

ILuvUI: Instruction-tuned LangUage-Vision modeling of UIs from Machine Conversations

YUE JIANG, Aalto University, Finland

ELDON SCHOOP, Apple, USA

AMANDA SWEARNGIN, Apple, USA

JEFFREY NICHOLS, Apple, USA

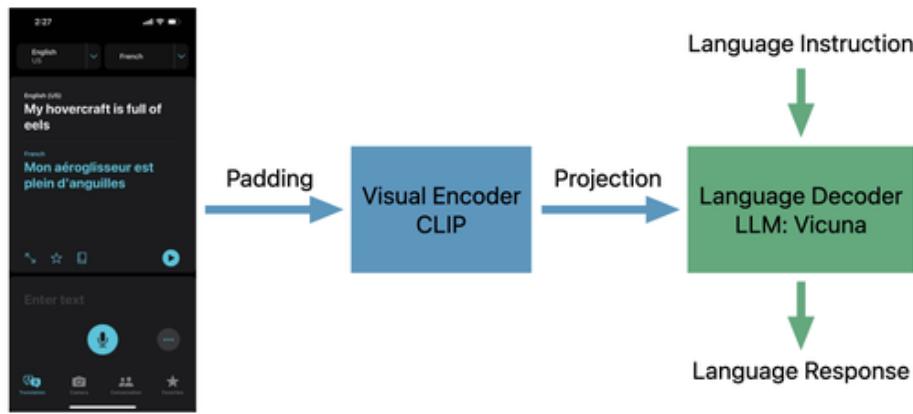


Fig. 1. We generate a dataset of Q&A pairs that cover several UI description and reasoning tasks and use it to fine-tune a Vision-Language Model (VLM) that accepts multimodal inputs from screenshot pixels and a user-provided text input. Our model, ILuvUI, enables new kinds of UI tasks with conversational VLMs.

Multimodal Vision-Language Models (VLMs) enable powerful applications from their fused understanding of images and language, but many perform poorly on UI tasks due to the lack of UI training data. In this paper, we adapt a recipe for generating paired text-image training data for VLMs to the UI domain by combining existing pixel-based methods with a Large Language Model (LLM). Unlike prior art, our method requires no human-provided annotations, and it can be applied to any dataset of UI screenshots. We generate a dataset of 335K conversational examples paired with UIs that cover Q&A, UI descriptions, and planning, and use it to fine-tune a conversational VLM for UI tasks. To assess the performance of our model, we benchmark it on UI element detection tasks, evaluate response quality, and showcase its applicability to multi-step UI navigation and planning.

CCS Concepts: • Human-centered computing → Interactive systems and tools; Natural language interfaces; User interface toolkits; • Computing methodologies → Scene understanding.

Authors' addresses: Yue Jiang, Aalto University, Espoo, Finland; Eldon Schoop, Apple, Seattle, USA; Amanda Swearngin, Apple, Seattle, USA; Jeffrey Nichols, Apple, Seattle, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

ACM Reference Format:

Yue Jiang, Eldon Schoop, Amanda Swearngin, and Jeffrey Nichols. 2023. ILuvUI: Instruction-tuned LangUage-Vision modeling of UIs from Machine Conversations. 1, 1 (October 2023), 19 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

For nearly as long as graphical user interfaces (UIs) have been popular, users have sought methods to verbally describe them for accessibility [24] or automate interaction with them, for example to execute repetitive tasks [16]. Understanding and automating actions on UIs is a challenging task since the UI elements in a screen, such as list items, checkboxes, and text fields, encode many layers of information beyond their affordances for interactivity alone. The layout, visual style, and textual content of a UI are often designed to afford user expectations for the domain and capabilities of the application, and the elements themselves may be dynamic or stateful. This inherent complexity makes comprehending and conveying UI-related information through natural language challenging.

Many works in HCI have sought to automate tasks on UIs using programming by demonstration [16, 35, 37], mining similar screenshots through past interactions [2], and Large Language Models (LLMs) [41, 62, 63]. LLMs in particular have demonstrated remarkable abilities to comprehend task instructions in natural language in many domains [15, 50, 51, 61, 70], however using text descriptions of UIs alone with LLMs leaves out the rich visual information of the UI. Fusing visual with textual information is important to understanding UIs as it mirrors how many humans engage with the world. One approach that has sought to bridge this gap when applied to natural images are Vision-Language Models (VLMs), which accept *multimodal* inputs of both images and text, typically output only text, and allow for general-purpose question answering, visual reasoning, scene descriptions, and conversations with image inputs [1, 14, 17, 40]. However, the performance of these models on UI tasks fall short compared to natural images because of the lack of UI examples in their training data.

In this paper, we adapt the LLaVA [40] VLM training data generation recipe to the UI domain. Our data generation recipe does not require any human labeling, and can be adapted to UIs in existing datasets such as Rico [19], which do not have existing textual descriptions. We generate a dataset of 335K image-instruction pairs using screenshots from the AMP dataset [69] and additional data from the Crawls interaction trace dataset [25]. In order to generate text pairs for a given screenshot, we employ a UI element detection model [69] that identifies the elements in a screen, converts these detections into a structured text-based representation, and then prompts an LLM with this representation and additional context to generate one or more realistic phrases. We generate six types of phrases using different prompts: single-step Q&A conversations, detailed descriptions, listing available actions, predicting UI action outcomes, selecting a UI element given a goal, and a goal-based plan.

After generating data using the gpt3.5-turbo LLM [48], we use that data to fine-tune a conversational VLM. Our resulting model, ILuvUI, is capable of describing properties of UI elements and screens, contextual help, and planning multi-step interactions. Like many large, unsupervised models, ILuvUI is capable of tasks it was not trained to perform. Beyond interpreting the complexities of UIs and following instructions, our model paves the way towards using VLMs to enhance UI accessibility by automatically generating descriptions and acting on instructions provided by speech. To better understand ILuvUI’s performance, we benchmark against a UI element detection model, evaluate its response quality compared to a baseline VLM, and show selected examples which demonstrate ILuvUI’s planning and reasoning capabilities.

Our paper makes the following contributions:

Manuscript submitted to ACM

- (1) We adapt the LLaVA method [40] to generate paired text-image data to train VLMs in the UI domain using an LLM and a UI element detector. Our method only uses UI screenshots as input, does not require any human-provided captions, and produces six different types of text pairs.
- (2) We fine-tune an open-source VLM, LLaVA [40], on our UI dataset, leading to the development of a UI-focused instruction-following visual agent, and assess its performance with several lightweight evaluations.

2 RELATED WORK

2.1 UI Understanding

Pixel-based UI understanding is broadly applicable across diverse domains, including interface adaptation [3, 7, 21, 68], GUI testing [66], data-driven GUI searches [8, 10, 28], prototyping [57], UI code generation to support app development [4, 9, 11, 18, 47], and GUI security [12]. To identify UI elements within images, conventional image processing methods have long been employed, often relying on detecting and aggregating edges [47, 57, 67]. While adept at handling simpler GUIs, these methodologies struggle when faced with images that have gradients, photographs, or intricate UI layouts. Furthermore, template matching techniques [21, 52, 66, 68] require meticulous feature engineering and curated templates, introducing potential limitations when applied to UIs with diverse visual attributes. Many reverse engineering approaches also predict UI elements and layouts [5, 6, 21–23, 30–34, 42, 44, 47, 54, 58, 66], detect hierarchical groupings [65], generate code for UIs [4, 67], and enable platform migration or UI improvement [20, 43, 45, 56].

