Controllable 3D UI Exploration and it's Future

Md. Akil Raihan Iftee
Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Bangladesh
iftee1807002@stud.kuet.ac.bd

I. Overview

Unlike traditional 2D UIs, Augmented Reality (AR) interfaces present 3D components anchored in the physical world, requiring spatial reasoning and new data modalities. While foundational 2D GUI datasets like *Rico* [1] and *WebUI-25K* [2] exist, no equivalent standardized 3D AR interface datasets currently exists. Existing datasets focus on: ShapeNet (CAD models), ScanObjectNN (real-world scans), and Matterport3D (indoor scenes) but no UI-specific metadata.

To address this, we propose **"Controllable 3D UI Exploration"** approach focus only on user interfaces for mobile or web contexts, and our contributions include:

- A method to collect and annotate 3D AR UI data.
- A synthetic dataset of AR interfaces with conversational annotations.
- An instruction-tuned Vision-Language Model (VLM) for AR UI understanding, enabling tasks such as element detection, action prediction, and multi-step UI planning in AR contexts.

II. Dataset Creation

We define a data collection and annotation pipeline for AR UI scenes, including both real captures and synthetic scenes.

Real-UI Capture: We build an AR app (via ARKit/ARCore) that lets users place UI elements (buttons, sliders, text, etc.) in real environments. During scene creation, we record each element's spatial transform and capture the RGB camera view.

Synthetic Generation: Using a 3D engine (e.g., Unity [3]), we programmatically generate AR scenes with a library of 3D UI models. We randomize layouts (e.g., on walls, tables, or floating), labels, and backgrounds to create diverse and scalable synthetic data.

Annotation Format: For each AR UI scene, we annotate every UI component with its type (e.g., "button", "slider", "text label"), text content (if any, as extracted by OCR or known from generation), and its 3D bounding box coordinates in world space, defined from corner (x_1, y_1, z_1) to (x_2, y_2, z_2).

III. Methodology

Conversational Data Generation: We adapt the LLM to create conversational examples about AR UIs. Using a 3D object detector (VoteNet [4]) and GPT-4/GPT-3.5, we generate multi-turn dialogues covering UI descriptions, Q&A, action lists, outcome predictions, and UI generation tasks.

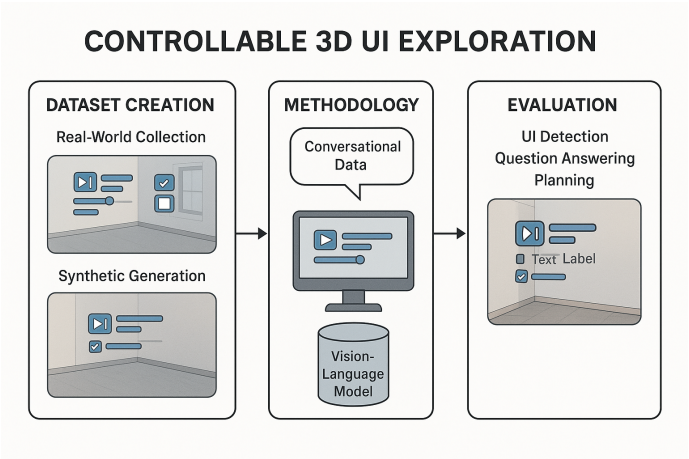

Fig. 1: The Framework of Proposed Gaze-Enhanced Multimodal Interaction

Vision-Language Model: We fine-tune an instruction-following VLM (CLIP visual encoder + Vicuna-13B) on this AR UI data. The resulting model answers AR UI questions and can propose UI layouts from language.

IV. Evaluation

We evaluate the model on AR-specific UI tasks:

- **Detecting UI elements** from an AR view.
- **Answering element-level questions** about the AR scene.
- **Planning multi-step interactions** with the interface.

We also consider its use for AR UI design:

- The model can **generate AR UI specifications** in response to natural-language goals.
- This supports **usability research** and design exploration.

V. Further Extension

Future work may extend this with [5], [6] research through **FlexDoc-3D**, enabling adaptive optimization of both 3D UI content and spatial layout in AR environments. Similarly, **Dreamstruct-3D** could synthesize rich 3D UI and slide datasets, enhancing multimodal understanding for instruction-tuned language-vision models in immersive interfaces.

References

[1] B. Deka, Z. Huang, C. Franzen, J. Hibschman, M. Afergan, Y. Li, J. Nichols, and R. Kumar, "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*, pp. 845–854, ACM, Oct. 2017.

[2] J. Wu, S. Wang, S. Shen, Y.-H. Peng, J. Nichols, and J. P. Bigham, "Webui: A dataset for enhancing visual ui understanding with web semantics," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, ACM, Apr. 2023.
[3] Unity Technologies, "Unity real-time development platform," 2025. https://unity.com/.
[4] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9277–9286, 2019.
[5] Y. Peng, F. Huq, Y. Jiang, J. Wu, X. Li, J. Bigham, and A. Pavel, "Dreamstruct: Understanding slides and user interfaces via synthetic data generation," in *European Conference on Computer Vision*, (Cham), pp. 466–485, Springer Nature Switzerland, September 2024.
[6] Y. Jiang, C. Lutteroth, R. Jain, C. Tensmeyer, V. Manjunatha, W. Stuerzlinger, and V. Morariu, "Flexdoc: Flexible document adaptation through optimizing both content and layout," in *Proceedings of the 2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 217–222, IEEE, September 2024.