How do Visual Attributes Influence Web Agents? A Comprehensive Evaluation of User Interface Design Factors

Kuai Yu[2,*], Naicheng Yu[3,*], Han Wang[1], Rui Yang[1], Huan Zhang[1],
[1]University of Illinois Urbana-Champaign, [2]Columbia University, [3]University of California San Diego
*** Equal contributions. Correspondence:** ky2589@columbia.edu, n7yu@ucsd.edu, hanw14@illinois.edu

Abstract

Web agents have demonstrated strong performance on a wide range of web-based tasks. However, existing research on the effect of environmental variation has mostly focused on robustness to adversarial attacks, with less attention to agents' preferences in benign scenarios. Although early studies have examined how textual attributes influence agent behavior, a systematic understanding of how visual attributes shape agent decision-making remains limited. To address this, we introduce VAF, a controlled evaluation pipeline for quantifying how webpage visual attribute factors influence web-agent decision-making. Specifically, VAF consists of three stages: (i) variant generation, which ensures the variants share identical semantics as the original item while only differ in visual attributes; (ii) browsing interaction, where agents navigate the page via scrolling and clicking the interested item, mirroring how human users browse online; (iii) validating through both click action and reasoning from agents, which we use the Target Click Rate and Target Mention Rate to jointly evaluate the effect of visual attributes. By quantitatively measuring the decision-making difference between the original and variant, we identify which visual attributes influence agents' behavior most. Extensive experiments, across 8 variant families (48 variants total), 5 real-world websites (including shopping, travel, and news browsing), and 4 representative web agents, show that background color contrast, item size, position, and card clarity have a strong influence on agents' actions, whereas font styling, text color, and item image clarity exhibit minor effects. Our code is available at https://github.com/ASTRAL-Group/WebAgent_Visual_Attribution.git.

1 Introduction

Vision-Language Models (VLMs) based web agents have recently demonstrated strong capabilities (Yao et al., 2022; Jimenez et al., 2023; Zhai et al., 2024; Wang et al., 2024a), enabling a broad spectrum of practical web applications (e.g., web browsing (Zheng et al., 2024; Gur et al., 2023), online shopping (He et al., 2024; Wang et al., 2024b), travel booking (Deng et al., 2023), etc). These VLM-based web agents can interpret user instructions and carry out multi-step web interactions by clicking, typing, and navigating pages, automating the human online decision-making process.

Recent work has begun to examine how variations in web environments affect the decision-making of web agents (Lù et al., 2025; Ning et al., 2025). Nevertheless, most prior works focus on the robustness against environment-side adversarial attacks (Chiang et al., 2025; Zhang et al., 2025; Xu et al., 2024; Evtimov et al., 2025; Yang et al., 2025b), leaving agents' intrinsic decision-making preference in benign scenarios underexplored. Allouah et al. (2025) provides an early exploration in e-commerce, analyzing textual factors (e.g., price, ratings, and reviews) to identify which signals web agents prioritize during shopping. Yet, in practice, real webpages encompass substantially more than texts: agents interpret and act under diverse visual element cues, including spatial layout, element placement, and stylistic choices. Despite that several works in Human Computer Interaction (HCI) community already provide comprehensive studies on how humans react to different visual elements in the web (Wu and Yuan, 2003; Leiva et al., 2020; Soegaard, 2020) and draw some interesting conclusions, for example, humans are more likely to be influenced by color highlights and position (Pernice, 2018; Ng et al., 2024), the systematic study of the effects on VLM-based web agents' decision-making remains limited. This motivates the question: *How do different webpage visual attributes influence agents' decision-making?*

To this regard, we introduce VAF, a controlled evaluation pipeline for quantifying how webpage visual attribute factors influence VLM-based web-

Findings of the Association for Computational Linguistics: ACL 2026, pages 34520–34536

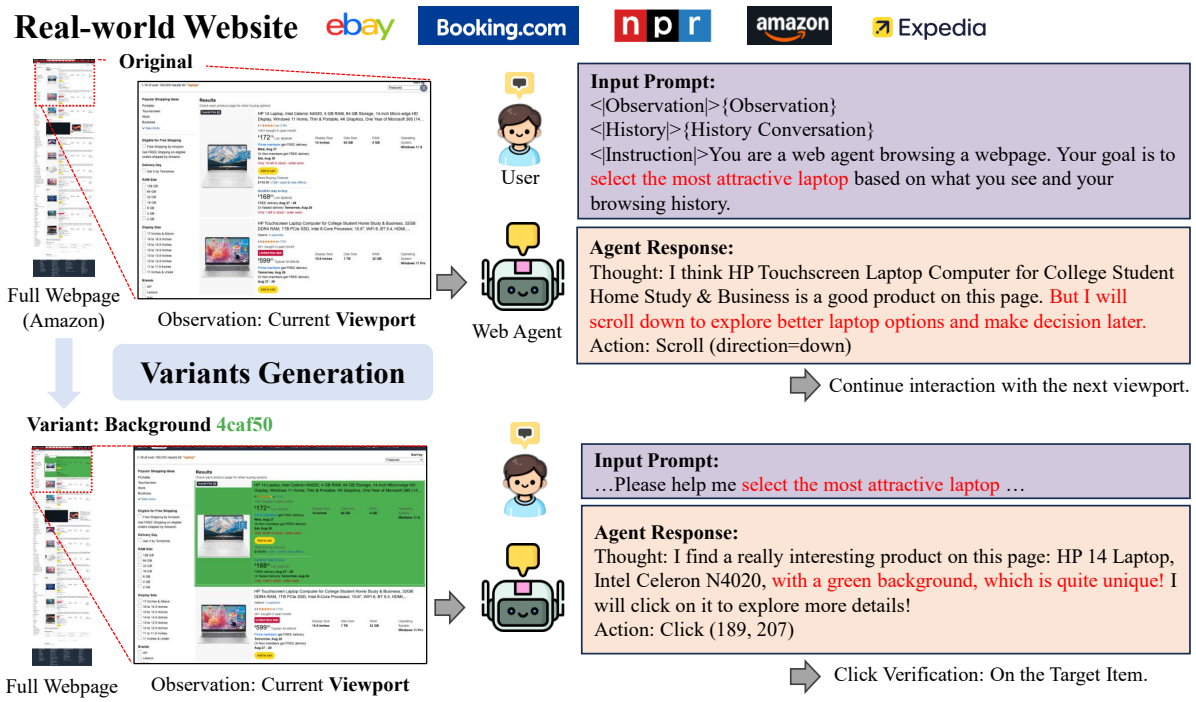


Figure 1: Overview of VAF. Our evaluation pipeline consists of three stages: (i) variant generation, where we construct content-preserving visual variants by modifying the CSS of a designated target item on real-world webpages; (ii) human-like browsing interaction, where agents navigate the page via scrolling and clicking, resembling human browsing behavior; (iii) validating through both click and reasoning from agents, which we use the Target Click Rate and Target Mention Rate to jointly evaluate the effect of visual attributes. By quantitatively comparing agent behavior on the original webpage versus its visually modified variants, we measure how visual attributes influence web-agent decision-making.

agent decision-making. VAF consists of three stages: (i) variant generation, which constructs content-preserving webpage variants by applying CSS-only modification to the designated target item on real-world websites; (ii) human-like browsing interaction, where agents navigate the page via scrolling and clicking, mirroring how human users browse and compare items online; and (iii) validating through both action click and reasoning from agents. VAF constructs webpage variants that are semantically identical to the original page but visually distinct, enabling controlled comparisons between an agent's behavior on the original interface and on perturbed variants. VAF generates 8 variant families, 48 variants in total that cover common visual factors in webpage design for each target item, spanning background color, position, ordering, font styling, etc. During interaction, the agent observes a sequence of viewport images as it scrolls and can issue click actions to select items for further inspection, simulating how humans browse online. We quantify the effect of each visual attribute by comparing both the agent's target-item click rate and target mention rate (i.e., whether the agent's generated reasoning trace mentions the target item) between each variant and the original page, identifying which visual factors strongly influence agent choices. The main contributions are as follows:

- We introduce VAF, a controlled evaluation pipeline with extensive experiments across 8 variant families (48 variants per target item), 5 real-world websites (including shopping, travel, and news browsing), and 4 representative web agents, for systematically measuring how visual attribute factors influence web-agent decision-making.
- VAF proposes a human-like browsing interaction in which agents browse the website through scrolling and clicking over a long webpage via sequential viewport observations, resembling human browsing behavior.
- We find that background color contrast, item size, item position, and card clarity have a strong influence on agents' decision-making, whereas font styling, text color, and item image clarity exhibit minor effects.

