Supplementary Material of SeekUI: Predicting Visual Search Behavior on Graphical User Interfaces with a Reward-Augmented Vision Language Model

1 OVERVIEW

This supplementary material mainly provides three contents: (1) the implementation details of how we train and evaluate SeekUI, facilitating reproducibility; (2) the examples illustrating how we extract sequences of UI elements from explanation–scanpath pairs; and (3) more examples of layout optimization for target search efficiency.

2 IMPLEMENTATION DETAILS

We provide the following details for SeekUI: (1) the first training stage, (2) the second training stage, and (3) the evaluation.

2.1 Details on Stage (1): Instruction Tuning

We initialize SeekUI using Qwen2.5-VL-7B-Instruct [1]. During this stage, we freeze the Vision Transformer encoder while fine-tuning both the MLP-based projector and the LLM backbone. The training spans 100 epochs using the Adam optimizer. We employ a cosine learning rate scheduler with a warmup ratio of 0.03 and a weight decay of 0.01. Notably, we apply different learning rates: 1×10^{-5} for the projector and 2×10^{-7} for the LLM. The global batch size is set to 16, achieved with 2 gradient accumulation steps. For infrastructure, we utilize Flash Attention 2 [2] and bfloat16 precision for memory optimization, alongside DeepSpeed ZeRO-3 [3] for efficient state sharding. The process was completed in approximately 8 hours on a single NVIDIA H200 GPU.

2.2 Details on Stage (2): Reinforcement Learning

Building upon the instruction tuning stage, we further align SeekUI using reinforcement learning. The training is conducted for 20 epochs with a global batch size of 32 and an initial learning rate of 1×10^{-6} utilizing a linear decay scheduler. We set the group size $G = 4$, sampling four distinct candidate responses per prompt to compute the normalized advantages (see Eq. 11). The model trained with instruction tuning stage serves as the reference policy (π^{ref}), and we set the KL-divergence coefficient $\beta = 0.01$ (see Eq. 12) to ensure the policy remains stable relative to the reference. To manage computational costs for the SeekUI, the maximum generation length is capped at 512 tokens. For infrastructure, we utilize Flash Attention 2 and bfloat16 precision to optimize memory usage. Furthermore, DeepSpeed ZeRO-3 is employed to shard model states and gradients across GPUs. This training stage was completed in approximately 24 hours using 32 AMD MI250x GPUs.

2.3 Details on Evaluation

To mitigate potential bias arising from varying image resolutions, we standardize all scanpaths to a fixed resolution of 512×384. This ensures that scanpaths from larger images do not disproportionately influence the evaluation. For the ScanMatch calculation, we discretize this space into 16×12 bins (horizontal × vertical). Given the standardized resolution, this results in a uniform spatial bin size of 32×32 pixels. Furthermore, to evaluate saliency-based metrics, we transform the discrete scanpaths into continuous saliency heatmaps. This is achieved by convolving the fixation points with a Gaussian filter, setting the standard deviation σ to 22.4 pixels in both horizontal and vertical directions.

3 UI ELEMENT EXTRACTION FROM EXPLANATIONS AND SCANPATHS

Figures 1 and 2 present examples of UI element extraction from the given explanation-scanpath pairs. It demonstrates that the structure of the explanation elements aligns with the visual exploration path elements.

4 EXAMPLES OF LAYOUT OPTIMIZATION FOR TARGET SEARCH EFFICIENCY

As demonstrated in Figure 1, SeekUI supports a global optimization workflow. To evaluate search efficiency, the system takes specific targets, such as Academics" and Union College" in this instance, and generates scanpaths for each layout variant. This process quantifies the search cost associated with different spatial arrangements, providing designers with a direct metric to compare and select the most efficient layout candidates.

REFERENCES

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023).

[3] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–16.

