Beyond Screenshots: Evaluating VLMs' Understanding of UI Animations

Chen Liang, Xirui Jiang, Naihao Deng, Eytan Adar, Anhong Guo
University of Michigan
{clumich, xirui, dnaihao, eadar, anhong}@umich.edu

Abstract

AI agents operating on user interfaces must understand how interfaces communicate state and feedback to act reliably. As a core communicative modality, animations are increasingly used in modern interfaces, serving critical functional purposes beyond mere aesthetics. Thus, understanding UI animation is essential for comprehensive interface interpretation. However, recent studies of Vision Language Models (VLMs) for UI understanding have focused primarily on static screenshots, leaving it unclear how well these models handle dynamic UI animations. To address this gap, we created **AniMINT**, a novel dataset of 300 densely annotated UI animation videos. We systematically evaluate state-of-the-art VLMs on UI animation understanding, including their abilities to perceive the animation effects, identify animation purposes, and interpret animation meaning. Our results show that VLMs can reliably detect primitive motion. However, their high-level animation interpretation remains inconsistent, with substantial gaps relative to human performance. Finally, we use Motion, Context, and Perceptual Cues (MCPC) to probe factors affecting VLM performance, revealing key bottlenecks and directions for future improvement.

1 Introduction

Recent work on AI agents has increasingly focused on building systems that can autonomously perceive, reason about, and act within user interfaces (UI) to complete complex tasks on users' behalf (Li et al., 2023a; Deng et al., 2023b; Zheng et al., 2024a; Liu et al., 2024; Wang et al., 2024; Zhang et al., 2025). In real-world settings, such agents must also develop a rich understanding of user interfaces, including the ways in which interfaces convey system state, provide feedback, and signal available interaction affordances to users.

A central yet underexplored aspect of this understanding is UI animation. Animation plays a fundamental role in modern user interface design to convey feedback and information (Chang and Ungar, 1993; Thomas and Calder, 2001; Tversky et al., 2002; Heer and Robertson, 2007; Chevalier et al., 2016). These short yet critical animations serve more than aesthetic or experiential purposes; they are often essential for interpreting both the interface and the user's interaction with it. For example, the MacOS dock icon bounces for notifications, and the password input box shakes on wrong password. In many cases, such animations are the primary or only channel through which this information is communicated. Unlike icons or illustrations, animation is defined more by *movements* that are drawn than by *drawings* that move (Baecker and Small, 1990). Because animation's meaning is typically encoded in motion rather than accompanying graphics, a still image alone is often insufficient to capture its intended message. Thus, a comprehensive UI understanding must account for both static content and dynamic animations.

In this work, we evaluate the capabilities of state-of-the-art VLMs to understand UI animations. Recent VLMs have shown strong performance on a range of user interface understanding tasks and have been applied to increasingly complex UI-centered problems (Shaw et al., 2023; Wu et al., 2024; OpenAI). However, to the best of our knowledge, no prior works have systematically studied their capabilities to understand UI animations.

To this end, we constructed **AniMINT**, the **first** UI AniMation INTerpretation dataset. It contains 300 UI animation videos sourced from web, mobile, and desktop platforms. Animations were carefully annotated by 3 UI/UX practitioners and by 300 diverse users, providing a complementary view of UI animations from both experts and general users. We release our dataset and annotations[1].

To systematically evaluate VLMs' ability to un-

[1] https://github.com/publicationacc/AniMINT

Findings of the Association for Computational Linguistics: ACL 2026, pages 12907–12927