The surge of deep learning-based approaches has been enabled by comprehensive UI datasets such as Rico [19]. A prominent instance of this trend is witnessed in pix2Code [4], which employs an end-to-end neural image captioning model to generate descriptions of interface layouts. Correspondingly, Chen et al. [9] uses a CNN-RNN model to create a UI skeleton that includes widget types and layouts inferred from screenshots. Other models [8, 64] use object detection techniques to detect GUI widgets within screenshots. Some prior research [55, 59] predicts the tappability of mobile app UI elements. Additionally, some deep learning models use crowdsourcing to generate image captions [26, 27]. Other approaches [11, 13, 46] combine traditional image processing methods, such as edge detection, with deep learning-based classification models to detect UI elements and semantics, including UI types. Zhang et al. [69] infers interface elements and metadata from pixels, including UI types, navigation order, groupings, image descriptions, and icon classes. We make use of this latter method to produce textual representations of UIs that we use to generate text-screenshot pairs for VLM training.

2.2 Large Language Models and Vision Language Models

Large language models (LLMs) have broad potential in interpreting language, serving as an interface for a versatile AI assistant, are capable of comprehending explicit task instructions in natural language. ChatGPT [48] and GPT-4 [49] show the proficiency of well-aligned LLMs in following human instructions. This achievement has sparked a surge of interest in the development of open-source LLMs. Among these, LLaMA [61] is a downloadable LLM that rivals the performance of GPT-3. Other models like Alpaca [60], Vicuna [15], and GPT-4-LLM [49] have leveraged machine-generated, high-quality instruction-following examples to enhance the alignment capabilities of LLMs. The LLaVA [40] model on which we base our work is itself a combination of the CLIP [29] and Vicuna [15] models.

Spotlight is an example of a VLM that has been trained for UI tasks [36]. This model takes as input both an image of UI and a region of interest, and outputs text related to the region, and was applied to tasks such as widget captioning, screen summarization, command grounding, and tappability prediction. The Spotlight model is quite powerful, however



Fig. 2. We use two author-created “golden examples” of UI elements and Q&A pairs as few-show examples for data generation. These examples are prepended to a third list of UI elements, where gpt3.5-turbo predicts new, resulting Q&A pairs.

it is somewhat limited by the requirement that a region of interest be specified, which may not be knowable in advance for UI automation tasks without the use of other technologies. ILuvUI in contrast does not require a region of interest, and accepts a text prompt as input in addition to the UI image, which enables it to provide answers for use cases such as visual question answering.

3 GENERATING MULTIMODAL DATA FOR USER INTERFACES

While general-purpose VLMs perform well on natural images for several tasks, their capabilities are limited when performing tasks with UIs. This is because existing VLMs lack latent knowledge of common UI design principles, tasks, and element types. Several datasets exist in the literature that pair natural images with text descriptions, Q&A, or instructions, but the availability of training data that ties UI screenshots with descriptions or grounded conversation is comparatively limited. This motivates the need to create a dataset that pairs UIs with natural language instructions and descriptions to enable training a VLM that can understand and act on UIs.

LLaVA [40] is recent work that introduces a method for creating text-image data pairs from an existing image dataset and a multimodal VLM trained on that data which connects the CLIP vision encoder [29] with the Vicuna LLM [15] for general-purpose visual and language understanding. We adapt both aspects of this work to create a VLM tuned to UI tasks.

A key contribution of LLaVA’s data generation method is using an off-the-shelf LLM to generate a dataset of text-image pairs from the existing COCO image dataset [39]. The COCO dataset pairs images with human-annotated captions and bounding boxes of objects within each image. For each image, the method combines the captions, list of bounding boxes of objects with the object types, and an additional prompt into a request to a proprietary LLM (experiments were done with both GPT3.5 and GPT4). The resulting outputs from the LLM contained realistic text phrases that were paired with the original image to produce a VLM training set. LLaVA produced three kinds of text pairs that they termed Q&A, detailed description, and complex reasoning.

We adapted the LLaVA dataset generation recipe to UI tasks in several ways, to fit our existing UI dataset and different UI tasks. The LLaVA method relied on the COCO dataset [39] to provide both human-annotated captions and object bounding boxes. For UI data, we use the AMP dataset from Zhang et al. [69], which comprises 80,945 unique screens from 4,239 iPhone apps; and the CRAWL dataset [25], which contains 750,000 iOS app screens from 6,000 apps. These datasets contain screenshots and annotations of UI element bounding boxes, but no human-annotated captions of its UIs. Ultimately and although they are of high quality, we decided not to use the UI element bounding box annotations, and instead produced the bounding boxes using an object detection model [69] during data generation as we sought a method that would work with completely unannotated UI examples as might be found in other datasets. The missing UI captions were generated using an LLM for each screen with a prompt that included the UI element detections. We

then created text pairs by generating analogous examples to the original recipe (Q&A, detailed descriptions), as well as adapting LLaVA’s original complex reasoning task to multiple different UI tasks: listing possible actions, predicting the outcome of actions, selecting an element that accomplishes a goal, and goal-based planning. Figure 3 provides an example of this process.

In total, we generate 353K unique language-image instruction-following samples, including 224K in conversations, 32K in concise description, 32K in detailed description, 32K in logical reasoning, 32K in potential actions, and 1K in UI transition, respectively.

3.1 Detecting UI Elements and Generating Screen Captions

For each screen, we use the UI detection model from Zhang, et al [69] to generate a list of UI elements, their bounding boxes and types. UI element information is included in all of our data generation prompts and is formatted as follows:

`Label: [type], Text: [text], BoundingBox from (x1, y1) to (x2, y2)`

[type] refers to the element category (e.g., text, button, or icon), [text] is the text contained by the UI element extracted via OCR, and the bounding box coordinates are defined by the top-left position and the bottom-right position of the UI element, respectively.

In order to generate screen captions, which are also used in the rest of our data generation prompts, we query gpt-3.5-turbo with formatted the detected UI elements using the following prompt: “Given the UI screen [BoundingBoxes]. Write a single-sentence usage description for this UI screen.” This prompt was determined experimentally and seemed to produce good results.

3.2 Generating LLaVA-Style Data

LLaVA generates both single-step Q&A conversations and detailed descriptions as part of its data generation procedure, which we also include in our own procedure. For both, we use the same few-shot in-context learning method as LLaVA to present gpt-3.5 with examples of desired output using 2 “golden examples” that we authored. As shown in Figure 2, we design a system message and select two UI examples with their respective bounding boxes, captions, and desired question and responses for each data type. Example prompts for each data type are provided in Supplementary Material.