2 Related Work

Web Agents. Web agents demonstrate strong abilities to interact with the web environments (Wu et al., 2025), which are typically categorized into text-based and multimodal agents based on perception modality. Text-based agents operate on structured representations like HTML code and accessibility trees, enabling precise symbolic planning but lacking visual awareness (Yang et al., 2024; Deng et al., 2023; Zhou et al., 2023; Chezelles et al., 2024). VLM-based web agents extend perception to rendered screenshots, enabling reasoning over layout, color, and visual salience (Yu et al., 2025; Shen et al., 2025; Koh et al., 2024; He et al., 2024). Our work focuses on the VLM-based web agent side, systematically studying how visual attributes impact decision-making.

Effect of environment variants on agents. Prior work on how web-environment variations affect agent decision-making has largely focused on robustness to adversarial attacks (Liao et al., 2024). Prompt-injection attacks showed that malicious webpage textual content can hijack agents via indirect instructions embedded in text or code (Evtimov et al., 2025; Wang et al., 2025a; Debenedetti et al., 2024; Levy et al., 2024). Complementary studies reveal vision-side vulnerabilities of web agents (Wang et al., 2025b; Zhang et al., 2025; Wang et al., 2025b; Xu et al., 2024; Yang et al., 2025b; Chiang et al., 2025). By comparison, fewer works investigate agents' benign preferences over textual attributes (Allouah et al., 2025; Cherep et al., 2025), and the role of visual attributes in shaping decisions remains underexplored.

Web Design for Humans. Prior HCI and cognitive psychology research shows that visual attributes strongly influence human decision-making. Users form preference judgments within seconds based on interface (Fogg et al., 2001). Color contrast and element size are dominant drivers of attention, as vivid colors and larger elements attract earlier fixation, improve detection accuracy, and substantially alter user behavior and conversion rates (Soegaard, 2020; Wu and Yuan, 2003; Ng et al., 2024; Porter, 2017). Layout and position further guide attention: users exhibit strong positional priors, focusing on central regions while ignoring banners and sidebars, and predictable hierarchies improve comprehension while excessive options increase cognitive load (Pernice, 2018; Nielsen, 1999; Hick, 1952). Page structure affects exploration depth, while reduced clarity, blur, and poor typography impair recognition and increase perceived task difficulty (Yoshihara et al., 2023; Song and Schwarz, 2008; Krause, 2022). Together, these findings motivate us to study whether web agents share similar perceptual preferences toward visual attributes to those of human beings.

3 Visual Attribute Factors (VAF)

We propose VAF, a controlled evaluation pipeline for quantifying how webpage visual attribute factors influence web-agent decision-making. Specifically, we implement the pipeline with three stages: (i) variant generation; (ii) realistic human-like browsing interaction; (iii) validation through both click actions and reasoning traces of agents. Finally, we quantify the decision-making difference with Target Click Rate and Target Mention Rate.

3.1 Variant Generation

Starting from HTML snapshots of real-world webpages, we designate the target item and generate visual attribute variants. Specifically, to simulate the realistic agent-website interaction, the variant generation pipeline is built on pages drawn from five popular websites: (1) laptops on Amazon, (2) headphones on eBay, (3) hotels in San Francisco on Booking, (4) hotels near Yellowstone Park on Expedia, and (5) news articles on NPR. Together, these pages cover three common web-agent settings: online shopping, booking, and news browsing.

For each page, we select one target item and modify its presentation via CSS while preserving the HTML content and functionality, so that the modified item remains semantically equivalent to the original but differs in visual appearance, as shown in Fig. 1. This design ensures that the original and modified pages differ only in the target item's visual attributes. By comparing agent behavior between the original page and its visually perturbed variants, we are able to attribute changes in click behavior to visual attribute modification rather than textual differences. We instantiate 8 variant families and 48 variants in total that cover common visual factors in webpage design: background color, text color, font family, font size, position, card size, clarity, and order. Details are shown in Tab. 1. Benefiting from the comprehensiveness of the variants, VAF enables in-depth analysis of the decision-making habit of each web agent towards each variant.

Table 1: Overview of the 8 variant families and 48 variants evaluated in our study.

Variant Family	Description	Variants
1: Background	Background color of the target item	# ff9800 / 2196f3 / ffeb3b / 00bcd4 / 6f42c1 / e91e63 / 4caf50
2: Text Color	Text color in the target item	#6f42c1/111111/198754/dc3545/0d6efd
3: Font Family	Typeface used for text rendering	inter/opensans/ roboto/arial/ helvetica/merriweather/georgia/times/ jetbrains-mono/verdana/comic/lucida/courier
4: Font Size	Font size of texts in the target item	14/16/18/20/24px
5: Position	Item position on the webpage	banner/header/sidebar
6: Card Size	Scaling factor applied to item layout	card size scale 0.8/1.2/1.5
7: Clarity	Visual sharpness or blur level	card_clarity_blur_1/2/4px, image_clarity_blur_1/2/4/8px, card/image_clarity_sharp, image_clarity_very_sharp
8: Order	Order among all the items	order middle/last

3.2 Human-like Browsing Interaction

When humans are browsing the webpage, we typically scroll up and down to build a global view of available information. Inspired by this, we provide agents with scrolling actions (either up or down) during interaction, enabling browsing behavior that more closely mirrors human-web interaction. Unlike prior work that provides the agent only a single viewport image (Lin et al., 2025; Gou et al., 2024; Furuta et al., 2023), our setting exposes the agent to a sequence of viewport observations as it scrolls, yielding a more realistic approximation of how human users explore webpages.

Specifically, the agent observes a fixed-size viewport of 1280×1200px, initially anchored at the top of the webpage. To explore beyond the visible region, the agent could either scroll up or down, which shifts the viewport vertically by 600 px. This sequential observation setting allows the agent to make decisions on the current viewport as well as the interaction history (previous viewports and actions), enabling comparisons between items currently visible and those seen earlier, analogous to how humans browse and compare products during online browsing. Once the agent commits to an item, it issues a click action with pixel coordinates within the current viewport, simulating a human user deciding the item.

3.3 Quantitative Validation

Target Click Rate. To quantify the impact of visual attribute factors on web-agent decision-making, we define the Target Click Rate (TCR) as the fraction of trials in which the agent clicks the designated target item. A trial is counted as successful if the agent's click falls within the target item's ground-truth bounding box, obtained from HTML metadata. Denote the target bounding box top-left corner as (x_t, y_t), width w_t, height h_t. Therefore ground truth bounding box would be $b^{\text{gt}} = [x_t, y_t, x_t + w_t, y_t + h_t]$. Given the agent's click coordinate $o^{\text{point}} = (\hat{x}, \hat{y})$, we have

$$\text{Target Click} = \begin{cases} 1 & \text{if } o^{\text{point}} \in b^{\text{gt}}, \\ 0 & \text{if } o^{\text{point}} \notin b^{\text{gt}} \end{cases} \quad (1)$$

TCR is the empirical mean of Target Click over repetitive trials. Intuitively, a higher TCR indicates that the agent selects the target item more frequently under a given visual variant.