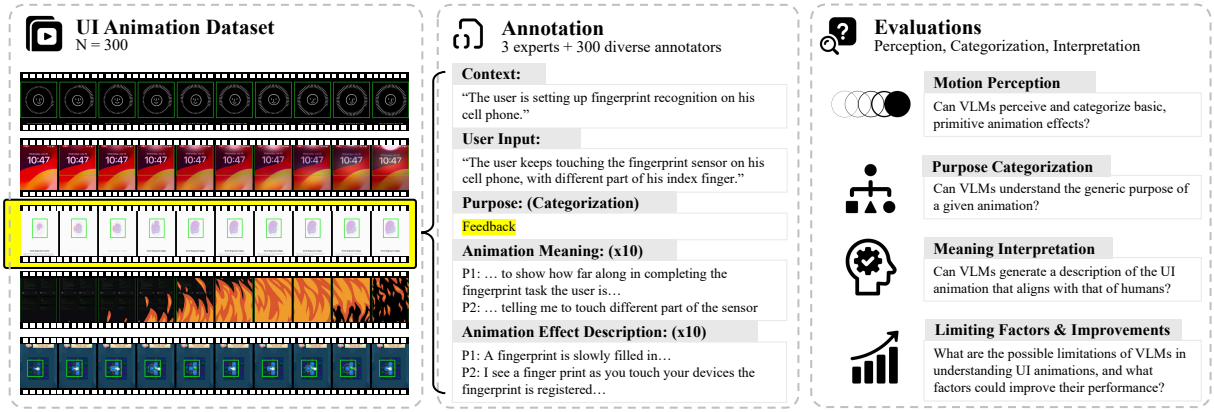


Figure 1: Overview of AniMINT, a UI animation dataset with multi-level human annotations. Each clip includes contextual information, an animation purpose label (highlighted in yellow), and ten annotations of the animation's meaning and effect, supporting evaluations of VLMs across perception, purpose categorization, and interpretation.

derstand UI animations, we formulate a set of research questions based on AniMINT that span both low-level animation recognition and higher-level animation understanding. We evaluated various state-of-the-art VLMs, including models from GPT, Gemini, and other model families.

Our evaluation shows that, although most VLMs reliably recognize primitive animation effects, they struggle with higher-level purpose categorization and meaning interpretation. To diagnose further, we investigate how Motion, Context, and Perceptual Cues (MCPC) affect UI animation understanding. We augment inputs with motion blending, interaction context, and perceptual captions, then re-evaluate categorization and interpretation. The findings reveal the bottleneck in motion perception, while also highlight the importance of grounding motion in interaction context and higher-level semantic meaning for accurate interpretation.

To summarize, our primary contributions are:

- We introduce **AniMINT**, the first dataset for UI animation understanding, with diverse annotations from both experts and everyday users.
- Using AniMINT, we conduct a systematic evaluation of nine state-of-the-art VLMs on both primitive animation perception and high-level animation categorization and understanding, revealing substantial gaps in current models' capabilities.
- We investigate factors that improve VLMs' capabilities on UI animation understanding and show their effectiveness on Gemini-2.5-Flash.

2 Related Work

UI Animation. Based on (Baecker and Small, 1990; Betrancourt and Tversky, 2000; Chevalier et al., 2016), we define UI animation as follows to guide data collection for AniMINT:

UI animation is a deliberately constructed, dynamic transformation of a user interface element that visualizes information or evokes a perceptual or cognitive response in the user. The transformation extends beyond the immediate next frame.

UI animations serve functional roles within the interface (Thomas and Calder, 2001; Chang and Ungar, 1993; Liddle, 2016), such as clarifying state transitions (Dessart et al., 2011; Schlienger et al., 2007), visualizing information (Tversky et al., 2002; Dessart et al., 2011; Schlienger et al., 2007), and enhancing user comprehension and experiences (Merz et al., 2016; Thomas and Calder, 2001). Drawing from prior literature (Baecker and Small, 1990; Chevalier et al., 2016; Novick et al., 2011; Avila-Munoz et al., 2021), we categorize animation purposes as Transition, Demonstration, Guidance, Feedback, Visualization, Highlight, and Aesthetic. We also categorize animation motion effects from the prior literature (Thomas and Calder, 2001; Novick et al., 2011) into 7 primitive effects, including move, rotate, size, color, fade, blur, and morph. These categorizations guide our data collection and annotation process, and detailed definition can be found in Appendix D.

UI Animation Datasets. There have been datasets that include UI interaction recordings for UI understanding and computer use agent training, such as Rico (Deka et al., 2017), MONDAY (Jang et al., 2025), GUI World (Chen et al., 2025), and others across different platforms and tasks (Rawles et al., 2023; Zhao et al., 2025; Man et al., 2025). To our knowledge, there is no dataset that specifi-