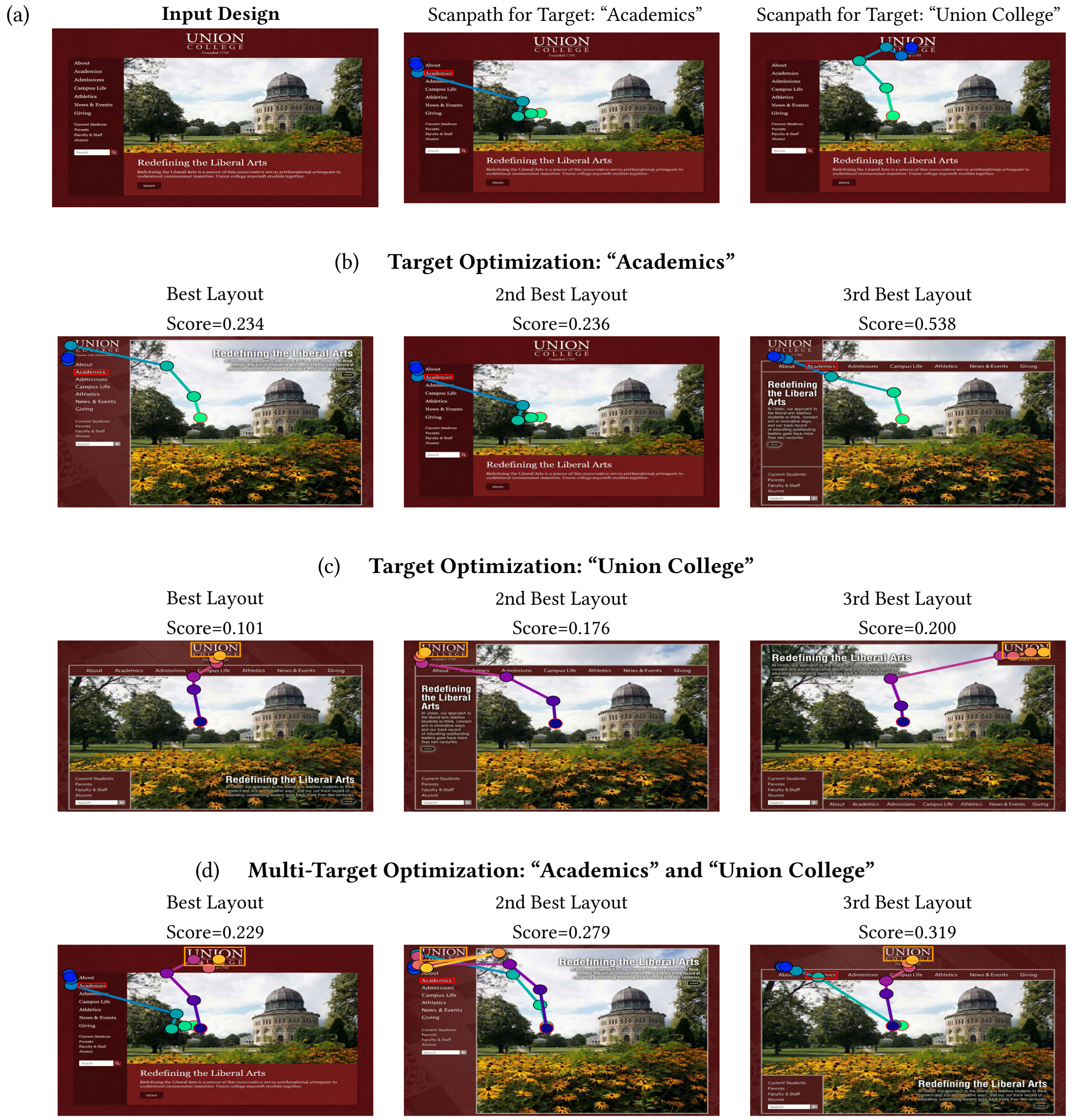
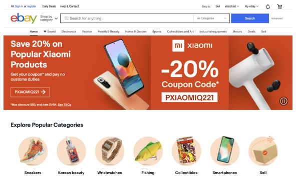
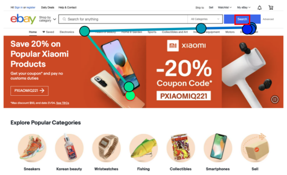


Figure 1: Application of layout optimization for target search efficiency. (a) The original interface and SeekUI’s predicted scanpaths for two targets (“Academics” and “Union College”). (b) Top-ranked reflowed layouts that minimize predicted search time for “Academics” alone. (c) Top-ranked reflowed layouts that minimize predicted search time for “Union College” alone. (d) Layouts selected through joint optimization, minimizing the combined search-efficiency score for both targets simultaneously. Lower scores indicate more efficient predicted visual search.