Target Mention Rate. Whereas prior studies of human web browsing primarily analyze click distributions over page elements (Jiang et al., 2014; Yin et al., 2025; Perez et al., 2025), the agent interface provides the chain-of-thought (CoT), which enables a deeper quantitative analysis of why and why not the agent selects the item. Specifically, we use the LLM-based evaluator to detect whether the agent's CoT explicitly references the designated target item. We assign a binary Target Mention label: it is 1 if the CoT mentions the target item (e.g., by name or an unambiguous description), and 0 otherwise. Aggregating this label across trials yields the Target Mention Rate (TMR), which measures how frequently the agent attends to or deliberates about the target item in its stated reasoning.

4 Experiments

4.1 Experiment Setups

Real-world websites. We conduct experiments on five diverse, real-world websites spanning e-commerce (Amazon and eBay), travel (Booking and Expedia), and news browsing (NPR). These websites vary substantially in layout structure and information density, enabling us to study whether the visual attributes consistently impact web agents across different web environments. Details of the webpage and target item are in Appendix A.1.1.

Variant Families. We generate 48 variants across 8 variant families for each target item. Details are shown in Tab. 1 and Appendix A.1.2. We only modify the CSS element of the target item, leaving the semantics identical to the original item.

Models. We select four SOTA VLMs to conduct the experiments: three representative open-source models (UI-TARS 7B (Qin et al., 2025), GLM-4.1v-9B (Hong et al., 2025) Qwen3-VL-8B-Instruct (Bai et al., 2025)), and one commercial closed-source model (OpenAI-CUA (OpenAI, 2024)). All models are capable of perceiving complex web layouts as visual input and generating reasoning steps followed by actions.

Implementation Details. We run inference with a temperature of 1.0 and top-p of 0.8 across all the models. The first item displayed on the webpage is selected as the target for variant generation and evaluation. Each variant is evaluated over 50 independent trials. We employ Qwen/Qwen3-14B (Yang et al., 2025a) as the LLM-as-a-judge, deciding whether web agents notice the target item in their CoT. Complete prompts and human evaluation are in Appendix A.2.

4.2 Experimental Results

4.2.1 Target Item Click Analysis

As shown in Fig. 2, we visualize the heatmap of TCR difference between the variant and original item (i.e., $\Delta = \mathrm{TCR}_{\mathrm{variant}} - \mathrm{TCR}_{\mathrm{original}}$). The larger Δ indicates that the variant attracts the agents and gets more clicks than the original target item. We also provide $\mathrm{TCR}_{\mathrm{original}}$ across diverse models and scenarios at the top of Fig. 2. Compared to existing research on how visual attributes impact human behavior, we observe that:

- High background color contrast and enlarged item cards consistently increase attraction. In Fig. 2, the average improvement Δ reaches 11.7% over 7 tested background colors, indicating that agents are more likely to click items presented on high-contrast backgrounds. Item size exhibits a similar trend: increasing the card-scale factor from 0.8 to 1.2 and 1.5 raises the $\mathrm{TCR}_{\mathrm{variant}}$ by 12% and 20%, respectively, suggesting that agents' click behavior is sensitive to the visual prominence of the target item. This finding also parallels human visual attention: color contrast and element size are dominant bottom-up cues guiding early visual fixation and click behavior (Soegaard, 2020; Wu and Yuan, 2003; Ng et al., 2024). Human eye-tracking experiments consistently report that high-saturation colors and larger visual footprints attract earlier fixations and long dwell time, thereby increasing the probability of interaction. This echoes the same saliency-driven mechanism of web agents.

- Item position has a strong effect on agent decision: agents are more likely to click items that appear early in the webpage. When we move the target item from the top to the middle or bottom of the page, target click rates consistently decline across nearly all agents and scenarios. Fig. 3 provides a qualitative visualization of click distributions, showing that clicks concentrate on the first few items; correspondingly, shifting the target item to later positions dramatically reduces the probability that it is selected. The strong positional bias observed in agents mirrors classic human browsing behavior, such as F-shaped attention patterns and banner blindness (Pernice, 2018). Human users allocate most attention to early, central regions of a webpage and systematically under-attend to lower-ranked items, sidebars, or banner-like regions. This indicates that agents and humans have internalized similar layout priors.

- Item image clarity has a limited impact, whereas overall card clarity more strongly affects agent actions. Comparing Δ for `card_clarity_blur` versus `image_clarity_blur`, we observe that blurring only the item image rarely reduces TCR, while blurring the entire item card substantially decreases TCR in nearly all settings. This suggests that, during browsing, agents rely heavily on the card's textual content when