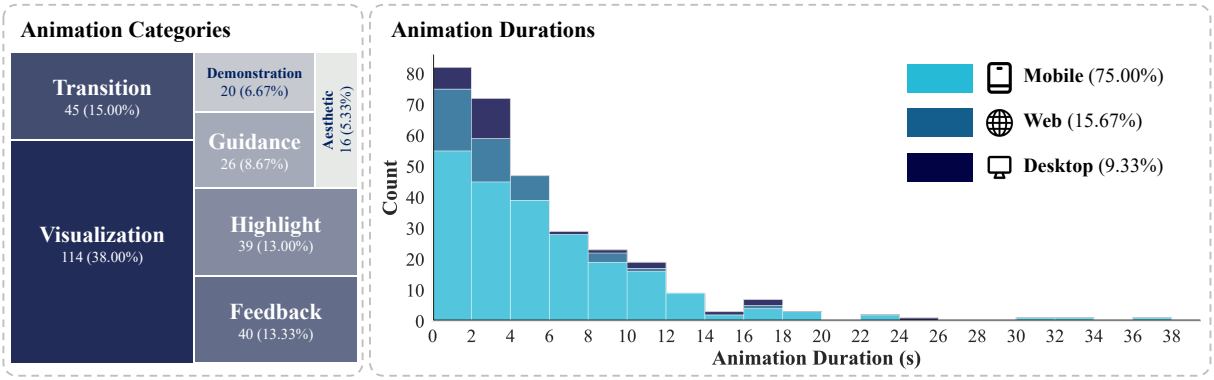


Figure 2: Dataset statistics. (Left) Distribution across seven animation purposes based on prior taxonomy. (Right) Animation duration by platform (mobile, web, and desktop). The median duration is 3.59s.

cally focuses on UI animation understanding. Existing datasets are typically designed for specialized tasks such as evaluating state transitions, UI adaptability, or visual signifiers for interaction discoverability (Mackamul et al., 2025; Dessart et al., 2011). Although recordings may include animation, they lack the diversity and annotation needed to evaluate animation understanding. In contrast, AniMINT sources diverse UI animation videos annotated by 3 domain experts and 300 general users.

VLMs and VLM Agent in UI Understanding. VLMs have emerged as powerful tools across various multimodal tasks, including visual scene comprehension, image captioning, and instructional task execution (Grattafiori et al., 2024; Bai et al., 2025). More recently, VLM-based agents have extended these capabilities to interactive settings, enabling models to perceive, reason about, and act within complex visual environments by iteratively grounding language instructions in visual observations (Xie et al., 2024; Wu et al., 2025). Despite this, their application to UI understanding, particularly regarding dynamic animations, is less explored. Existing studies primarily focus on static UI elements, such as visual component identification, interface semantics extraction, and static screen analysis, rather than the dynamic UI properties (Henderson, 2015; Trapp and Yasmin, 2013). In this work, we comprehensively evaluate a diverse set of VLMs on UI animation understanding.

3 AniMINT: Dataset for UI AniMation INTerpretation

Dataset and Annotations. We crafted a dataset of 300 animation videos collected across mobile, desktop, and web platforms. Mobile animations are mostly collected from the top 100 apps on the App Store and Google Play Store. Figure 2 visualizes the dataset distribution. The dataset is labeled by 3 domain experts and 300 diverse participants recruited on Prolific. First, each animation is labeled with metadata, including its temporal range, region of interest (ROI), and interaction context. Second, experts assign a purpose category to each animation based on majority voting. Third, we collect open-ended descriptions of each animation's meaning from general users, obtaining 10 independent responses per animation. In total, this results in 3,000 user-generated descriptions. Participants are compensated $3 for every 10 responses. The study is IRB approved. Detailed study setup and annotator demographics are provided in Appendix B.

Research Questions. Based on AniMINT, we formulate three research questions to evaluate VLMs' capabilities in understanding UI animations. Specifically, Can VLMs perceive and categorize primitive animation effects (**RQ1**, Section 4)? Can VLMs understand the UI animation purpose (**RQ2**, Section 5)? Can VLMs interpret UI animation meaning (**RQ3**, Section 6)? Guided by these questions, we further analyze how motion, context, and perceptual cues affect VLM performance to identify key factors for improvement.

Model Selections. As listed in Table 1, we evaluate 9 state-of-the-art models, including both commercial and open-sourced models. The detailed selection rationale is listed in Appendix A. We highlight that AniMINT serves as an evaluation similar to Zhou et al. (2023); Rein et al. (2024). Therefore, all experiments are conducted in a zero-shot setting, without any task-specific fine-tuning.

Video Preprocessing. We sample the 60 fps source videos at 10 fps, the minimum threshold

Model Name	Size	Context Length	License
GPT-5	Unk	400K	Closed
GPT-5-mini	Unk	400K	Closed
GPT-o4-mini	Unk	200K	Closed
GPT-o3	Unk	200K	Closed
Gemini-2.5-Pro	Unk	1M	Closed
Gemini-2.5-Flash	Unk	1M	Closed
Claude-Sonnet-4	Unk	1M	Closed
Qwen-2.5-VL-72B	73.4B	128K	Open
GLM-4.5V	106B	64K	Open

Table 1: Model details of VLMs tested in this work.