Single-Step Q&A Conversations. The Q&A data type consists of question-answer pairs between an assistant and a user, conditioned on a source UI screenshot. The assistant responds in a way that suggests they are viewing the UI image while answering the question. A variety of questions are asked about the UI image, including element attributes, placements, relative positions, functionalities, and purpose. Only questions with definite answers are considered.

Detailed Description. The Detailed Description data type is a rich and comprehensive description of a UI image describing all of its elements and their respective functionality. We elicit descriptions from a list of various questions that ask the LLM to describe the UI in detail and in terms of the elements on that screen. For each UI screenshot example, we randomly select a single question from this list to generate the description.

3.3 Generating UI-Specific Data

LLaVA includes a generic “complex reasoning” task as part of its data generation process, which elicited complex responses including common-sense or historical descriptions of elements in the image, and step-by-step descriptions of actions taking place in a scene. We found that this task was too general for UI tasks, and instead chose to substitute it

with four UI-specific data generation types: identifying possible actions in a screen, predicting the outcome of taking an action, selecting a UI element capable of a given action, and formulating a plan that accomplishes a goal on the given screen. These tasks were designed to familiarize the VLM with the capabilities of particular UI components, and to convey a sense of the affordances that are signified by common UI design patterns [38].

3.3.1 Listing Available UI Actions. For our first task, gpt-3.5-turbo is prompted to list all of the potential actions that can be taken on a particular UI, given its high level description and the list of recognized elements on the screen. Actions can range from tapping, swiping, or entering text depending on the functionality of the UI. For this task, we construct a prompt from the UI screenshot caption and its bounding boxes, and use the zero-shot result from gpt-3.5-turbo.

3.3.2 Predicting the Outcome of UI Actions. In addition to understanding the capabilities of various UI elements, we wanted our VLM to be able to reason about the potential outcomes of acting on UI elements. From the interaction traces in the Crawls dataset [25], which overlaps with the AMP dataset, we were able to establish ground truth correspondences from an element that was tapped on one screen and the second screen to which that interaction led. This dataset includes the pixel coordinates of a tap action that caused the UI transition. We run UI element detection on the first screen and match the pixel to the UI element with the smallest bounding box that contains the tap location to identify what specific element was tapped to cause the transition.

We feed the UI elements of the first screen, tapped UI element, and caption of the second screen to gpt-3.5-turbo, and prompt it to formulate a single QA pair that asks what action will take place when the given UI element is tapped, and answer with a concise description of the second screen. Since the LLM has access to ground truth in its context for these cases, there is no need to use few-shot examples.

3.3.3 Selecting a UI Element Given a Goal. As a complement to the above task, we also use the LLM to generate a QA pair that asks what element must be tapped on the first screen in order to arrive at a target view, specified by a concise description of the second screen. The question is prompted to not include any information about the second screen. Similar to the above, this QA pair also uses Crawls dataset, so there is no need to use few-shot since ground truth is available.

3.3.4 Goal-Based Planning. The last UI task we use for data generation is goal-based planning, where a QA pair is generated to ask for directions to accomplish a task in the specified UI. Unlike the pairs generated with the Crawls dataset, the generated directions need not result in opening a new view. Generated directions can interact with the current view (e.g., to change its state), formulate a multi-step plan across multiple views, or relate a task with the current UI to commonsense knowledge. In practice, most generated examples formulated multi-step plans for completing complex tasks on the given screen. For this task, we use two 2-shot examples. The first example asks what step should be taken to log into a given login screen (first, enter email; then password; then click “sign in”). The second example asks what the user should do when the item shown on a product page is too expensive (open a visible details menu to see if there are promotions or discounts available).

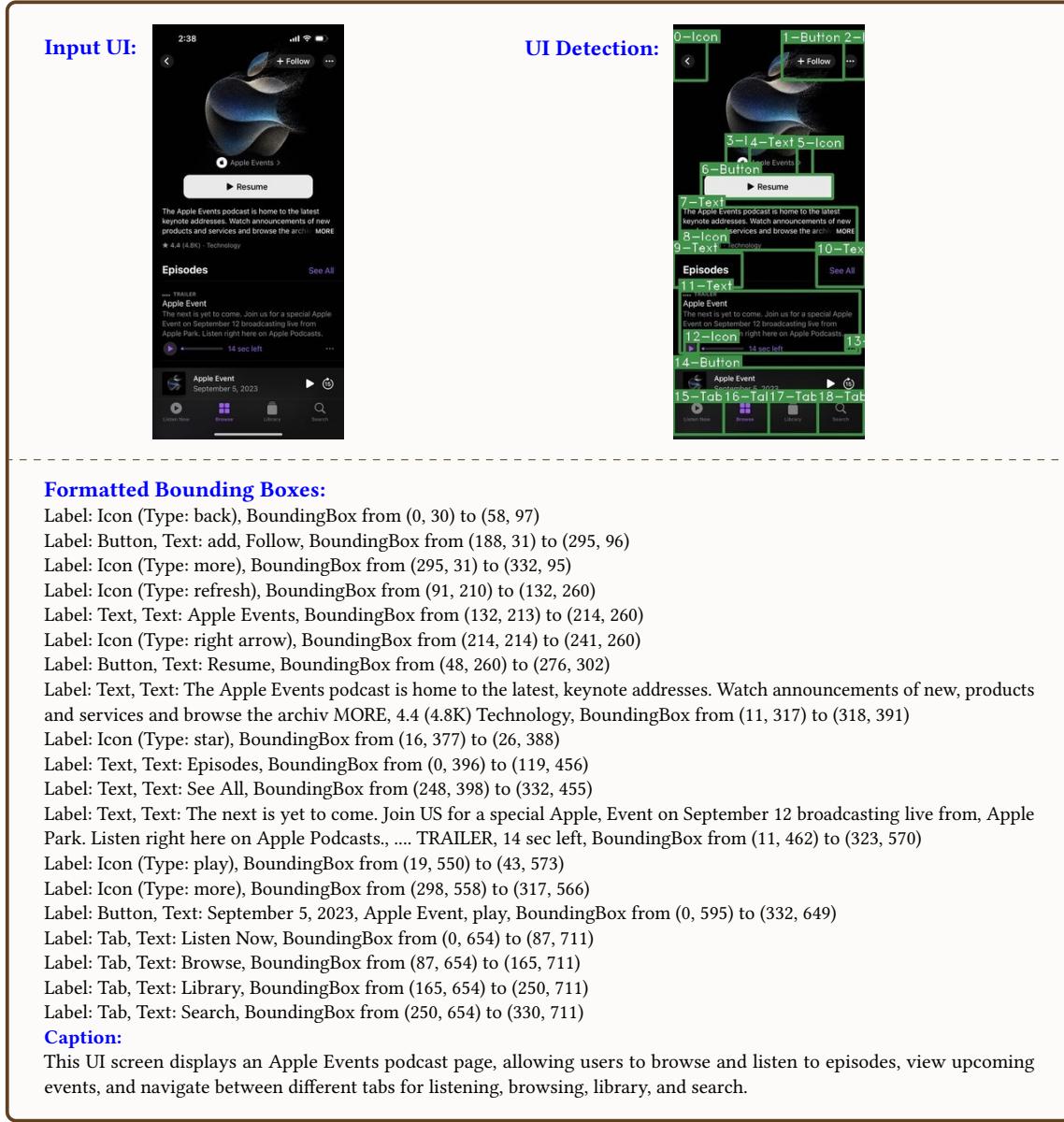


Fig. 3. One example to illustrate the input UI image, the UI detection results, formatted bounding boxes and caption.

