation. Avoid biases and strive for an even distribution of scores across the spectrum of quality. Your scoring must be as rigorous as possible and adhere to the following rules:
 - Overall, the higher the quality of the model's response, the higher the score, with factual accuracy and meeting user needs being the most critical dimensions. These two factors largely dictate the final composite score.
 - If the model's response is irrelevant to the question, contains fundamental factual errors, or generates harmful content, the total score must be 1.
 - If the model's response has no severe errors and is essentially harmless, but of low quality and does not meet user needs, the total score should be 2.
 - If the model's response generally meets user requirements but performs poorly in certain aspects with medium quality, the total score should be 3.
 - If the model's response is close in quality to the reference answer and performs well in all dimensions, the total score should be 4.
 - Only when the model's response surpasses the reference answer, fully addresses the user's problem and all needs, and nearly achieves a perfect score in all dimensions, can it receive a score between 5.
 - As an example, the golden answer could receive a 4-5.

Here is the response for you to judge:

Question: {question}

Golden Answer: {golden_answer}

Response: {response}

Now, directly output your response in json format.

Figure 24: Prompt for LLM-as-a-Judge: Judging Free-form and Conversational Tasks .

2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159

Prompt for LLM-as-a-Judge: Judging Multiple-Choice QA Tasks

You are a helpful assistant tasked with judging a Multiple Choice Question Answering exercise. I will provide a correct answer with only one option, and a response that requires evaluation. If the response matches the correct answer, simply output "Yes"; If it does not, output "No". Please avoid including any irrelevant information. Here are some examples:

Example 1:

Question: Based on the GUI video, why might the 'Loading' animation continue without reaching the next stage? A. The user has not yet entered their login credentials. B. There is a system update being installed. C. The server is taking time to authenticate the login credentials. D. The 'Log In' button is malfunctioning.

Answer: C

Response: C. The server is taking time to authenticate the login credentials.

Output: Yes

Example 2:

Question: If the user wants to resume the group video call after checking messages, what action should they take? A. Turn their head to the right. B. Close the messaging app interface. C. Say a voice command to switch applications. D. Turn their head to the left.

Answer: A

Response: B

Output: No

Example 3:

Question: What action does the user take to start playing music in the video? A. Closed the music player application B. Moved the music player to a new position C. Clicked the play button D. Adjusted the system volume

Answer: [[B]]

Response: C

Output: No

Here is the question, answer, and response for you to judge:

Question: {question}

Answer: {answer}

Response: {response}

Now, directly output "Yes" or "No".

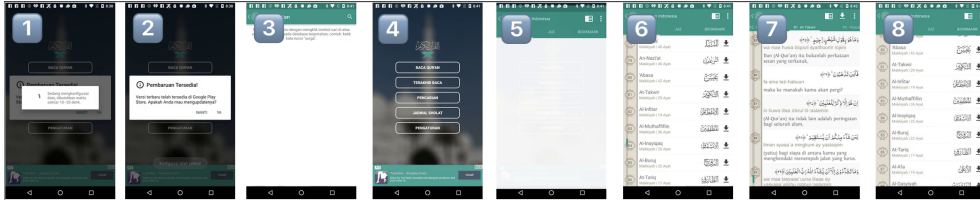
Figure 25: Prompt for LLM-as-a-Judge: Judging Multiple-Choice QA Tasks.

2160 F CASE STUDY
2161

2162 In this section, we provide detailed case studies for six GUI scenarios, each divided into two parts.
2163 [Figure 26](#) and [Figure 27](#) show example frames and various tasks associated with them. [Figure 28](#)
2164 and [Figure 29](#) for IOS, [Figure 30](#) and [Figure 31](#) for multiple-windows interaction, [Figure 34](#) and
2165 [Figure 35](#) for website, and [Figure 36](#) and [Figure 37](#) for XR respectively.
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213

2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267

(Part 1) Android



'Description1': "The video starts with a view of a mobile app interface, specifically an Al-Qur'an Indonesia app with a list of Surahs. A user scrolls through the Surah list, selects a specific Surah, and the view transitions to a page displaying the verses in Arabic and their Indonesian translation. The user then returns to the Surah list. Following this, a pop-up appears indicating a configuration process that takes approximately 10-20 seconds. After the configuration, another pop-up notifies of an available update from the Google Play Store, which the user can accept or postpone. The user dismisses this pop-up. The main menu is accessed, showing options such as 'Baca Quran', 'Terakhir Baca', 'Pencarian', 'Jadwal Sholat', and 'Pengaturan'. The user selects the 'Pencarian' option, which leads to a search page. Here, the user is instructed to enter a keyword to search in the translation database. Finally, the user navigates back to the main menu and then to the Surah list once more."

'Sequential-QA': **'Question'**: 'After selecting a specific Surah from the list, what does the user do next before a pop-up notification appears?'

'Answer': 'After selecting a Surah, the user views the verses in Arabic and their Indonesian translation. Then, the user returns to the Surah list before a pop-up notification appears.'