Motion	Start	End	Explanation
Blur			Changes sharpness or clarity
Color			Modifies hue/saturation/value
Fade			Adjusts transparency/opacity
Size			Scales along an axis
Rotate			Rotates orientation
Morph			Transforms shape/form
Move			Changes position

Table 2: Visual primitives and their effects. A stationary black square (left) serves as a spatial reference.

to avoid significant performance degradation for humans (Chen and Thropp, 2007). The sampling strategy is to enable fair comparison among models with and without native video support. More results for video input are in Appendix E and F. We use green bounding boxes as a visual prompt to highlight the ROIs for models. All frames are resized to a maximum dimension of 480 pixels. This protocol is applied across all models and tasks.

4 RQ1: Can VLMs Perceive Primitive Animation Effects?

Setup. We task VLMs with classifying primitive motion sequences into the most representative category among move, rotate, size change, color change, fade, blur, or morph, as shown in Table 2. Prompt and video details are listed in Appendix C.

Answer: Yes. Five out of nine models correctly classify all animation effects, including Claude Sonnet 4, GLM-4.5V, GPT-5, GPT-5-mini, and GPT-o4-mini. Figure 3 reports the corresponding accuracy scores. The results indicate that most models capture fundamental motion concepts with only minor errors, such as GPT-o3 misclassifying "fade" as "color change" in one case.

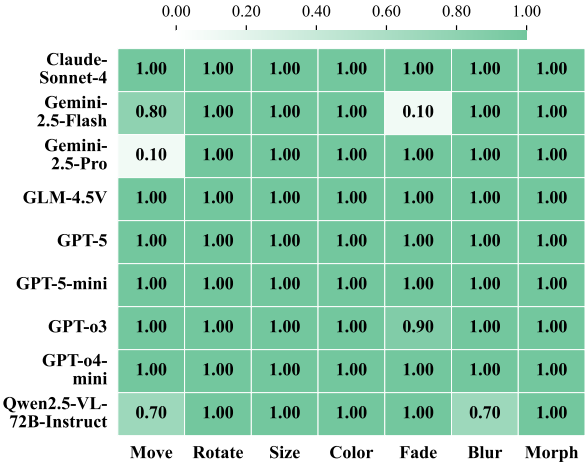


Figure 3: RQ1: VLM accuracy per animation effect. To mitigate position bias (Zheng et al., 2024b), we average 10 trials per prompt with randomized answer orders, keeping the randomization consistent across all models.

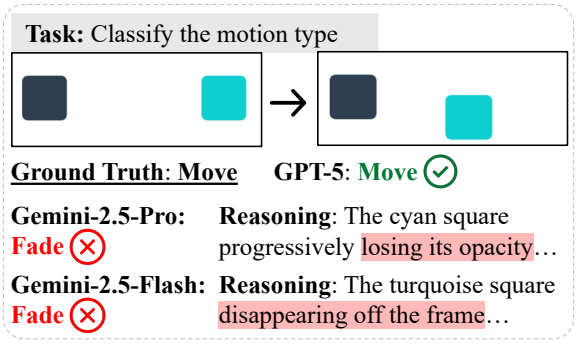


Figure 4: RQ1: An example where the "move" motion is incorrectly interpreted as "fade" by Gemini-2.5 Pro and Flash. These two models hallucinate the motion and reason the object "progressively losing capacity" or "disappearing off the frame."

4.1 Error Analysis

Hallucination errors. Despite correctly recognizing motion patterns, Gemini-2.5-Pro exhibits hallucination errors. In several cases, it describes non-existent visual elements, such as a "faint, translucent, rounded object" that does not appear in the animation. It also consistently misclassifies "move" as "fade" and hallucinates that a square is "progressively losing its opacity" (Figure 4). These behaviors suggest potential hallucination or misinterpretation of visual cues that aligns with prior observations (Li et al., 2023b; Gunjal et al., 2024).

Conceptual confusion. Gemini-2.5-Flash shows a pattern where it consistently labels fade as color change, whereas other models selected correctly. This suggests difficulty distinguishing subtle differences between closely related animation effects.

Model	Accuracy	Macro F1
Gemini-2.5-Pro	**0.64**	**0.55**
GPT-5	0.64	0.53
GPT-o4-mini	0.63	0.51
GPT-o3	0.62	0.54
Gemini-2.5-Flash	0.61	0.53
GPT-5-mini	0.58	0.48
Claude-Sonnet-4	0.57	0.46
GLM-4.5V	0.45	0.4
Qwen2.5-VL-72B-Instruct	0.39	0.32

Table 3: RQ2 Model performance comparison. Gemini-2.5-Pro achieves the highest accuracy (0.64) in identifying the generic purpose of UI animations, indicating significant room for improvement. Appendix E.3 provides the pair-wise statistical test.