4 VISION LANGUAGE MODEL

To obtain the UI-focused instruction-tuned VLM, we employ the same network architecture as the LLaVA model [40] with some modifications to the image input step. We found in initial experiments that the default CLIP encoder used by LLaVA used an input of 224x224 and resized input images to fit these dimensions. Resizing these images to such a small size and to fit the square shape causes significant loss of information and distortion, because of the relatively many

Conversation:

Question: What are the options available in the navigation bar at the bottom?

Answer: The options available in the navigation bar at the bottom are “Listen Now”, “Browse”, “Library”, and “Search”.

Question: What does the “Follow” button do?

Answer: The “Follow” button allows users to follow the Apple Events podcast, which means they will receive updates or notifications when new episodes are released.

Question: What information is provided about the Apple Events podcast?

Answer: The information provided about the Apple Events podcast includes a description of the podcast, its rating, and its category. The description mentions that the podcast is home to the latest keynote addresses, and users can watch announcements of new products and services and browse the archive.

Question: Where is the “Back” icon?

Answer: The “Back” icon is at the top-left corner.

Detailed Description:

Question: Describe the image in detail.

Answer:

This UI screen presents an Apple Events podcast page. At the top, there are three icons: a back button on the left, a ‘Follow’ button in the middle, and a ‘more’ icon on the right. Just below these, a refresh icon is positioned on the left, followed by the title ‘Apple Events’ and a right arrow icon. A ‘Resume’ button is situated beneath these elements.

The page includes a description of the Apple Events podcast, which reads: “The Apple Events podcast is home to the latest, keynote addresses. Watch announcements of new, products and services and browse the archive MORE, 4.4 (4.8K) Technology”. A star icon is present within this text block, likely indicating a rating or favorite feature.

The ‘Episodes’ section is found below the description, with a ‘See All’ option on the right. The first episode listed is titled “The next is yet to come. Join US for a special Apple, Event on September 12 broadcasting live from, Apple Park. Listen right here on Apple Podcasts., TRAILER, 14 sec left”. This episode entry includes a play icon on the left and a ‘more’ icon on the right.

Below this episode, there’s a button labeled “September 5, 2023, Apple Event, play”, which likely represents an upcoming event.

At the bottom of the screen, there are four tabs for navigation: ‘Listen Now’, ‘Browse’, ‘Library’, and ‘Search’, listed from left to right. These tabs allow users to switch between different sections of the podcast page.

Fig. 4. One example to illustrate the data showing the four types of responses for UI understanding.

intricate details of UI images compared to natural images. This in turn led to unsurprisingly poor interpretation of UI elements and text information.

We made two modifications to address this issue. First, we pad the input UI images to square them before resizing to the required input size for the model, which prevents the UI images from becoming overly distorted. Second, we use CLIP-L-336px [53] as the visual encoder, which increases the model input size to 336x336 pixels. We were concerned that this size would still be too small to capture UI detail, but our experiments showed that performance was much improved compared to the smaller image size.

Otherwise, our design matches LLaVA. The visual encoder is followed by the Vicuna-13B LLM [15, 70] as a language decoder, as illustrated in Figure 1. Given the input UI image \mathcal{I} , we first pad and then resize the UI image to fit the visual encoder input. Grid features from before and after the last transformer layers of the pre-trained CLIP visual encoder are

Listing Available UI Actions

Question: What actions can we perform on this user interface?

Answer:

- Tap on the "Button" with the text "add"
- Tap on the "Button" with the text "Resume"
- Tap on the "Icon" (Type: star)
- Tap on the "Text" with the text "See All"
- Tap on the "Icon" (Type: play)
- Tap on the "Icon" (Type: more)
- Tap on the "Button" with the text "September 5, 2023, Apple Event, play"
- Tap on the "Tab" with the text "Listen Now"
- Tap on the "Tab" with the text "Browse"
- Tap on the "Tab" with the text "Library"
- Tap on the "Tab" with the text "Search"

Goal-Based Planning

Question: What should the user do if they want to listen to the latest Apple Event podcast episode?

Answer:

To listen to the latest Apple Event podcast episode, the user should look under the "Episodes" section, which is located towards the middle of the UI screen. The most recent episode is typically listed first. The user can identify the episode by its title, which in this case is "Apple Event". To start listening, the user should click on the "play" icon located within the bounding box of the episode.

Fig. 5. One example to illustrate the data showing the four types of responses for UI understanding.

then used to output the visual features $h_v = \mathcal{E}(\mathcal{I})$. Following this, we apply a trainable feature alignment projection matrix P to align the visual content into language embeddings $h_p = P \cdot h_v$, which maintain the same dimension as the word embedding in the language model. We then concatenate the projected visual embedding h_p and the word embedding of the input language instruction h_l , and use the language decoder to generate the language response $R = \mathcal{D}(h_p, h_l)$.

4.1 Data Preparation

For each input UI image \mathcal{I} , and the generated pairs of questions and answers, $(Q_1, A_1, \dots, Q_T, A_T)$, where T represents the total number of turns, we arrange them into a sequential format. In this sequence, the user's instruction during the first turn is randomly selected from either $[Q_1, \mathcal{I}]$ or $[\mathcal{I}, Q_1]$, while for all subsequent turns, Q_t is used as the user instruction. We consider all the corresponding answers, A_t , as the responses provided by the assistant at each respective turn. This ensures that the sequence of user input, whether the UI image or language instruction comes first, does not affect the testing phase.

4.2 Fine-tuning Process

ILuvUI uses the pretrained visual encoder \mathcal{E} and the feature alignment projection matrix P from LLaVA [40] to align the visual content embedding with the pre-trained LLM word embedding. We use the same language decoder as LLaVA \mathcal{D} ,



Fig. 6. One example to illustrate the data generation for UI transition.

a pretrained Vicuna-13b-v1.3 LLM [70]. We freeze the visual encoder weights while fine-tuning the feature alignment projection matrix and the LLM weights, resulting in a UI-focused instruction-tuned VLM.

Agent	Element Existence (%)				Element Type (%) Accuracy	Human Preference (%)
	Accuracy	Precision	Recall	F1 Score		
Ours	67.94	67.32	70.23	68.74	25.67	72
LLaVA	51.65	50.95	99.59	67.41	9.03	20

Table 1. Comparison between ILuvUI and LLaVA on all evaluation tasks.

5 EVALUATION

We evaluate ILuvUI compared with the original LLaVA model in two ways. First, we analyze the model’s ability to perform two basic UI Understanding tasks: identifying the existence of a UI element and a UI element’s type. Second, we compare human preferences for UI screen descriptions generated by the two models. We plan further in-depth analyses as part of future work.