'Prediction': **'Question'**: 'What would likely be the next step in the app interface after the user is directed to the search page?'

'Answer': "After being directed to the search page, the likely next step would be for the user to enter a keyword into the search bar to find a translation or meaning within the app's database."

'Conversation': **'User 1'**: 'I need to find a specific verse in the Quran. How do I search for it?'

'Assistant 1': "You can search for a specific verse by tapping on the 'Pencarian' option in the main menu, which will take you to a search page where you can enter a keyword."

'User 2': 'I see. And how do I get back to the main menu from the search page?'

'Assistant 2': 'To return to the main menu from the search page, you can tap on the back arrow located at the top left corner of the screen.'

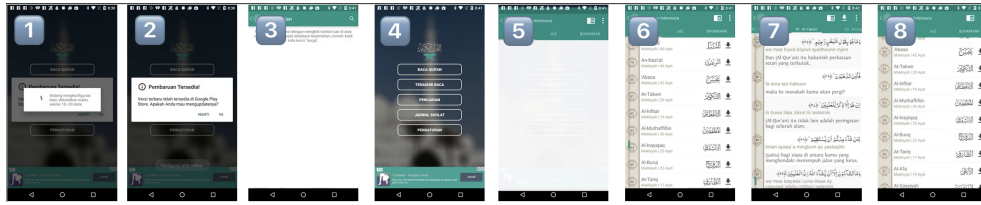
'Reasoning': **'Question'**: "If the user wants to read the verses of Surah 'Al-Infitar', what should they do after launching the app?"

'Options': ["A. Go to the settings and select 'Al-Infitar'.", "B. Scroll through the Surah list and select 'Al-Infitar'.", "C. Choose the 'Pencarian' option and type 'Al-Infitar'.", "D. Wait for a pop-up and select 'Al-Infitar' from there."]

'Correct Answer': "B. Scroll through the Surah list and select 'Al-Infitar'."

2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321

(Part 2) Android



'Description2': "The video begins by displaying a mobile GUI with a list of chapters from the Quran in Indonesian. Each chapter has a downward arrow suggesting expandable content. As the video progresses, a popup appears with a loading icon and a message in Indonesian indicating a configuration is in progress, which takes about 10-20 seconds. After this, another popup appears notifying of a new update available on the Google Play Store with options to update or postpone. Subsequently, the screen shows a search interface where users can input keywords for searching within the Quran's translated database. The main menu is then accessed, with options such as 'Read Quran', 'Last Read', 'Search', 'Prayer Schedule', and 'Settings'. The GUI transitions back to the list of chapters, and a specific chapter, At-Takwir, is selected. The video then displays the verses of this chapter, both in Arabic and Indonesian translation, with an option to listen to the audio. Finally, it navigates back to the list of chapters."

'Caption': "Navigating through a Quran app's GUI, interacting with chapter lists, update notifications, search function, and viewing specific verses with translations."

'static QA': **'Question'**: 'What options are available in the main menu of the mobile Quran application?'

'Answer': "The main menu of the mobile Quran application provides several options for the user to choose from. These include 'BACA QURAN' (Read Quran) for accessing the chapters to read, 'TERAKHIR BACA' (Last Read) to resume reading from where the user left off last time, 'PENCARIAN' (Search) to search the Quran's database for specific keywords, 'JADWAL SHOLAT' (Prayer Schedule) to check the prayer times, and 'PENGATURAN' (Settings) to modify app settings. This menu provides a simple and efficient way for users to navigate through the app's features and customize their reading and learning experience."

'MCQA': **'Question'**: 'What happens after the user is notified about the new update available on the Google Play Store?'

'Options': 'A': 'The app closes automatically.', 'B': 'The search interface is displayed.', 'C': 'The list of chapters disappears.', 'D': 'An advertisement for shopping deals is shown.'

'Correct Answer': '[[B]] The search interface is displayed.'

Figure 27: Case study for Android (part 2).