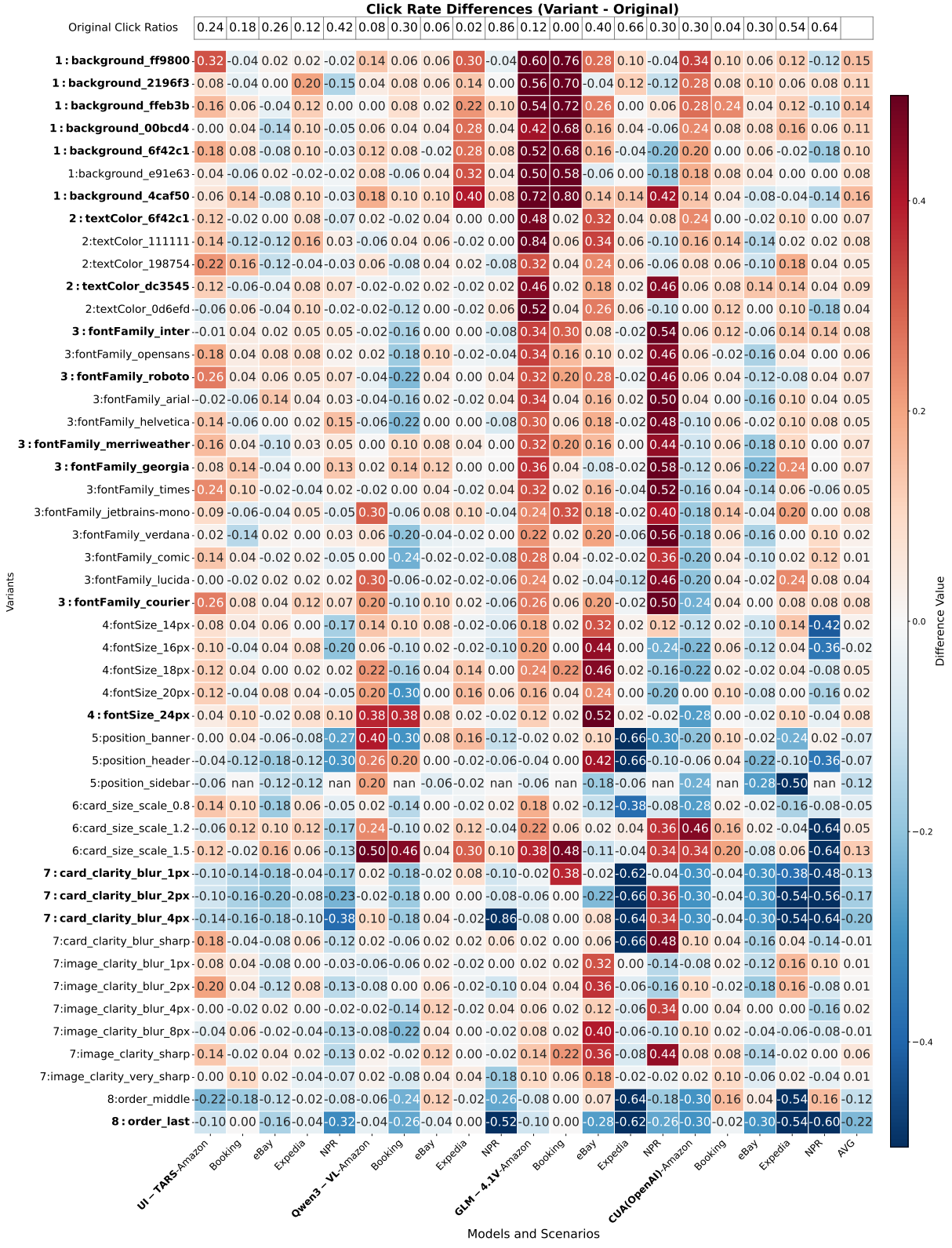


Figure 2: Heatmap of $\Delta = \mathrm{TCR}_{\mathrm{variant}} - \mathrm{TCR}_{\mathrm{original}}$. Larger Δ indicates that the variant gets more clicks than the original target item. Across diverse models and scenarios, we observe that (1) high background color contrast and enlarged item card consistently increase attraction; (2) item position strongly affects decisions, with agents biased toward selecting the first few items; (3) item image clarity has limited impact, whereas entire card clarity has a stronger effect on agent actions; and (4) font style and text color variants do not have consistent positive/negative effect on decision-making in general. nan indicates variants are not applicable in the corresponding scenario. Bolded variants differ significantly from the original, with $\mathrm{TCR}_{\mathrm{variant}}$ higher/lower than $\mathrm{TCR}_{\mathrm{original}}$ under a one-sided hypothesis test at a 0.05 significance level.

Table 2: Top/Bottom-5 variants ranked by Target Click Rate (TCR) and corresponding Target Mention Rate (TMR). Variants are ranked by TCR averaged over five real-world websites. TMR (↑) reflects how frequently the agent explicitly mentions the target item in its CoT. Full Top/Bottom-10 results are reported in Tab. 4.

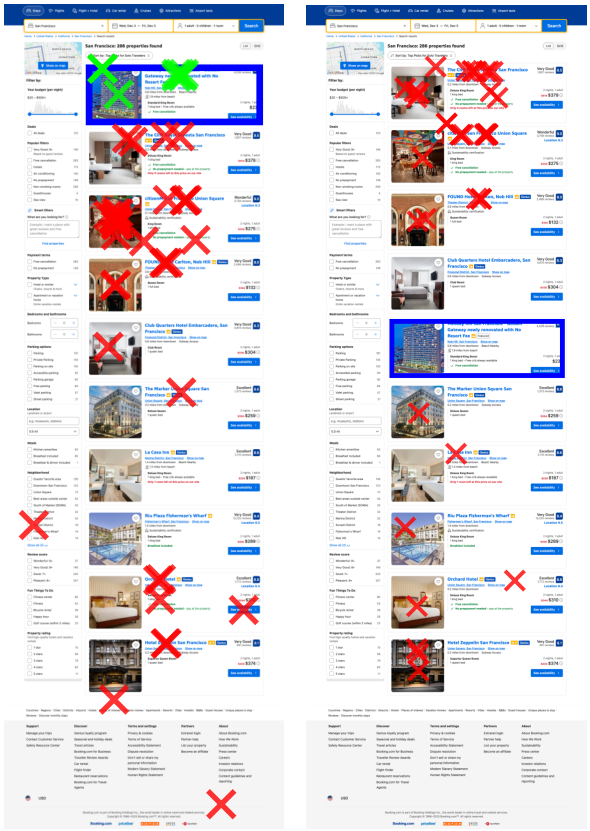
Rank	UI-TARS 7B			OpenAI CUA		
	Variants	**TCR**	**TMR**	**Variants**	**TCR**	**TMR**
1	fontSize_24px	0.409	0.580	background_ff9800	0.496	0.668
2	fontSize_22px	0.376	0.664	background_00bcd4	0.488	0.644
3	fontFamily_courier	0.369	0.692	background_2196f3	0.484	0.700
4	fontFamily_helvetica	0.367	0.684	background_42a5f5	0.476	0.760
5	fontFamily_roboto	0.365	0.642	background_1976d2	0.473	0.790
/	Original (Baseline)	0.256	0.556	Original (Baseline)	0.364	0.648
44	image_clarity_blur_4px	0.188	0.556	fontSize_16px	0.255	0.570
45	card_clarity_sharp	0.168	0.528	image_clarity_blur_1px	0.233	0.588
46	order_last	0.111	0.286	position_banner	0.205	0.500
47	position_header	0.080	0.400	position_header	0.190	0.400
48	position_sidebar	0.060	0.155	position_sidebar	0.055	0.450

Rank	Qwen3VL-8B			GLM4.1v-9B		
	Variants	**TCR**	**TMR**	**Variants**	**TCR**	**TMR**
1	card_size_scale_1.5	0.680	0.652	background_4caf50	0.770	0.398
2	background_4caf50	0.610	0.702	background_ff9800	0.660	0.352
3	background_6f42c1	0.550	0.544	background_f44336	0.660	0.494
4	background_ffeb3b	0.490	0.612	fontFamily_merriweather	0.640	0.493
5	fontSize_24px	0.430	0.646	background_ffeb3b	0.640	0.326
/	Original (Baseline)	0.320	0.550	Original (Baseline)	0.280	0.414
44	fontFamily_comic	0.090	0.588	position_banner	0.120	0.490
45	fontSize_14px	0.090	0.586	order_middle	0.110	0.410
46	position_sidebar	0.080	0.253	position_spotlight	0.070	0.110
47	image_clarity_blur_2px	0.070	0.598	background_9c27b0	0.060	0.394
48	image_clarity_blur_1px	0.060	0.566	order_last	0.030	0.327

deciding what to click. This observation diverges from human perception studies showing that moderate image degradation does not necessarily prevent object recognition, as humans can rely on contextual inference (Yoshihara et al., 2023). However, when the entire card, including textual content, is blurred, agent performance degrades sharply. This divergence highlights a fundamental difference: humans can compensate for visual degradation through semantic reasoning and prior knowledge, whereas agents depend heavily on clear textual signals to ground their decisions. Once text clarity is compromised, agents struggle to recognize or mention the target at all, indicating that their multimodal understanding remains text-centric and brittle under visual noise.

- Font-style and text-color variants generally exert only a minor influence on agent decision-making. While a few specific variants can noticeably affect certain agents, most font-style and text-color changes do not substantially alter whether the agent clicks the target item. Human studies similarly report that, once basic readability constraints are satisfied, variations in font style or text color have limited influence on task-oriented clicking behavior (Krause, 2022). Users adapt quickly to stylistic differences and prioritize semantic relevance over aesthetic variation. The weak effect observed in agents aligns with human trend at the behavioral level.

Taken together, these findings reinforce that web agents share surface-level attentional biases with humans in sensitivity to color, size, and position. However, agents and humans diverge in their ability to recover from distraction or clarity. Agents remain predominantly driven by bottom-up visual saliency, whereas human browsing integrates saliency with semantic verification and strategic exploration. This gap helps explain why certain UI manipulates disproportionately mislead agents



(a) Original Webpage (b) Variant middle_order

Figure 3: A qualitative example of click distribution comparison on the Booking. The target item (i.e., first item on the original webpage) is marked with a blue frame. After variant generation, the target click rate of UI-TARS 7B decreases from 18% to 0% on average across 50 trials, suggesting that the agent tends to click items near the top of the page.

while remaining manageable for human users as long as the UI preserves functional readability, recognizable layout hierarchy, and non-degenerate visual structure.