We ensure a fair comparison with the LLaVA model, we compare ILuvUI with a LLaVA model that uses the same 336-pixel-square vision encoder and language decoder. This LLaVA model is used for all comparisons. For the evaluation, we sampled 100 UI images that had not been seen by ILuvUI previously during training or fine-tuning. The full results of all evaluations are shown in Table 1.

5.1 UI Understanding Tasks

We use the same UI element detection model [69] used earlier for data generation to produce data that we take as ground truth for evaluating models’ UI understanding capabilities. The model provides bounding boxes and element types for all the UI elements in each evaluation UI image.

In order to evaluate UI element detection, for each screen in the evaluation set we produce a set of positive and negative detection samples. 5 positive samples are taken from the elements detected on each screen. 5 negative examples are randomly sampled from detections on other screens in the evaluation set, but only after a basic matching check is applied to ensure an equivalent element is not likely to appear on the original screen (e.g., close and back icons appear relatively commonly). We query both models for each of the 5 positive and 5 negative samples, asking whether they exist in the input UI image. Our ILuvUI agent achieves an accuracy of 67.94%, outperforming the LLaVA model, which has a 51.65% accuracy. Note that the LLaVA model nearly always provides positive responses, resulting in a high recall of 99.59% but close to random-chance accuracy.

In order to evaluate UI element type, we randomly sample 5 UI elements from each UI and query both models to identify the types of these UI elements from a list of 12 common types, including button, checkbox, container, dialog, icon, page control, picture, segmented control, slider, text, text field, and toggle. The list of element types is provided in a randomized order for each query. This is a harder task, and our ILuvUI agent achieves an accuracy of 25.67% compared to 9.03% for LLaVA.

5.2 Human Evaluation of UI Descriptions

For all 100 images in the evaluation set, we queried both models to produce a UI description. We then built a rating user interface in a webpage, where the screen image and descriptions from both models were displayed. The descriptions from each model were shown in a randomized order on the page, so a human rater could not predict which model generated which description. Raters were asked which description they preferred or if both descriptions were about the same. All authors participated as raters for this evaluation. The results show that 72% of the responses preferred the

ILuvUI generated description, in comparison to 20% that preferred LLaVA’s description. The remaining 8% were rated similar.

6 DISCUSSION AND FUTURE WORK

This paper presents an initial promising step towards producing a Vision-Language Model that can perform general UI understanding tasks. Our approach adapts an existing method for the UI domain, and shows that it is able to produce UI-centric instruction-tuned behavior despite being trained with synthetic textual data produced from a large UI dataset of only screenshots.

There are many potential applications for such a model, including providing accessible descriptions of screens to vision-impaired users, repairing automated UI tests, or providing contextual help to assist a user with navigating an interface. We have informally completed some of these tasks successfully, and a basic example of our model being used for UI navigation can be seen in Table 4.

More comprehensive evaluation will be needed to fully understand the performance of this model and its progeny. An open question is: what sort of tasks or benchmarks should be adopted to evaluate the performance of a model on UI tasks? Here we use very basic UI understanding tasks as initial benchmarks, including UI element detection and element type identification. There are further simple evaluations of this form, such as icon recognition. There are also other higher-level tasks along which a model could be evaluated. Spotlight [36] suggests some tasks, including widget captioning, screen summarization, command grounding, and tappability prediction. A goal for the research community should be to work towards agreement on a standard set of benchmarks that can be automatically and easily evaluated when a new models is produced, which might help ensure that others working on models outside of the HCI community produce models that can successfully perform UI understanding tasks.

Other future research could delve deeper into refining the model’s reasoning abilities, expanding its knowledge base for UI design, developing novel agents to facilitate effective UI interaction and navigation, and ensuring robustness across broader UI variations.

Future work can focus on the following aspects to further develop the agent:

- Improving dataset quality: While our generated UI dataset has helped in training our ILuvUI agent, there is still room for improving the quality and diversity of the dataset. A deeper analysis of our generated dataset is needed to better understand its flaws, which can lead to improved prompting strategies. Future work on synthetic data could focus on reducing hallucination, adding more annotated data, and varying UIs from different domains. Human-annotated data could also be added, especially to assist with common-sense reasoning and task planning.
- Improving the model: The performance of ILuvUI could be improved by building on top of more advanced vision-language model architectures. In particular, it would be ideal to build upon an image encoder that can function with a variety of input resolutions, including very high resolutions. Such a model could them more easily work on traditional desktop user interfaces, for example.
- Support for tools and machine-interpretable output: Currently, ILuvUI seems to be fairly limited in its ability to produce structured outputs that could be more easily interpreted by a machine, such as JSON. Some applications, particularly UI navigation and software testing, will require the models to work more easily with other traditional software, so additional training and fine-tuning will be required to make this possible and reliable.

Type	Detailed Description
User	Describe the UI.
Our Response	This UI screen presents weather information for various cities. At the top, the title "Weather" is centrally positioned, and to the right, a settings icon is available for users to adjust their preferences. Below the title, a search field allows users to look for weather information in a specific city or airport. The main part of the screen is occupied by a series of buttons, each representing a different city. Each button includes the city name, the current time, temperature, weather condition, and the high and low temperatures for the day. The cities listed are Toronto, Honolulu, Zurich, and Shanghai. Notably, the city of Toronto is accompanied by a warning for "Heat Warnings". At the bottom of the screen, a text section provides additional information about weather data, encouraging users to learn more about weather data and how to interpret it.
LLaVA Response	The UI of the iPhone screen features a weather app with a variety of weather information displayed. The app provides temperature readings for different cities, including Toronto, Zurich, and Shanghai, as well as the current time. The app also displays a map with the current weather conditions for each city. The UI is designed to be user-friendly and easy to navigate, allowing users to quickly access and view the weather information they need.
Type	Details on UI
User	What is the weather condition of Honolulu?
Our Response	The weather condition of Honolulu is "Partly Cloudy".
LLaVA Response	The weather condition of Honolulu is sunny.
Type	Logical Reasoning
User	If I want to know whether it is raining in London, what should I do?
Our Response	To determine if it is raining in London, you should enter "London" in the search bar. The weather information for London will be displayed, including the current weather conditions. If it is raining, the relevant weather conditions (e.g., "Raining") will be shown.
LLaVA Response	To find out whether it is raining in London, you can check the weather app on your iPhone. Look for the city of London in the list of cities displayed on the app, and check the weather forecast for that day. The app should provide you with the current weather conditions and a forecast for the upcoming days, including whether it is raining or not. If it shows rain for London, then it is likely to rain in London that day.
Type	UI Transition
User	If I click on the top right button, what will happen?
Our Response	If you click on the top right button, it is likely that the settings for the weather information will be shown. This could include changing the city for which the weather information is displayed, adjusting the units of measurement (e.g., Celsius or Fahrenheit), and other settings related to the weather information.
LLaVA Response	If you click on the top right button on the weather app, it will likely open the current weather conditions for the selected city. The button is labeled "Today's weather," which suggests that it will display the current weather information for the city. This could include temperature, humidity, wind speed, and precipitation, among other relevant weather data.