<table>
<tr><th>Input UI</th><th>Scanpath on UI</th></tr>
<tr><td>
</td><td>
ebay
Search
Save 20% on Popular Xiaomi Products
xiaomi
-20%
Coupon Code*
PXIAOMIQ221
Explore Popular Categories

</td></tr>
<tr><th colspan="2">Explanation</th></tr>
<tr><td colspan="2">The scanpath begins with the user focusing on the central promotional banner featuring Xiaomi products, likely due to its prominent placement and eye-catching design. From there, their gaze naturally moves upward towards the top navigation bar, where various categories such as Electronics, Fashion, Health & Beauty, Home & Garden, Sports, Collectibles & Art, Industrial equipment, Motors, Deals, and Sell are displayed. The user may briefly scan these options to determine if they need to navigate to another section but quickly realizes they are already in the desired area. Next, the user's attention shifts slightly further right along the top navigation bar, possibly searching for additional filtering or sorting options related to their search query. Finally, after briefly exploring the surrounding elements, the user's focus returns to the central promotional banner, reinforcing its relevance to their current task. Throughout this process, the user demonstrates a logical progression from the main content area to the navigation tools, indicating a systematic approach to locating the target "Search" function efficiently.</td></tr>
<tr><th colspan="2">UI Element Extraction</th></tr>
<tr><th>Explanation</th><th>Scanpath</th></tr>
<tr><td>Xiaomi product, navigation bar, Electronics, Fashion, Health & Beauty, Home & Garden, Sports, Collectibles & Art, Industrial equipment, Motors, Deals, Sell, banner</td><td>Xiaomi product, navigation bar, Fashion, Health & Beauty, Home & Garden, Sports, Collectibles and Art, Industrial equipment, Motors, Deals, Sell, Search</td></tr>
</table>

Table 1: The example of UI element extraction from explanations and scanpaths, given the target "Search".

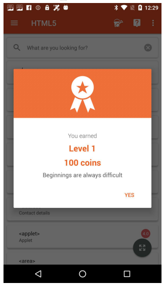
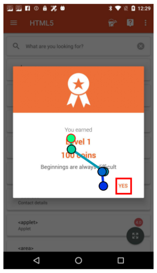
<table>
<tr><th>Input UI</th><th>Scanpath on UI</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><th colspan="2">Explanation</th></tr>
<tr><td colspan="2">The scanpath begins with the user focusing on the central message area, which prominently displays the achievement of earning Level 1 and 100 coins. This message naturally draws attention due to its bold text and central placement. From there, the user's gaze moves slightly downward, exploring the surrounding context provided by the phrase "Beginnings are always difficult." This movement suggests an attempt to understand the implications or next steps related to the achieved level. Subsequently, the user shifts focus further down towards the bottom-right corner of the interface, where the word "YES" appears next to a small icon. This final fixation indicates a deliberate search for actionable options or confirmation prompts associated with the achieved milestone. Throughout this sequence, the user demonstrates a logical progression from recognizing the primary accomplishment to seeking additional information or interaction opportunities, culminating in locating the target "YES."</td></tr>
<tr><th colspan="2">UI Element Extraction</th></tr>
<tr><th>Explanation</th><th>Scanpath</th></tr>
<tr><td>Level 1, 100 coins, Beginnings are always difficult, YES, a small icon</td><td>Level 1, 100 coins, Beginnings are always difficult, YES</td></tr>
</table>

Table 2: The example of UI element extraction from explanations and scanpaths, given the target "YES".