2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375

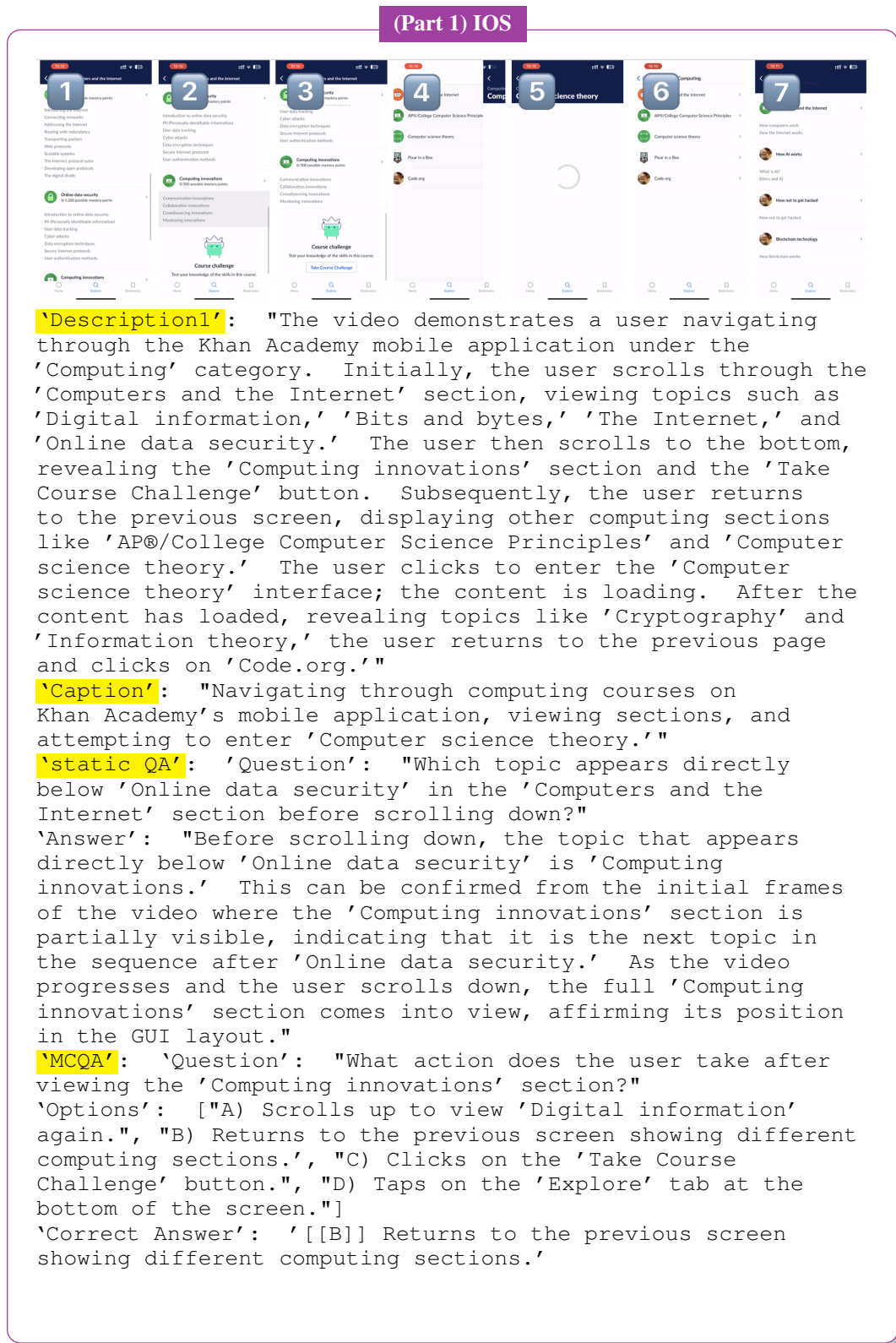
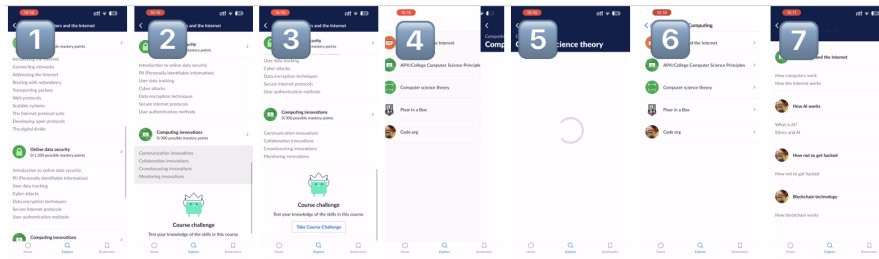


Figure 28: Case study for IOS (part 1).

2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429

(Part 2) IOS



Description2: "The video begins with the user viewing the 'Computers and the Internet' course section within the Khan Academy application. The user scrolls through various subsections such as 'Digital information,' 'Computers,' 'The Internet,' and 'Online data security,' each with a list of topics and a status of possible mastery points. The user continues to scroll down to the 'Computing innovations' section and then further down to a 'Course challenge' prompt. The user then scrolls back up, revealing previously seen sections in reverse order. The user eventually navigates back to the main 'Computing' category screen, showing an overview of all computing-related courses. From there, the user selects 'Computer science theory,' which briefly loads before displaying topics within that course such as 'Cryptography' and 'Information theory.' Following this, the user returns to the main 'Computing' category screen."

Sequential-QA: 'Question': "What action does the user take after scrolling through the 'Online data security' section, and what is displayed as a result of this action?", 'Answer': "After scrolling through the 'Online data security' section, the user scrolls down to the 'Computing innovations' section. As a result of this action, topics such as 'Communication innovations', 'Collaboration innovations', 'Crowdsourcing innovations', and 'Monitoring innovations' are displayed, followed by a 'Course challenge' prompt."