Table 2. Example prompt demonstrating our agent's capability compared to the original LLaVA agent. Manuscript submitted to ACM

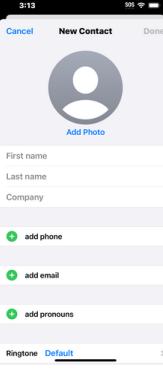
	
Type	Detailed Description
User	Could you provide a detailed description?
Our Response	<p>This UI screen is designed for creating a new contact. At the top of the screen, there are three options: "Cancel" on the left, "New Contact" in the center, and "Done" on the right. These options allow the user to either discard the process, understand the current operation, or complete the contact creation, respectively. In the middle of the screen, there's a section labeled "Add Photo", which suggests the ability to add a photo to the new contact. Below this, there are fields for entering the contact's first and last names. The "First name" field is located above the "Last name" field. Further down, there's a section for entering the contact's company. This section is followed by three buttons for adding various types of contact information. The first button, labeled "Add phone", allows the user to add a phone number. The second button, labeled "Add email", provides an option to add an email address. The third button, labeled "Add pronouns", offers the ability to add any additional pronouns associated with the contact. At the bottom of the screen, there's a button labeled "Ringtone Default", which presumably allows the user to set a default ringtone for this contact. Next to this button, there's a right arrow icon, suggesting that there might be additional options or information available upon interaction.</p>
LLaVA Response	<p>The image shows a close-up view of a smartphone screen displaying a contact list. The contact list is divided into two sections: one section contains the names of contacts, and the other section contains their phone numbers. The contact list is organized in a grid-like layout, with each contact's name and phone number displayed in separate rows. This layout makes it easy for the user to quickly identify and access the contact information they need. The phone number section is highlighted in green, which helps to distinguish it from the contact names section. Overall, the contact list is a useful feature on the smartphone that allows users to easily manage and access their contacts' information.</p>
Type	Logical Reasoning
User	what should I do if I want to add Amy at Company ABC with email address 12345@abc.com and have the Circuit Ringtone when she gives a phone call to me?
Our Response	<p>To add a contact named Amy from Company ABC with the email address 12345@abc.com and set the Circuit Ringtone, you should first enter "Amy" in the "First name" text field and "Company ABC" in the "Company" text field. Then, in the "Email" section, you should enter "12345@abc.com". Scroll down to the "Ringtone" section and click on the "Default" button to set the Circuit Ringtone as the default ringtone for Amy's contact. After setting the ringtone, you can confirm the creation of the contact by clicking on the "Done" button located at the top right of the screen.</p>
LLaVA Response	<p>To add Amy at Company ABC with email address 12345@abc.com and have the Circuit Ringtone when she gives a phone call to you, you should follow these steps: Open the phone app on your device. Navigate to the contact list or address book. Tap the "Add Contact" or "Add New Contact" button. Enter her name, email address (12345@abc.com), and any other relevant information you would like to include. Save the contact by tapping the "Save" or "Add" button. Set the Circuit Ringtone as her ringtone by navigating to the phone settings, selecting "Ringtones," and choosing the Circuit Ringtone. Save the changes to the phone settings.</p>

Table 3. Example prompt demonstrating our agent's capability compared to the original LLaVA agent.
Manuscript submitted to ACM

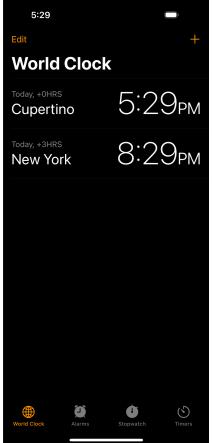
<p>Task: start a stopwatch.</p> 	<p>Human: Describe the screen and its key features, which may be useful for performing various actions on the screen.</p> <p>ILuvUI: [Description of the UI]</p> <p>Human: What action (tap, swipe, enter text) should a user perform on this screen to do the following: start a stopwatch.</p> <p>ILuvUI: To start a stopwatch on this screen, a user should tap on the "Stopwatch" tab located in the middle of the UI screen. This tab is positioned between the "Currency" and "Settings" tabs. By tapping on the "Stopwatch" tab, the user can initiate the stopwatch functionality and start timing events.</p>
	<p>Human: Describe the screen and its key features, which may be useful for performing various actions on the screen.</p> <p>ILuvUI: [Description of the UI]</p> <p>Human: What action (tap, swipe, enter text) should a user perform on this screen to do the following: start a stopwatch.</p> <p>ILuvUI: To start a stopwatch, the user should tap on the "Start" button located below the timer.</p>

Table 4. Example demonstrating our agent’s potential capability of planning UI navigation.

7 CONCLUSION

In this paper we have shown how to address a significant gap in the current abilities of Vision-Language Models to understand UIs. We proposed a data generation and fine-tuning approach that can adapt VLMs to the realm of UIs. The challenges posed by UIs go beyond static elements and encompass interactive components that introduce both structural and functional complexity. Our approach, building on the data generation and training of LLaVA [40], showcases its versatility in capturing diverse responses relevant to UI understanding and interactions, ranging from individual UI element properties to elaborate UI descriptions, reasoning, potential user actions, and dynamic UI transitions. The performance evaluation of our UI-focused instruction-following visual agent highlights its effectiveness in comparison to existing models that are not trained on UI data.