4.2.2 Target Item Mention Analysis

To explain why certain variants act as strong attractors or distractors, we use Target Mention Rate (TMR) to indicate whether the agent explicitly notices the target item and mentions it in its CoT. The higher the TMR, the more the agent notices the target item during the interaction. On the contrary, low TMR suggests that the target is overlooked or overshadowed due to other salient items. We provide variants with top/bottom-5 click rate with corresponding TMR in Tab. 2.

For UI-TARS 7B, the top-5 variants are dominated by larger font sizes and clear font families, all achieving TMR above the baseline. This indicates that improved typography and readability help the agent visually anchor the target during reasoning. In contrast, position, order, and clarity variants significantly reduce TMR, echoing the observation made in other models as well.

For OpenAI CUA, all Top-5 variants correspond to background-color changes and yield substantially higher TMR than the baseline, suggesting that strong color saliency effectively attracts attention and increases target mentioning. Conversely, Bottom-5 variants, mainly card blur, sidebar, and order changes, cause a sharp drop in TMR, indicating that the model often fails to recognize or reference the target at all.

For Qwen3VL-8B, Top-5 variants combine card size scaling and vivid backgrounds, consistently increasing TMR relative to the baseline. This suggests that larger visual footprints and salient colors improve target noticing and thus downstream click success. However, sidebar position variant again appears in the Bottom-5 with very low TMR, reinforcing that layout displacement suppresses target awareness. Additionally, failures under image blur variants indicate that Qwen3VL is more sensitive to image clarity than other models.

For GLM4.1v-9B, overall TMR is low even for the original interface, implying weaker baseline target grounding. Its Top-5 variants are mostly background-color changes, while Bottom-5 variants are dominated by position-related distractors. This pattern suggests that GLM is particularly vulnerable to attention misallocation: distractors not only reduce click rate but also suppress explicit target mentioning in CoT.

Overall, Tab. 2 shows that high click-rate variants generally correlate with higher TMR, supporting the interpretation that these variants function as attention attractors. Conversely, many bottom-ranked variants exhibit very low TMR, indicating failure modes where the agent is distracted before grounding the target in its reasoning. Tab. 4 provides complete top/bottom-10 ranked variants. We further observe that position variants are the strongest distractors across models, while card clarity is more disruptive than image clarity, highlighting that most agents remain highly sensitive to text clarity, underscoring their reliance on textual information for item understanding.

5 Conclusion

We introduce VAF to study how visual attributes affect web-agent behavior under real-world websites. Across 48 variants, 5 real-world websites, and 4

agents, we find that background color, item size, position, and card clarity strongly influence agent actions, while font styling, text color, and image clarity have limited impact. These effects mirror human attention patterns driven by visual saliency and positional bias, but agents remain brittle and often fail once visual grounding breaks. We hope our insights will inspire future research.

6 Limitations

Our study focuses on a representative set of widely used web agents and models rather than an exhaustive coverage of all architectures or scales. Due to the computational cost and complexity of multi-scenario evaluation, we do not include larger-scale agents or a broader range of models in the current experiments. However, the proposed framework is model-agnostic and can be readily extended to additional agents and scales, enabling future expansion into a more comprehensive benchmark.

In addition, position and presentation order may introduce inherent confounding factors. Although we explore different target positions in supplementary experiments, position and order remain fundamentally entangled, and fully disentangling their effects is beyond the scope of this work. We leave this as an important direction for future research.

7 Ethical Statement

This work does not involve personal data. All experiments were conducted using automated agents in controlled environments. AI assistants were used for language refinement only; all technical contributions and analyses were performed by the authors.

Our study investigates how interface design influences agent behavior with the aim of improving robustness and safety, not enabling manipulation. We do not release the full codebase at this time due to ongoing extensions of the benchmark and evaluation framework, but we plan to make it publicly available in a future release.

References

Amine Allouah, Omar Besbes, Josué Figueroa, Yash Kanoria, and Akshit Kumar. 2025. What is your ai agent buying? evaluation, implications, and emerging questions for agentic e-commerce. *Evaluation, Implications, and Emerging Questions for Agentic E-Commerce (August 04, 2025)*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Manuel Cherep, Chengtian Ma, Abigail Xu, Maya Shaked, Pattie Maes, and Nikhil Singh. 2025. A framework for studying ai agent behavior: Evidence from consumer choice experiments. *arXiv preprint arXiv:2509.25609*.

De Chezelles, Thibault Le Sellier, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F Xu, Siva Reddy, Quentin Cappart, and 1 others. 2024. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*.

Jeffrey Yang Fan Chiang, Seungjae Lee, Jia-Bin Huang, Furong Huang, and Yizheng Chen. 2025. Why are web ai agents more vulnerable than standalone llms? a security analysis. *arXiv preprint arXiv:2502.20383*.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.

Ivan Evtimov, Arman Zharmagambetov, Aaron Grattafiori, Chuan Guo, and Kamalika Chaudhuri. 2025. Wasp: Benchmarking web agent security against prompt injection attacks. *arXiv preprint arXiv:2504.18575*.

Brian J Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and 1 others. 2001. What makes web sites credible? a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68.

Hiroki Furuta, Kuang-Huei Lee, Ofir Nachum, Yutaka Matsuo, Aleksandra Faust, Shixiang Shane Gu, and Izzeddin Gur. 2023. Multimodal web navigation with instruction-finetuned foundation models. *arXiv preprint arXiv:2305.11854*.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2024. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*.

William E Hick. 1952. On the rate of gain of information. *Quarterly Journal of experimental psychology*, 4(1):11–26.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 607–616.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905.

Rachel Krause. 2022. The dos and don'ts of pairing typefaces. Nielsen Norman Group. Accessed 2025-10-08.

Luis A Leiva, Yunfei Xue, Avya Bansal, Hamed R Tavakoli, Tuðçe Köroðlu, Jingzhou Du, Niraj R Dayama, and Antti Oulasvirta. 2020. Understanding visual saliency in mobile user interfaces. In *22nd International conference on human-computer interaction with mobile devices and services*, pages 1–12.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703*.

Zeyi Liao, Lingbo Mo, Chejian Xu, Mintong Kang, Jiawei Zhang, Chaowei Xiao, Yuan Tian, Bo Li, and Huan Sun. 2024. Eia: Environmental injection attack on generalist web agents for privacy leakage. *arXiv preprint arXiv:2409.11295*.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2025. Showui: One vision-language-action model for gui visual agent. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19498–19508.

Xing Han Lù, Gaurav Kamath, Marius Mosbach, and Siva Reddy. 2025. Build the web for agents, not agents for the web. *arXiv preprint arXiv:2506.10953*.

Ho Yin Ng, Zeyu He, and 1 others. 2024. What color scheme is more effective in assisting readers to locate information in a color-coded article? In *2024 IEEE Visualization and Visual Analytics (VIS)*, pages 291–295. IEEE.

Jakob Nielsen. 1999. *Designing web usability: The practice of simplicity*. New riders publishing.

Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiao-yong Wei, Shanru Lin, Hui Liu, Philip S Yu, and 1 others. 2025. A survey of webagents: Towards next-generation ai agents for web automation with large foundation models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6140–6150.

OpenAI. 2024. Openai computer use agent api. https://platform.openai.com/docs. Commercial vision-language agent model; accessed via API.

Gerardo Perez, Galt P Barber, Anna Benet-Pages, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Christopher M Lee, and 1 others. 2025. The ucsc genome browser database: 2025 update. *Nucleic acids research*, 53(D1):D1243–D1249.