Prediction: 'Question': "If the user were to select the 'Cryptography' section after it was displayed, what kind of content could you predict will be shown next?", 'Answer': "If the user selects the 'Cryptography' section, one could predict that the next content shown would be detailed topics under Cryptography, such as 'Ancient cryptography,' 'Ciphers,' 'Cryptography challenge 101,' 'Modern cryptography,' 'Modular arithmetic,' 'Primality test,' and 'Randomized algorithms,' along with the status of possible mastery points for that section."

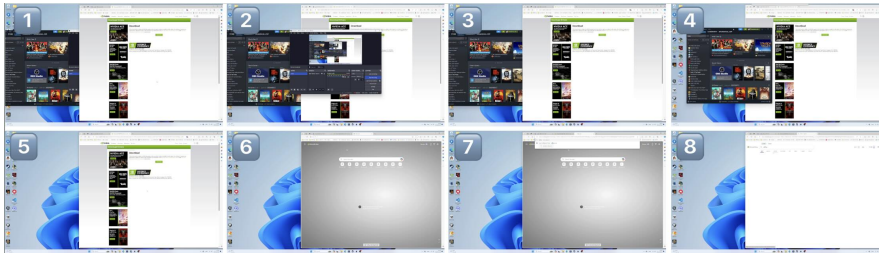
Conversation: 'User 1': 'Can you take me to the section about user authentication methods?'
'Assistant 1': "Sure, you will need to scroll down the 'Computers and the Internet' course section until you reach the 'Online data security' subsection, where you will find the topic 'User authentication methods.'"
'User 2': "I scrolled too far down and reached the course challenge. How do I find 'User authentication methods' from here?"
'Assistant 2': "No problem, you'll need to scroll back up past the 'Computing innovations' section until you see the 'Online data security' subsection again. 'User authentication methods' is listed there among other topics."

Reasoning: 'Question': "After browsing through the course topics in 'Computers and the Internet,' the user returns to a broader category view. Based on this behavior, what could be the reason for the user returning to the broader category view?"
'Options': ['A. The user wants to take a course challenge.', 'B. The user is looking for a different computing-related course.', 'C. The application automatically redirected the user.', 'D. The user intends to log out of the Khan Academy application.']
'Correct Answer': 'B'

Figure 29: Case study for IOS (part 2).

2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483

(Part 1) Multiple-Windows Interaction



'Description1': "The video begins with a Windows desktop displaying multiple open applications, including Steam, OBS Studio, and a web browser with NVIDIA's website loaded. The user starts by clicking on the back page of the browser, which partially obscures the OBS window. Then, the user clicks on the OBS application, bringing it to the forefront. The user minimizes OBS, followed by dragging the Steam window to the center of the screen and minimizing it as well. A new web page is opened in the Edge browser's navigation bar, and the user types 'office' into the search bar. The browser navigates to the Bing search interface, and 'office' is successfully searched."

'Caption': 'Navigating and Managing Multiple Applications on Windows Including Steam, OBS Studio, and Edge Browser'

'static QA': **'Question'**: "Which web browser is used in the video and which website is prominently featured before the search for 'office'?"

'Answer': "The web browser used in the video is Microsoft Edge. The prominently featured website before the search for 'office' is NVIDIA's official website where the 'Download Drivers' page is displayed."

'MCQA': **'Question'**: 'What action is taken after the OBS application is minimized?', **'Options'**: ['A. The Steam window is closed.'

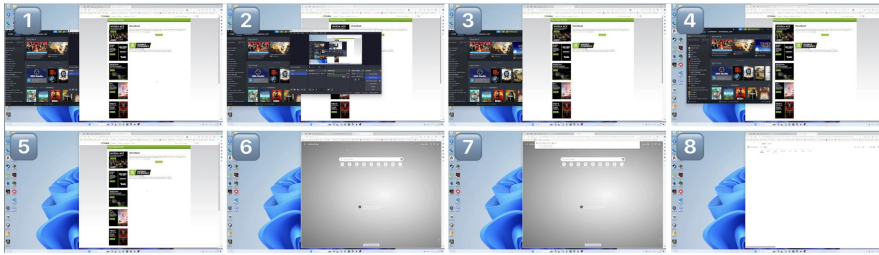
'B. The Steam window is moved to the center of the screen and minimized.', 'C. The Edge browser is closed.', 'D. A file is opened from the desktop.']

'Correct Answer': '[[B]] The Steam window is moved to the center of the screen and minimized.'

Figure 30: Case study for multiple-windows interaction (part 1).

2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537

(Part 2) Multiple-Windows Interaction



'Description2': "The video shows a Windows desktop with several application windows open, including Steam, OBS, and Edge. Initially, the OBS window is partially covering the Edge browser. The user clicks on the Edge browser, bringing it to the foreground, and then minimizes the OBS window, clearing the view. Next, the user moves the Steam window to the center of the screen and minimizes it as well. Afterward, the user opens a new tab in Edge and enters 'office' into the search bar, which leads to a Bing search results page for 'office'. The video ends with the user on the Bing search page with 'office' results displayed."