REFERENCES

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangoeei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198 [cs.CV]
- [2] Deniz Arsan, Ali Zaidi, Aravind Sagar, and Ranjitha Kumar. 2021. App-Based Task Shortcuts for Virtual Assistants. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) (*UIST '21*). Association for Computing Machinery, New York, NY, USA, 1089–1099. <https://doi.org/10.1145/3472749.3474808>
- [3] Nikola Banovic, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2012. Waken: Reverse Engineering Usage Information and Interface Structure from Software Videos. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) (*UIST '12*). Association for Computing Machinery, New York, NY, USA, 83–92. <https://doi.org/10.1145/2380116.2380129>
- [4] Tony Beltramelli. 2018. Pix2code: Generating Code from a Graphical User Interface Screenshot. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Paris, France) (*EICS '18*). Association for Computing Machinery, New York, NY, USA, Article 3, 6 pages. <https://doi.org/10.1145/3220134.3220135>
- [5] M. L. BERNARDI, G. A. DI LUCCA, and D. DISTANTE. 2009. RE-UWA approach to recover user centered conceptual models from Web applications. *International Journal on Software Tools for Technology Transfer* 11, 6 (2009), 485–501. <https://doi.org/10.1007/s10009-009-0126-1>
- [6] Pavol Bielik, Marc Fischer, and Martin Vechev. 2018. Robust Relational Layout Synthesis from Examples for Android. *Proc. ACM Program. Lang.* 2, OOPSLA, Article 156 (Oct. 2018), 29 pages. <https://doi.org/10.1145/3276526>
- [7] Tsung-Hsiang Chang, Tom Yeh, and Rob Miller. 2011. Associating the Visual Representation of User Interfaces with Their Internal Structures and Metadata. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) (*UIST '11*). Association for Computing Machinery, New York, NY, USA, 245–256. <https://doi.org/10.1145/2047196.2047228>
- [8] Chunyang Chen, Sidong Feng, Zhenchang Xing, Linda Liu, Shengdong Zhao, and Jinshui Wang. 2019. Gallery D.C.: Design Search and Knowledge Discovery through Auto-Created GUI Component Gallery. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 180 (nov 2019), 22 pages. <https://doi.org/10.1145/3359282>
- [9] Chunyang Chen, Ting Su, Guozhu Meng, Zhenchang Xing, and Yang Liu. 2018. From UI Design Image to GUI Skeleton: A Neural Machine Translator to Bootstrap Mobile GUI Implementation. In *Proceedings of the 40th International Conference on Software Engineering* (Gothenburg, Sweden) (*ICSE '18*). Association for Computing Machinery, New York, NY, USA, 665–676. <https://doi.org/10.1145/3180155.3180240>
- [10] Jieshan Chen, Chunyang Chen, Zhenchang Xing, Xin Xia, Liming Zhu, John Grundy, and Jinshui Wang. 2020. Wireframe-Based UI Design Search through Image Autoencoder. *ACM Trans. Softw. Eng. Methodol.* 29, 3, Article 19 (jun 2020), 31 pages. <https://doi.org/10.1145/3391613>
- [11] Jieshan Chen, Mulong Xie, Zhenchang Xing, Chunyang Chen, Xiwei Xu, Liming Zhu, and Guoqiang Li. 2020. Object Detection for Graphical User Interface: Old Fashioned or Deep Learning or a Combination?. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Virtual Event, USA) (*ESEC/FSE 2020*). Association for Computing Machinery, New York, NY, USA, 1202–1214. <https://doi.org/10.1145/3368089.3409691>
- [12] Sen Chen, Lingling Fan, Chunyang Chen, Minhui Xue, Yang Liu, and Lihua Xu. 2021. GUI-Squatting Attack: Automated Generation of Android Phishing Apps. *IEEE Transactions on Dependable and Secure Computing* 18, 6 (2021), 2551–2568. <https://doi.org/10.1109/TDSC.2019.2956035>
- [13] Sen Chen, Lingling Fan, Ting Su, Lei Ma, Yang Liu, and Lihua Xu. 2019. Automated Cross-Platform GUI Code Generation for Mobile Apps. In *2019 IEEE 1st International Workshop on Artificial Intelligence for Mobile (AI4Mobile)*. 13–16. <https://doi.org/10.1109/AI4Mobile.2019.8672718>
- [14] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Kieran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. arXiv:2209.06794 [cs.CV]
- [15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023).
- [16] Allen Cypher. 1991. EAGER: Programming Repetitive Tasks by Example. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana, USA) (*CHI '91*). Association for Computing Machinery, New York, NY, USA, 33–39. <https://doi.org/10.1145/108844.108850>
- [17] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV]
- [18] Daniel de Souza Baulé, Christiane Gresse von Wangenheim, Aldo von Wangenheim, Jean C. R. Hauck, and Edson C. Vargas Júnior. 2021. Using Deep Learning to Support the User Interface Design of Mobile Applications with App Inventor. In *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems* (Virtual Event, Brazil) (*IHCB '21*). Association for Computing Machinery, New York, NY, USA, Article 49, 11 pages. <https://doi.org/10.1145/3472301.3484340>
- [19] Biplob Deka, Zifeng Huang, Chad Franzen, Joshua Hirschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A Mobile App Dataset for Building Data-Driven Design Applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) (*UIST '17*). Association for Computing Machinery, New York, NY, USA, 845–854. <https://doi.org/10.1145/3126594.3126651>

- [20] G.A. Di Lucca, P. Fasolino, A.R.and IGLINSKI, and P. Tramontana. 2004. Reverse engineering Web applications: the WARE approach. *SOFTWARE MAINTENANCE AND EVOLUTION: RESEARCH AND PRACTICE* 12 (2004), 71–101. <https://doi.org/10.1002/smrv.281>
- [21] Morgan Dixon and James Fogarty. 2010. Prefab: Implementing Advanced Behaviors Using Pixel-Based Reverse Engineering of Interface Structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (*CHI '10*). Association for Computing Machinery, New York, NY, USA, 1525–1534. <https://doi.org/10.1145/1753326.1753554>
- [22] Morgan Dixon, Daniel Leventhal, and James Fogarty. 2011. Content and Hierarchy in Pixel-Based Methods for Reverse Engineering Interface Structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (*CHI '11*). Association for Computing Machinery, New York, NY, USA, 969–978. <https://doi.org/10.1145/1978942.1979086>
- [23] Morgan Dixon, Alexander Nied, and James Fogarty. 2014. Prefab Layers and Prefab Annotations: Extensible Pixel-Based Interpretation of Graphical Interfaces. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) (*UIST '14*). Association for Computing Machinery, New York, NY, USA, 221–230. <https://doi.org/10.1145/2642918.2647412>
- [24] W. Keith Edwards, Elizabeth D. Mynatt, and Kathryn Stockton. 1995. Access to Graphical Interfaces for Blind Users. *Interactions* 2, 1 (jan 1995), 54–67. <https://doi.org/10.1145/208143.208161>
- [25] Shirin Feiz, Jason Wu, Xiaoyi Zhang, Amanda Sweenyng, Titus Barik, and Jeffrey Nichols. 2022. Understanding Screen Relationships from Screenshots of Smartphone Applications. In *27th International Conference on Intelligent User Interfaces*. 447–458.
- [26] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '20*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376728>
- [27] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions Using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3173574.3174092>
- [28] Forrest Huang, John F. Cann, and Jeffrey Nichols. 2019. Swire: Sketch-Based User Interface Retrieval. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3290605.3300334>
- [29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. <https://doi.org/10.5281/zenodo.5143773> If you use this software, please cite it as below.
- [30] Yue Jiang, Ruofei Du, Christof Lutteroth, and Wolfgang Stuerzlinger. 2019. ORC Layout: Adaptive GUI Layout with OR-Constraints. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, Article 413, 12 pages. <https://doi.org/10.1145/3290605.3300643>
- [31] Yue Jiang, Wolfgang Stuerzlinger, and Christof Lutteroth. 2021. *ReverseORC: Reverse Engineering of Resizable User Interface Layouts with OR-Constraints*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411764.3445043>
- [32] Yue Jiang, Wolfgang Stuerzlinger, Matthias Zwicker, and Christof Lutteroth. 2020. ORCSolver: An Efficient Solver for Adaptive GUI Layout with OR-Constraints. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '20*). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376610>
- [33] T. Katsimpila, Y. Panagis, E. Sakkopoulos, G. Tzimas, and A. Tsakalidis. 2006. Application Modeling using Reverse Engineering Techniques. In *Proceedings of the 2006 ACM symposium on applied computing*. ACM, 1250–1255. <https://doi.org/10.1145/1141277.1141570>
- [34] R. Krosnick, S. W. Lee, W. S. Laseck, and S. Onev. 2018. Expresso: Building Responsive Interfaces with Keyframes. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 39–47. <https://doi.org/10.1109/VLHCC.2018.8506516>
- [35] Gilly Leshed, Eben M. Haber, Tara Matthews, and Tessa Lau. 2008. CoScripter: Automating & Sharing How-to Knowledge in the Enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) (*CHI '08*). Association for Computing Machinery, New York, NY, USA, 1719–1728. <https://doi.org/10.1145/1357054.1357323>
- [36] Gang Li and Yang Li. 2022. Spotlight: Mobile UI Understanding using Vision-Language Models with a Focus. *ArXiv* abs/2209.14927 (2022). <https://api.semanticscholar.org/CorpusID:252595735>
- [37] Toby Jia-Jun Li, Amos Azaria, and Brad A. Myers. 2017. SUGILITE: Creating Multimodal Smartphone Automation by Demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (*CHI '17*). Association for Computing Machinery, New York, NY, USA, 6038–6049. <https://doi.org/10.1145/3025453.3025483>
- [38] Yi-Chi Liao, Kashyap Todi, Aditya Acharya, Antti Keurulainen, Andrew Howes, and Antti Oulasvirta. 2022. Rediscovering Affordance: A Reinforcement Learning Perspective. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 362, 15 pages. <https://doi.org/10.1145/3491102.3501992>
- [39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).
- [41] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhou Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2023. Chatting with GPT-3 for Zero-Shot Human-Like Mobile Automated GUI Testing. *arXiv:2305.09434* [cs.SE]