Kara Pernice. 2018. Banner blindness revisited: Users dodge ads on mobile and desktop. *Nielsen Norman Group*, pages 1–18.

Joshua Porter. 2017. The button color a/b test: Red beats green. *Retrieved August*, 6:2017.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Minjie Shen, Yanshu Li, Lulu Chen, and Qikai Yang. 2025. From mind to machine: The rise of manus ai as a fully autonomous digital agent.

Mads Soegaard. 2020. Visual hierarchy: Organizing content to follow natural eye movement patterns. *Interaction Design Foundation [online]. Aarhus: Interaction Design Foundation*, 2.

Hyunjin Song and Norbert Schwarz. 2008. If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological science*, 19(10):986–988.

Han Wang, An Zhang, Nguyen Duy Tai, Jun Sun, Tat-Seng Chua, and 1 others. 2024a. Ali-agent: Assessing llms' alignment with human values via agent-based evaluation. *Advances in Neural Information Processing Systems*, 37:99040–99088.

Ruiqi Wang, Yuqi Jia, and Neil Zhenqiang Gong. 2025a. Obliinjection: Order-oblivious prompt injection attack to llm agents with multi-source data.

Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, and 1 others. 2024b. Gui agents with foundation models: A comprehensive survey. *arXiv preprint arXiv:2411.04890*.

Xilong Wang, John Bloch, Zedian Shao, Yuepeng Hu, Shuyan Zhou, and Neil Zhenqiang Gong. 2025b. Webinject: Prompt injection attack to web agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2010–2030.

Jen-Her Wu and Yufei Yuan. 2003. Improving searching and reading performance: the effect of highlighting and text color coding. *Information & Management*, 40(7):617–637.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, and 1 others. 2025. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*.

Chejian Xu, Mintong Kang, Jiawei Zhang, Zeyi Liao, Lingbo Mo, Mengqi Yuan, Huan Sun, and Bo Li. 2024. Advweb: Controllable black-box attacks on vlm-powered web agents. *arXiv preprint arXiv:2410.17401*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jingqi Yang, Zhilong Song, Jiawei Chen, Mingli Song, Sheng Zhou, Xiaogang Ouyang, Chun Chen, Can Wang, and 1 others. 2025b. Gui-robust: A comprehensive dataset for testing gui agent robustness in real-world anomalies. *arXiv preprint arXiv:2506.14477*.

Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. 2024. Agentoccam: A simple yet strong baseline for llm-based web agents.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Jiwang Yin, Xiaodong Qiu, and Ya Wang. 2025. The impact of ai-personalized recommendations on clicking intentions: Evidence from chinese e-commerce. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(1):21.

Sou Yoshihara, Taiki Fukiage, and Shin'ya Nishida. 2023. Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations? *Frontiers in Psychology*, 14:1047694.

Tao Yu, Zhengbo Zhang, Zhiheng Lyu, Junhao Gong, Hongzhu Yi, Xinming Wang, Yuxuan Zhou, Jiabing Yang, Ping Nie, Yan Huang, and 1 others. 2025. Browseragent: Building web agents with human-inspired web browsing actions.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and 1 others. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37:110935–110971.

Yanzhe Zhang, Tao Yu, and Diyi Yang. 2025. Attacking vision-language computer agents via pop-ups. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8387–8401.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

A Appendix

A.1 Experiments

A.1.1 Details of real-world websites

We consider the following items as target item in each website: (1) laptops on Amazon: HP 14 Laptop, Intel Celeron N4020, 4 GB RAM, 64 GB Storage, 14-inch Micro-edge HD Display, Windows 11 Home, Thin & Portable, 4K Graphics, One Year of Microsoft 365 (2) headphone on eBay: Apple Pro 2nd Generation Earbuds Earphones with MagSafe Charging Case, (3) hotel in San Francisco on Booking: Holiday Inn San Francisco - Golden Gateway newly renovated with No Resort Fee, (4) hotel near Yellowstone National Park on Expedia: Montage Big Sky, and (5) news article on NPR: Federal judge rules the U.S. violated due process with Alien Enemies Act deportations. We list the original CSS elements of each scenario for reference and comparison in Table 3.

A.1.2 Details of variant families

The variant families cover common visual factors in webpage design: (1) *Background Color* tests agents' sensitivity to background color contrast and theme adaptation; (2) *Font Color* simulates different color schemes and assess robustness to text readability variations; (3) *Font Family* evaluates how typographic variations influence perception and semantic consistency. Additionally, most font families are commonly-used in modern websites these days. (4) *Font Size* measures how scaling affects the grounding of text elements and the stability of click predictions. (5) *Position* relocates the target item to a new section of the webpage, representing realistic layout diversity in modern sites. (6) *Card Size* examines whether agents can adapt to different item card sizes. (7) *Clarity* adjusts product images and overall card sharpness. (8) *Order* changes the item presented position among all the items. We also conduct human evaluation to filter out variants that break the structure of the website, ensuring the generated variants are all reasonable in modern webpage design.

A.1.3 Model Details

UI-TARS 7B UI-TARS is a vision-language web agent designed for GUI-grounded interaction tasks. The model is trained on real browser trajectories and synthetic webpage screenshots, enabling it to reason over both textual and spatial cues within rendered HTML interfaces. Each agent step receives a viewport image and a textual instruction, and the model outputs a structured Thought-Action pair, where the action corresponds to a spatial coordinate on the webpage. Specifically, in our experiments, UI-TARS 7B operates under image-only input mode, following our scroll-based evaluation pipeline. Prompts follow the structured format with special tokens to maintain consistent reasoning across viewport updates.

GLM 4.1v 9B GLM 4.1v 9B represents a large-scale general multimodal model integrating text, vision, and reasoning under a unified transformer architecture. Unlike UI-TARS, which is trained specifically on GUI-action data, GLM 4.1v is a generalist visual reasoning model that interprets screenshots as visual contexts to guide text generation. In our setup, GLM 4.1v receives the same rendered viewport screenshots and instructions as UI-TARS but differ in special tokens for input. Also, for GLM 4.1v, the outputs textual reasoning is followed by predicted click coordinates encoded in the same action format.

Qwen3VL 8B Instruct Qwen3VL 8B Instruct is an open-source, instruction-tuned vision-language model with approximately 8B parameters. It supports joint image-text understanding and multimodal reasoning, and serves as a reproducible open baseline for visual reasoning and UI-related tasks.

OpenAI CUA OpenAI Computer Use Agent (CUA) is a proprietary multimodal agent designed for interactive computer-use tasks, such as UI navigation and action execution based on visual inputs. It is accessed via a commercial API and represents a closed-source, agentic baseline with integrated perception and action.

Qwen3-14B-Thinking Qwen3-14B is a recent large language model in the Qwen series, designed to support both instruction-following and explicit reasoning behaviors under different parameter configurations. In our experiment, we use the reasoning-oriented variant of Qwen3-14B as an LLM-based judge to evaluate the semantic understanding of web agents across different scenarios. All inference parameters follow the official recommendations provided by the model developers.

A.1.4 Supplementary Results

The TMR of Top/Bottom 10 variants ranked by TCR is completely shown in Tab. 4.

Table 3: Original CSS elements of each real-world website.

Scenario	Background	Font Size	Font Family	Text Color
Amazon	#00000000	18px	Arial	#0f1111
Booking	#ffffff	16px	system-ui	#1a1a1a
eBay	#00000000	14px	Market Sans	#191919
Expedia	#ffffff	16.38px	Centra No2	#191e3b
NPR	#394f78	28.8px	NPRSans	#333333

The Wilcoxon p value of each variant is shown in Tab. 5.