'Sequential-QA': 'Question': 'After moving the Steam window to the center, what did the user do next in the Edge browser?'
'Answer': "After moving the Steam window to the center, the user minimized the Steam application. Then, they opened a new tab in the Edge browser and entered 'office' into the search bar, which led them to the Bing search results page for 'office'."

'Prediction': 'Question': "What would be the likely next action the user performs after searching for 'office' on Bing?"
'Answer': "The likely next action the user would perform after searching for 'office' on Bing would be to click on one of the search results to find more information or to visit the Microsoft Office website to purchase or learn more about the office suite."

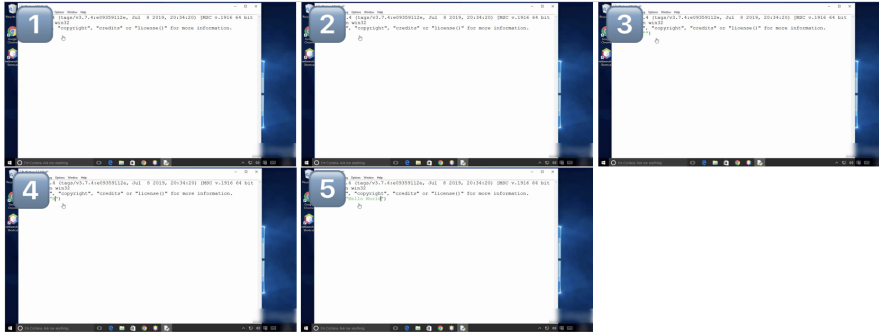
'Conversation': 'User 1': 'Can you minimize the OBS application for a better view of the browser?'
'Assistant 1': 'Certainly, the OBS application has been minimized, providing a clear view of the Edge browser.'
'User 2': 'Great, now can you search for Microsoft Office in the Edge browser?'
'Assistant 2': "Of course, a new tab has been opened in the Edge browser and 'office' has been entered into the search bar. The Bing search results for 'office' are now displayed."

'Reasoning': 'Question': 'If the user needs to record gameplay footage next, which application should they interact with and what would be their first step?'
'Options': ["A. They should open the Steam application and click on the 'STORE' tab.", "B. They should open the Edge browser and search for 'game recording software'.", "C. They should reopen the OBS application and click on the 'Start Recording' button.", "D. They should access the Windows Start menu and search for the 'Camera' app."]
'Correct Answer': 'C'

Figure 31: Case study for multiple-windows interaction (part 2).

2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591

(Part 1) Software



`'Description1'`: "The video shows a Python 3.7.4 Shell window on a Windows system. The user begins by typing the 'print' function followed by a pair of parentheses. Inside the parentheses, the user types a string, 'Hello World', which is enclosed in double quotes. Upon pressing Enter, the Python Shell executes the command and outputs the text 'Hello World' below the command line, indicating that the code ran successfully without any errors."

`'Caption'`: "Executing the print command in Python Shell to display 'Hello World'"

`'static QA'`: "Question": "What version of Python is shown running in the video?"

"Answer": "The version of Python running in the video is Python 3.7.4, as indicated by the text at the top of the Python Shell window."

`'MCQA'`: "Question": "What operation does the user perform after typing the print command?"

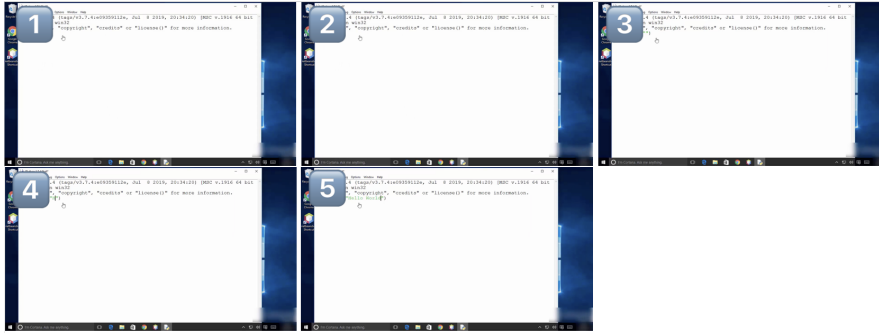
"Options": ["A. The user saves the file.", "B. The user compiles the code.", "C. The user executes the print command.", "D. The user closes the Python Shell.]"

"Correct Answer": "[C] The user executes the print command."

Figure 32: Case study for software (part 1).

2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645

(Part 2) Software



'Description2': "The video shows a sequence of a user interacting with the Python Shell on a Windows system. Initially, the shell is open and waiting for input. The user begins by typing 'print*' and then corrects the typo by removing the asterisk, typing 'print'. They then type two quotation marks to prepare for a string input. Subsequently, the user moves the cursor between the quotation marks and types 'Hello World' to complete the 'print(Hello World)' command. After pressing enter, the shell executes the command, and the phrase 'Hello World' is outputted below the command prompt."

