

---

# Supplementary Materials

---

## 1 Documentation and Intended Uses

This dataset covers six GUI scenarios and eight types of GUI-orientated questions in three formats, with 12,379 videos and more than 100k QA pairs focusing on both static and dynamic GUI content. All features are listed and explained below:

- video: the video of the GUI content.
- question: each video has 6 questions, with 4 free-form questions and 2 multiple-choice questions.
- golden answer: each free-form question has a golden answer, while the multiple-choice question has a definite correct option.
- multi-round conversation: each video has a multi-round conversation, with two roles: 'Assistant' and 'User'.
- human-selected keyframe: each video contains human-selected keyframes (except android), with annotation of subgoal, operation of mouse and keyboard.
- description (caption) of video: each video has two detailed captions and one concise caption, focusing on different aspects of the video content.
- description of keyframes (only for test split): In test split, we use GPT-4V to annotate all the keyframe for detailed and concise caption for ablation study.

This dataset will be used for benchmarking current MLLMs, pre-training datasets for GUI-Orientated models, and future research in various GUI-related areas, such as GUI grounding and General Virtual Agents.

## 2 Author Statement

As the authors of this dataset, we hereby declare that we bear full responsibility for any violations of rights, including but not limited to intellectual property rights, privacy rights, or any other legal rights that may arise from the distribution of this dataset. We assure that the dataset complies with all ethical guidelines and legal requirements in the creation and distribution of the dataset. We confirm that the dataset is distributed under the [CC BY 4.0] license.

### 32 **3 Plan**

33 The dataset and benchmark model weights have been uploaded to HuggingFace, and  
34 the dataset has been publicly released under the [CC-BY-4.0] license. HuggingFace  
35 is designated as the primary release platform, with plans for ongoing optimization  
36 and expansion of the dataset. Prior to the paper’s official publication, the dataset  
37 will be archived, and a Digital Object Identifier (DOI) will be assigned. Once a DOI is  
38 assigned to the dataset on HuggingFace, it becomes immutable, preventing actions  
39 such as deletion, renaming, transfer, or modification of visibility to private. This  
40 mechanism ensures the dataset’s enduring preservation and affords it a persistent,  
41 dereferenceable identifier.

42 **Note: We have decided to change the license from MIT mentioned in the paper**  
43 **to CC-BY-4.0**