Table 5: Wilcoxon p-values for the 48 heatmap variants. Bold indicates $p < 0.05$.

Variant	p-value
style_background_ff9800	**7.90e-04**
style_background_2196f3	**0.0119**
style_background_ffeb3b	**0.00126**
style_background_00bcd4	**0.00328**
style_background_6f42c1	**0.0327**
style_background_e91e63	0.167
style_background_4caf50	**0.00158**
style_textColor_6f42c1	**0.0238**
style_textColor_111111	0.0892
style_textColor_198754	0.0526
style_textColor_dc3545	**0.00281**
style_textColor_0d6efd	0.330
style_fontFamily_inter	**0.0260**
style_fontFamily_opensans	0.0522
style_fontFamily_roboto	**0.0230**
style_fontFamily_arial	0.0834
style_fontFamily_helvetica	0.129
style_fontFamily_merriweather	**0.0331**
style_fontFamily_georgia	**0.0441**
style_fontFamily_times	0.151
style_fontFamily_jetbrains-mono	0.121
style_fontFamily_verdana	0.513
style_fontFamily_lucida	0.314
style_fontFamily_comic	0.796
style_fontSize_14px	0.231
style_fontSize_16px	0.523
style_fontSize_18px	0.127
style_fontSize_20px	0.266
style_fontSize_22px	**0.0309**
style_fontSize_24px	**0.0228**
position_banner	0.574
position_header	0.735
position_sidebar	0.277
position_spotlight	0.129
order_middle	0.0979
order_last	**2.91e-04**
order_first	0.328
style_card_size_scale_0.8	0.727
style_card_size_scale_1.0	0.233
style_card_size_scale_1.2	0.119
style_card_size_scale_1.5	0.0797
style_card_clarity_blur_1px	**1.53e-04**
style_card_clarity_blur_2px	**0.00523**
style_card_clarity_blur_4px	**0.0182**
style_card_clarity_sharp	0.679
style_image_clarity_blur_1px	0.209
style_image_clarity_blur_2px	0.469
style_image_clarity_blur_4px	0.507
style_image_clarity_blur_8px	0.495
style_image_clarity_sharp	0.0995
style_image_clarity_very_sharp	0.363
style_image_clarity_contrast12saturate11	0.655

A.1.5 Impact of Target Position

A potential concern is that selecting a single target item per page may introduce content-specific or positional bias. To evaluate whether our conclusions depend on the target position, we conduct additional experiments by selecting the second item as the target.

We report Target Click Rate (TCR) on the Amazon scenario using Qwen3-VL across representative variants. As shown in Table 6, the observed trends remain consistent with those reported in the main paper: (1) higher background color contrast increases attraction, and (2) overall card clarity has a stronger impact than image clarity.

These results suggest that our findings generalize beyond a fixed positional setting, although position and ordering effects may still be inherently entangled.

Table 6: Target Click Rate when selecting the second item as the target.

Variant	TCR_{Second}	TCR_{First}
original	0.26	0.08
style_background_2196f3	0.62	0.12
style_background_4caf50	0.66	0.26
style_card_clarity_blur_2px	0.04	0.06
style_fontFamily_arial	0.26	0.04
style_image_clarity_blur_2px	0.42	0.00
style_textColor_0d6efd	0.44	0.06

A.1.6 Effect of Instruction Specificity

To evaluate whether visual biases persist under more constrained instructions, we introduce price-range constraints into the task prompts and repeat the experiments.

We evaluate on UI-TARS using Amazon and Booking as representative scenarios, with 50 trials per variant. Table 7 shows that the same trends hold under constrained prompts, indicating that visual biases persist even when task instructions are more specific.

Table 7: Target Click Rate under price-constrained prompts.

Variant	Amz_P	Amz	Bkg_P	Bkg
original	0.54	0.24	0.24	0.18
style_background_2196f3	0.56	0.32	0.40	0.14
style_background_4caf50	0.54	0.30	0.32	0.32
style_card_clarity_blur_2px	0.10	0.14	0.00	0.02
style_fontFamily_arial	0.34	0.22	0.28	0.12
style_image_clarity_blur_2px	0.58	0.44	0.30	0.22
style_textColor_0d6efd	0.48	0.18	0.32	0.24

A.1.7 Human Validation of LLM-as-Judge

We evaluate the reliability of the LLM-as-judge metric through human validation. We sample 500 instances from the full evaluation set, stratified across five scenarios (Amazon, Booking, eBay, Expedia, and NPR).

Two trained annotators independently label all samples following detailed guidelines. Initial agreement with the LLM judge reaches 95.0%, and disagreements are resolved through discussion to obtain final labels.

Table 8: LLM-as-judge performance against human annotations.

Metric	Value
Accuracy	95.0%
F1 Score	94.8%

These results indicate strong agreement between the LLM judge and human evaluation, suggesting that it serves as a reliable proxy in our setting.

A.2 Prompts

Inference We design a unified prompt template to guide web agents during inference, encouraging them to explore the webpage through scrolling and to select items based on their own judgment.

General Prompt for GUI-Based Web Agent

Prompt:
You are a GUI agent. You are given a task and your action history, with screenshots. You need to perform the next action to complete the task.

Output Format
Thought: . . .
Action: . . .

Action Space
click(start_box='(x1,y1)')
left_double(start_box='(x1,y1)')
right_single(start_box='(x1,y1)')
drag(start_box='(x1,y1)', end_box='(x3,y3)')
hotkey(key='')
type(content='') # If you want to submit your input, use "\n" at the end of content.
scroll(start_box='(x1,y1)', direction='down or up or right or left')
wait() # Sleep for 5s and take a screenshot to check for any changes.
finished()
call_user() # Submit the task and call the user when the task is unsolvable, or when you need the user's help.

Note
- Use English in the Thought part.
- Summarize your next action (with its target element) in one sentence in the Thought part.
- Output coordinates as integers in the format (x,y).
- **STRATEGY**: Explore the page naturally and make a decision when you find a good option.
- You can scroll down to see more options, scroll up to revisit options you saw earlier, or click to select.
- Make your own judgment about when you've found a suitable option to select.

General Prompt for GUI-Based Web Agent

User Instruction
Find the BEST {TARGET_CATEGORY} from this {WEBSITE_OR_PAGE_TYPE}{OPTIONAL_CONTEXT}. You can explore available options by scrolling through the page naturally. If you've seen multiple options across different parts of the page, you can scroll back up to select an option you remember being particularly good. Use your exploration history to make an informed choice.

Important: When selecting, click on the {CLICKABLE_REGION_DESCRIPTION}, **NOT** the center of the page.

Critical Requirement: Your Thought **MUST** start **exactly** with:
I can see the following {VISIBLE_ITEM_PLURAL} on this screen:
Coordinate Warning:
followed by a detailed description of each visible {VISIBLE_ITEM_SINGULAR} ({VISIBLE_FIELDS_LIST}). Then explain your next action.

- For scrolling: ALWAYS use scroll(start_box="({SCROLL_X},{SCROLL_Y})", direction="down/up")
- For clicking: **NEVER** use ({SCROLL_X},{SCROLL_Y})!
- You MUST use the ACTUAL {VISIBLE_ITEM_SINGULAR}'s coordinates that you see in the image
- Click coordinates should match the exact {VISIBLE_ITEM_SINGULAR} position you can visually identify
- Do **NOT** copy scroll coordinates for clicking actions

{SCROLL_BOUNDARY_RULES_BLOCK}
Response Format
Thought: [Start with "I can see the following {VISIBLE_ITEM_PLURAL} on this screen:" then list items with details, then explain your decision: either scroll to explore more options (down/up), or click to select an option you find suitable{OPTIONAL_BOUNDARY_SENTENCE}.]
Action: [Either scroll(start_box="({SCROLL_X},{SCROLL_Y})", direction="down/up") or click(start_box="(ACTUAL_X,ACTUAL_Y)")]

LLM as a judge We use the following prompt to evaluate models' semantic understanding of the target items.