'Sequential-QA': "Question": "Can you describe the sequence of actions the user took to correct the command and output 'Hello World'?"

"Answer": "Certainly. Initially, the user typed 'print*' but realized the asterisk was a mistake. They deleted the asterisk to correct the command to 'print'. Next, they entered two quotation marks to indicate the start of a string. The user then placed the cursor between the quotation marks and typed the string 'Hello World'. Finally, upon pressing enter, the Python Shell executed the command and displayed the output 'Hello World' beneath the command prompt."

'Prediction': "Question": "What would be the expected GUI change if the user were to type another command after the 'Hello World' output?"

"Answer": "If the user were to type another command after the 'Hello World' output, the expected GUI change would be that the command prompt indicator '>>>' would appear on a new line below the 'Hello World' output, signaling readiness for the next command input. Any subsequent command typed by the user would appear after this new prompt, awaiting execution."

'Conversation': "User 1": "I want to print a different message in Python. How can I do that?", "Assistant 1": "To print a different message, you can simply type 'print(Your Message Here)' where 'Your Message Here' is the text you wish to print. Make sure you enclose your message in quotation marks and hit enter to execute the command."

"User 2": "What if I want to print 'Goodbye, World' instead?"

"Assistant 2": "To print 'Goodbye, World', you would type 'print(Goodbye, World)' and press enter. The Python Shell will execute the command and display 'Goodbye, World' as the output below the command prompt."

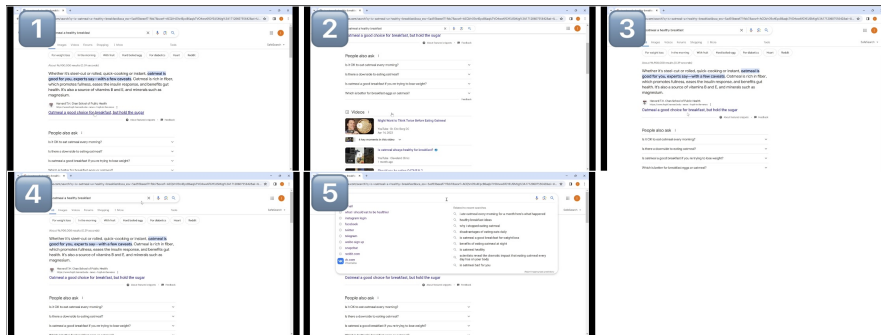
'Reasoning': "Question": "What command did the user execute to get the output in the Python Shell?"

"Options": ["A. print(Hello World)", "B. print(Hello World)", "C. print(Hello World)", "D. echo(Hello World)"]

"Correct Answer": "C",

2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699

(Part 1) Website



'Description1': "The video begins with the Google search results page visible on a Windows system browser, displaying the query 'is oatmeal a healthy breakfast'. The mouse cursor scrolls down the page, revealing additional search results, and the 'People also ask' section with related questions. The user then scrolls back up to the top of the page. Next, the cursor moves to the search bar, and the 'X' button is clicked to clear the previous search content, leaving an empty search bar. The browser's suggested searches drop-down menu appears with various related search queries. Finally, the video fades to black, indicating the end of the sequence."

'Caption': 'Navigating Google Search Results and Clearing the Search Query on a Windows System Browser'

'static QA': 'Question': "What feature snippet is displayed at the top of the Google search results for the query 'is oatmeal a healthy breakfast'?"

'Answer': "The featured snippet at the top of the Google search results for the query 'is oatmeal a healthy breakfast' is from the Harvard T.H. Chan School of Public Health website. It includes an excerpt stating 'Whether it's steel-cut or rolled, quick-cooking or instant, oatmeal is good for you, experts say|with a few caveats. Oatmeal is rich in fiber, which promotes fullness, eases the insulin response, and benefits gut health. It's also a source of vitamins B and E, and minerals such as magnesium.' This snippet provides a concise summary of the health benefits of oatmeal, according to experts, highlighting its nutritional value and potential impact on fullness and insulin response. The presence of this snippet offers a quick and authoritative answer to the user's query, showcasing Google's ability to extract relevant information from web pages and present it prominently for ease of access."

'MCQA': 'Question': 'What action did the user take after reviewing the search results?'

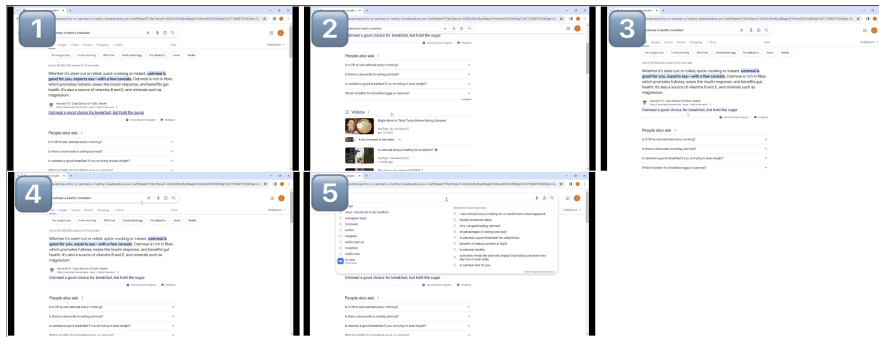
'Options': ['A. The user clicked on one of the search results.', 'B. The user scrolled through the 'People also ask' section.', 'C. The user cleared the search content in the search bar.', 'D. The user navigated to a different website.']