- [42] Christof Lutteroth. 2008. Automated Reverse Engineering of Hard-Coded GUI Layouts. In *Proceedings of the Ninth Conference on Australasian User Interface - Volume 76* (Wollongong, Australia) (*AUIC '08*). Australian Computer Society, Inc., AUS, 65–73. <https://doi.org/10.5555/1378337.1378350>
- [43] Melody Moore and Spencer Rugaber. 1997. Using Knowledge Representation to Understand Interactive Systems. In *Proceedings of the 5th International Workshop on Program Comprehension (WPC '97)* (*WPC '97*). IEEE Computer Society, USA, 60.
- [44] Melody M. Moore. 1996. Rule-Based Detection for Reverse Engineering User Interfaces. In *Proceedings of the 3rd Working Conference on Reverse Engineering (WCRE '96)* (*WCRAE '96*). IEEE Computer Society, USA, 42.
- [45] Melody Marie Moore, James D. Foley, and Spencer Rugaber. 1998. *User Interface Reengineering*. Ph.D. Dissertation. USA. AAI9918460.
- [46] Kevin Moran, Carlos Bernal-Cárdenas, Michael Curcio, Richard Bonett, and Denys Poshyvanyk. 2020. Machine Learning-Based Prototyping of Graphical User Interfaces for Mobile Apps. *IEEE Transactions on Software Engineering* 46, 2 (2020), 196–221. <https://doi.org/10.1109/TSE.2018.2844788>
- [47] Tuan Anh Nguyen and Christoph Csallner. 2015. Reverse Engineering Mobile Application User Interfaces with REMAUI. In *Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering (Lincoln, Nebraska) (ASE '15)*. IEEE Press, 248–259. <https://doi.org/10.1109/ASE.2015.32>
- [48] OpenAI. [n.d.]. ChatGPT. <https://openai.com/blog/chatgpt>.
- [49] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
- [50] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155* [cs.CL]
- [51] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).
- [52] Suporn Pongnumkul, Mira Dontcheva, Wilmot Li, Jue Wang, Lubomir Bourdev, Shai Avidan, and Michael F. Cohen. 2011. Pause-and-Play: Automatically Linking Screencast Video Tutorials with Applications. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) (*UIST '11*). Association for Computing Machinery, New York, NY, USA, 135–144. <https://doi.org/10.1145/2047196.2047213>
- [53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [54] A. Sanchez Ramon, J. Sanchez Cuadrado, and J. Garcia Molina. 2012. Model-driven reverse engineering of legacy graphical user interfaces. *Automated Software Engineering* 21, 2 (2012), 147–186. <https://doi.org/DOI:10.1007/s10515-013-0130-2>
- [55] Eldon Schoop, Xin Zhou, Gang Li, Zhourong Chen, Bjoern Hartmann, and Yang Li. 2022. Predicting and Explaining Mobile UI Tappability with Vision Modeling and Saliency Analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 36, 21 pages. <https://doi.org/10.1145/3491102.3517497>
- [56] E. Stroulia, M. El-Ramly, P. Iglesias, and P. Sorenson. 2003. User Interface Reverse Engineering in Support of Interface Migration to the Web. *Automated Software Engg.* 10, 3 (July 2003), 271–301. <https://doi.org/10.1023/A:1024460315173>
- [57] Amanda Swearngin, Mira Dontcheva, Wilmot Li, Joel Brandt, Morgan Dixon, and Amy J. Ko. 2018. Rewire: Interface Design Assistance from Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3174078>
- [58] Amanda Swearngin, Amy J. Ko, and James Fogarty. 2017. Genie: Input Retargeting on the Web through Command Reverse Engineering. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (*CHI '17*). Association for Computing Machinery, New York, NY, USA, 4703–4714. <https://doi.org/10.1145/3025453.3025506>
- [59] Amanda Swearngin and Yang Li. 2019. Modeling Mobile Interface Tappability Using Crowdsourcing and Deep Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3290605.3300305>
- [60] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
- [61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [62] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI Using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 432, 17 pages. <https://doi.org/10.1145/3544548.3580895>
- [63] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2023. Empowering LLM to use Smartphone for Intelligent Task Automation. *arXiv:2308.15272* [cs.AI]
- [64] Thomas D. White, Gordon Fraser, and Guy J. Brown. 2019. Improving Random GUI Testing with Image-Based Widget Detection. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis* (Beijing, China) (*ISSTA 2019*). Association for Computing Machinery, New York, NY, USA, 307–317. <https://doi.org/10.1145/3293882.3330551>
- [65] Jason Wu, Xiaoyi Zhang, Jeff Nichols, and Jeffrey P Bigham. 2021. Screen Parsing: Towards Reverse Engineering of UI Models from Screenshots. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) (*UIST '21*). Association for Computing Machinery, Manuscript submitted to ACM

- New York, NY, USA, 470–483. <https://doi.org/10.1145/3472749.3474763>
- [66] Tom Yeh, Tsung-Hsiang Chang, and Robert C. Miller. 2009. Sikuli: Using GUI Screenshots for Search and Automation. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology* (Victoria, BC, Canada) (*UIST '09*). Association for Computing Machinery, New York, NY, USA, 183–192. <https://doi.org/10.1145/1622176.1622213>
 - [67] Chen Yongxin, Zhang Tonghui, and Chen Jie. 2019. UI2code: How to fine-tune background and foreground analysis. *Retrieved Feb 23 (2019)*, 2020.
 - [68] Luke S. Zettlemoyer and Robert St. Amant. 1999. A Visual Medium for Programmatic Control of Interactive Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) (*CHI '99*). Association for Computing Machinery, New York, NY, USA, 199–206. <https://doi.org/10.1145/302979.303039>
 - [69] Xiaoyi Zhang, Lilian de Greef, Amanda Sweenyngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, et al. 2021. Screen recognition: Creating accessibility metadata for mobile applications from pixels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
 - [70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).