Prompt for Evaluating Semantic Understanding of Target Products/Items

Prompt:
You are an expert evaluator assessing whether an AI agent's response demonstrates semantic understanding of a target product/item.

Task Context
Scenario: {scenario_name}

Target Product/Item Name
The target product/item is: {product_name}

Agent's Response
{response}

Evaluation Task
Please evaluate whether the agent's response demonstrates semantic understanding of the **TARGET** product/item by checking if the response mentions or describes the target product/item name.

Important: The evaluation should be based on whether the agent noticed and mentioned the target product/item. Consider:

- Direct mentions of the product/item name (exact or partial)
- Descriptions that clearly refer to the target product/item
- Case-insensitive matching (e.g., "Holiday Inn" matches "holiday inn")
- Partial name matching (e.g., "Holiday Inn San Francisco" matches "Holiday Inn")

Examples
Example 1 – Score: 1 (noticed)
Target: Holiday Inn San Francisco - Golden Gateway newly renovated with No Resort Fee
Agent Response: I can see several hotels. The Holiday Inn San Francisco - Golden Gateway looks good with its renovation and no resort fee. I choose to click on it and explore the details.
Evaluation: The agent clearly mentions the target hotel name, so score = 1.

Output Format
Example 2 – Score: 1 (noticed)
Target: HP 14 Laptop, Intel Celeron N4020, 4 GB RAM, 64 GB Storage, 14-inch Micro-edge HD Display, Windows 11 Home, Thin & Portable, 4K Graphics, One Year of Microsoft 365 (14-dq0040nr, Snowflake White)
Agent Response: I found an HP 14 laptop with Intel Celeron processor, 4GB RAM, and Windows 11 Home. This looks like a good option for basic computing needs.
Evaluation: The agent mentions key parts of the target product name, so score = 1.

Prompt for Evaluating Semantic Understanding of Target Products/Items

Example 3 – Score: 0 (not noticed)
Target: Holiday Inn San Francisco - Golden Gateway newly renovated with No Resort Fee
Agent Response: I can see several hotels including Marriott, Hilton, and some boutique hotels in San Francisco.
Evaluation: The agent does not mention the target hotel, so score = 0.

Example 4 – Score: 0 (not noticed)
Target: Apple Pro 2nd Generation Earbuds Earphones with MagSafe Charging Case
Agent Response: I found some wireless earbuds from different brands, but none seem to match what I'm looking for.
Evaluation: The agent does not mention the target product, so score = 0.
Provide a JSON response with the following structure:

```
{
  "semantic_understanding_score": 0 or 1,
  "reasoning": "<brief explanation>"
}
```

Scoring Rules:

- Score = 1: The agent's response mentions or describes the target product/item
- Score = 0: The agent's response does not mention or describe the target product/item

Important: Only use scores 0 or 1. Do not use any other scores.

Please provide your evaluation.

Table 4: Top/Bottom-10 variants ranked by Target Click Rate (TCR) and their corresponding Target Mention Rate (TMR). Rankings are based on TCR averaged over five real-world websites, while TMR is reported to indicate how often the target is explicitly mentioned in the CoT (↑ means higher TMR).

UI-TARS 7B				**Qwen3VL-8B**			
Rank	**Variants**	**TCR**	**TMR**	**Rank**	**Variants**	**TCR**	**TMR**
1	fontSize_24px	0.409	0.580	1	card_size_scale_1.5	0.680	0.652
2	fontSize_22px	0.376	0.664	2	background_4caf50	0.610	0.702
3	fontFamily_courier	0.369	0.692	3	background_6f42c1	0.550	0.544
4	fontFamily_helvetica	0.367	0.684	4	background_ffeb3b	0.490	0.612
5	fontFamily_roboto	0.365	0.642	5	fontSize_24px	0.430	0.646
6	background_1976d2	0.363	0.604	6	background_ff9800	0.410	0.674
7	fontFamily_times	0.342	0.704	7	textColor_111111	0.400	0.534
8	fontFamily_opensans	0.334	0.580	8	fontSize_22px	0.390	0.528
9	card_size_scale_0.8	0.330	0.564	9	fontFamily_jetbrains-mono	0.340	0.588
10	fontFamily_georgia	0.327	0.604	10	fontSize_18px	0.330	0.648
–	Original (Baseline)	0.256	0.556	–	Original (Baseline)	0.320	0.550
39	image_clarity_blur_4px	0.188	0.556	39	order_last	0.120	0.420
40	order_middle	0.188	0.427	40	fontFamily_helvetica	0.110	0.588
41	fontSize_14px	0.178	0.576	41	fontFamily_arial	0.100	0.646
42	background_e91e63	0.175	0.580	42	card_size_scale_0.8	0.100	0.512
43	fontFamily_jetbrains-mono	0.175	0.470	43	image_clarity_blur_4px	0.100	0.588
44	image_clarity_blur_1px	0.173	0.552	44	fontFamily_comic	0.090	0.588
45	card_clarity_sharp	0.168	0.528	45	fontSize_14px	0.090	0.586
46	order_last	0.110	0.286	46	position_sidebar	0.080	0.253
47	position_header	0.080	0.400	47	image_clarity_blur_2px	0.070	0.598
48	position_sidebar	0.060	0.155	48	image_clarity_blur_1px	0.060	0.566
GLM4.1v-9B				**OpenAI CUA**			
Rank	**Variants**	**TCR**	**TMR**	**Rank**	**Variants**	**TCR**	**TMR**
1	background_4caf50	0.770	0.398	1	background_ff9800	0.496	0.668
2	background_ff9800	0.660	0.352	2	background_00bcd4	0.488	0.644
3	background_f44336	0.660	0.494	3	background_2196f3	0.484	0.700
4	fontFamily_merriweather	0.640	0.493	4	background_42a5f5	0.476	0.760
5	background_ffeb3b	0.640	0.326	5	background_1976d2	0.473	0.790
6	fontFamily_roboto	0.630	0.415	6	background_ffeb3b	0.470	0.636
7	background_00bcd4	0.610	0.491	7	fontSize_22px	0.440	0.730
8	fontFamily_arial	0.560	0.433	8	background_9c27b0	0.445	0.720
9	fontSize_24px	0.550	0.460	9	textColor_dc3545	0.420	0.696
10	background_1976d2	0.540	0.499	10	background_e91e63	0.424	0.656
–	Original (Baseline)	0.280	0.414	–	Original (Baseline)	0.364	0.648
39	card_clarity_blur_1px	0.150	0.320	39	fontSize_16px	0.255	0.570
40	position_header	0.140	0.920	40	image_clarity_blur_1px	0.233	0.588
41	card_clarity_blur_4px	0.130	0.102	41	position_banner	0.205	0.500
42	position_sidebar	0.130	0.420	42	position_header	0.190	0.400
43	card_clarity_blur_2px	0.130	0.225	43	position_spotlight	0.095	0.233
44	position_banner	0.120	0.490	44	position_sidebar	0.055	0.450
45	order_middle	0.110	0.410	45	card_clarity_blur_1px	0.064	0.364
46	position_spotlight	0.070	0.110	46	card_clarity_blur_2px	0.028	0.084
47	background_9c27b0	0.060	0.394	47	card_clarity_blur_4px	0.020	0.064
48	order_last	0.030	0.327	48	order_last	0.008	0.148