'Correct Answer': '[[C]] The user cleared the search content in the search bar.',

Figure 34: Case study for website (part 1).

2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753

(Part 2) Website



'Description2': "The video shows a sequence of actions on a Google search results page within a web browser on a Windows system. Initially, the mouse cursor moves over a search result discussing the health benefits of oatmeal. Next, the user scrolls down, revealing a 'People also ask' section with questions related to oatmeal and a 'Videos' section showcasing related content. Subsequently, the user scrolls back up to the original position, highlighting the same search result about oatmeal's health benefits.

Finally, the user moves the cursor to the search bar and clicks the 'X' to clear the previous search content, resulting in a blank search bar with suggestions and related searches listed below it. The screen then goes black, indicating the end of the video."

'Sequential-QA': 'Question': "After reading about oatmeal's health benefits, what section of the page did the user explore next and did they return to read more about the health benefits afterward?"

'Answer': "After reading about the health benefits of oatmeal, the user scrolled down to explore the 'People also ask' section and the 'Videos' section related to oatmeal. Following this, the user scrolled back up to revisit the original search result that discussed the health benefits of oatmeal."

'Prediction': 'Question': 'Based on the sequence of events, what action is the user likely to perform after clearing the search bar?'
'Answer': 'After clearing the search bar, the user is likely to type in a new search query or select one of the suggested searches that appear below the search bar.'

'Conversation': 'User 1': "I want to find more articles on oatmeal's health benefits. Can you scroll back up?"
'Assistant 1': "I've scrolled back up and the article on oatmeal's health benefits from the Harvard T.H. Chan School of Public Health is highlighted again."

'User 2': 'Great, now can you clear the search and look for something else?'
'Assistant 2': 'The search content has been cleared, and the search bar is now empty, showing a list of related searches and previous search history suggestions for a new query.'

'Reasoning': 'Question': 'If the user wants to perform a new search after clearing the search bar, which of the following actions would they need to take next?', 'Options': ['A. Scroll down to view more search results'

'B. Type a new query into the search bar', "C. Click on one of the 'People also ask' questions", 'D. Close the browser window']
'Correct Answer': 'B',

Figure 35: Case study for website (part 2).

2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807

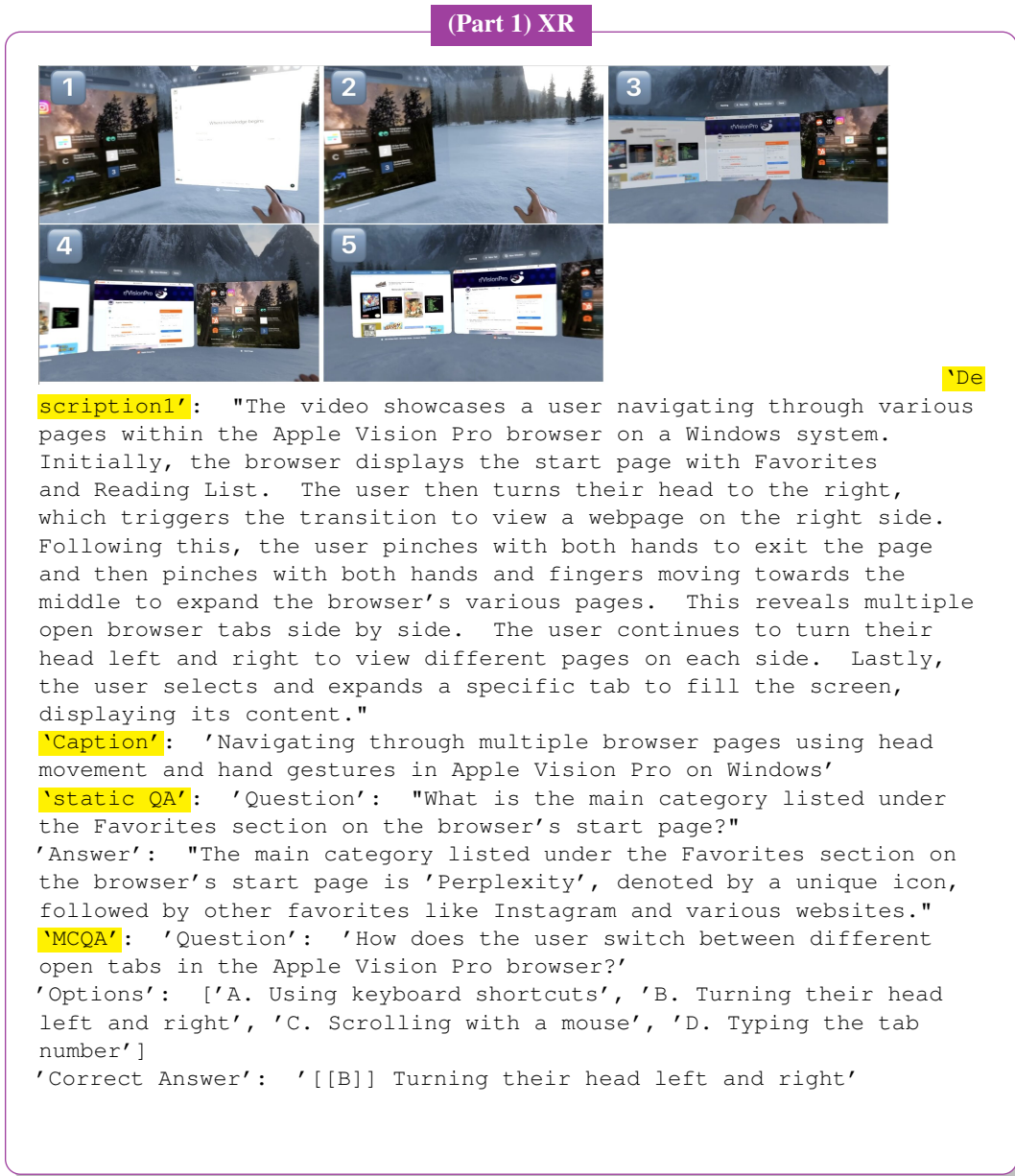
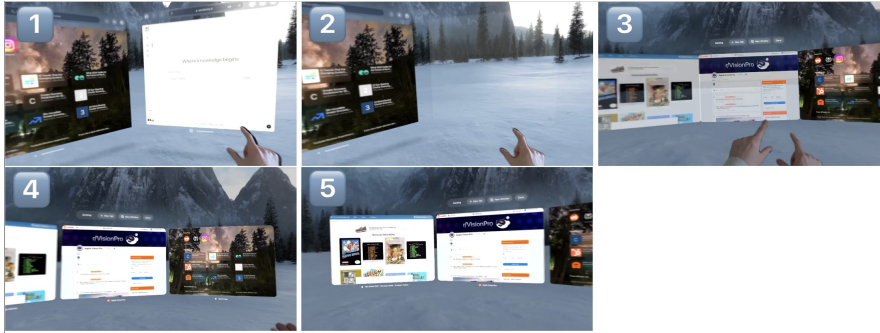


Figure 36: Case study for XR (part 1).

2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861

(Part 2) XR



'Description2': "The video starts with a full-screen view of a browser interface titled 'Apple Vision Pro' displaying various website thumbnails and bookmarks. The user then turns their head to the right, causing the right side of the browser to come into view. Next, the user looks at a dot at the bottom of the page and pinches with both hands, which causes the browser to exit the full-screen view and shrink to a smaller, windowed mode. The user then performs a pinching motion with both hands, bringing the fingers towards the middle, which causes the browser's various pages to expand, giving an overview of multiple open tabs. The user again turns their head to the right to view the right side page and then to the left to view the left side page. Throughout the video, the GUI elements such as tabs, the address bar, and website thumbnails respond dynamically to the user's head movements and hand gestures."

'Sequential-QA': 'Question': 'What action did the user take immediately after turning their head to the right to view the right side page?'

'Answer': 'After turning their head to the right to view the right side page, the user looked at a dot at the bottom of the page and pinched with both hands, which caused the browser to exit full-screen mode and shrink to a smaller windowed interface.'

'Prediction': 'Question': 'If the user pinches with both hands and moves their fingers towards the middle after the browser shrinks to a smaller window, what will happen next?'

'Answer': "If the user pinches with both hands and moves their fingers towards the middle after the browser shrinks, the action will cause the browser's various pages to expand, providing an overview of multiple open tabs."

'Conversation': 'User 1': 'How do I view all the pages I have open in the browser?'

'Assistant 1': 'You can view all the open pages by pinching with both hands and moving your fingers towards the middle. This will expand the browser and show an overview of all the open tabs.'

'User 2': 'Great, and how can I exit this overview?'

'Assistant 2': 'To exit the overview of open pages, you can look at the dot at the bottom of the page and pinch with both hands. This will exit the overview and return you to the individual page view.'

'Reasoning': 'Question': 'How can the user access the options to open a new tab or window from the current state?'

'Options': ['A. Turn their head to the left and select the plus sign.', 'B. Swipe left on the touchpad.', 'C. Turn their head to the right and select the 'Done' button.', 'D. Pinch with both hands to exit the current view and access the toolbar.']

'Correct Answer': 'D'

Figure 37: Case study for XR (part 2).