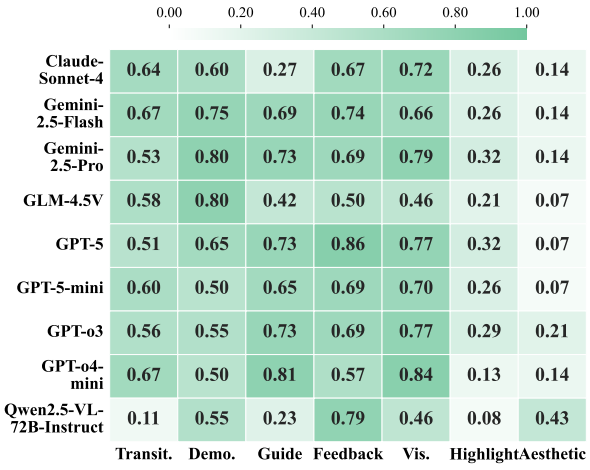


Figure 5: RQ2 per-category recall scores by model and animation purposes. While models perform better on animations with more direct functional purposes (such as Transition, Demonstration, Guidance, Feedback, and Visualization), they struggle with animations serving more subtle purposes, such as Highlight and Aesthetic.

5 RQ2: Can VLMs Understand the UI Animation Purpose?

Setup. We task VLMs to categorize the purpose of each animation into one of the seven classes: Transition, Demonstration, Guidance, Feedback, Visualization, Highlight, and Aesthetic. Appendix D lists the detailed definitions, examples, and the prompt. We report classification accuracy and macro-averaged F1 score in Table 3.

Answer: No. As shown in Table 3, the best-performing model, Gemini-2.5-Pro, can only reach an accuracy of 0.64. This shows that VLMs still have a significant gap in understanding the general purpose of UI animations.

Per-category performance. Figure 5 shows the per-category recall. Models capture direct functional purposes such as Transition (average recall of 0.54), Demonstration (0.63), Guidance (0.59), Feedback (0.69), and Visualization (0.69) relatively well, with all categories achieving recall above 0.50. However, performance drops on more subtle purposes, such as Highlight (0.24) and Aesthetic (0.16), which are harder for models to identify.

Per-category difficulty. We analyze the majority-vote results across the nine models in Figure 12. The models unanimously select the correct label for 56 animations (18.7%) and reach a correct majority consensus for 176 animations (58.7%). Consistent with the per-category recall results, direct functional categories, particularly Feedback (0.76), Visualization (0.73), and Guidance (0.69), show higher agreement and accuracy than more subtle categories Highlight and Aesthetic. Additional results and discussions can be found in Appendix E.

5.1 Error Analysis

VLMs focus on the static frame rather than the animation. As shown in Figure 6 (top), six out of nine VLMs incorrectly predict the animation category as Feedback rather than Aesthetic. The models seem to base their prediction on the final static frame. Specifically, the concluding frame displays the message "Your order is confirmed," which conveys feedback to the user. However, the animation features the McDonald's "M" logo bouncing into view, accompanied by the text "ba da ba," creating a playful and celebratory effect. These motion cues and visual elements serve an aesthetic and emotional purpose, reinforcing brand identity rather than communicating new or necessary information. This failure case highlights a limitation of existing VLMs that they tend to overemphasize salient textual cues in static frames rather than interpreting the animation holistically. As a result, visually rich but semantically lightweight animations can be overshadowed by static content, leading to incorrect interpretation.

Small ROIs pose challenges. As shown in Figure 6 (middle), VLMs often fail on animations where the countdown progress indicator occupies only a small ROI. Four models ignore the progress bar entirely and instead provide high-level descriptions of the surrounding webpage, and two other models identify an incorrect animation. Among samples that have a correct majority-voted an-

Figure 6: Examples in RQ2 where VLMs fail to identify the correct animation purpose.

swer, the mean ROI size is 24.3% of the screen, which is significantly larger (Mann–Whitney U test, $p = 0.03$) than that of abstained cases (i.e., instances where no consensus among models is reached), whose mean ROI size is 14.1%. These errors indicate that when animated elements are visually small, VLMs may fail to localize the relevant motion, leading to incorrect inferences about the animation's purpose.

VLMs overlook the context. The animation shown in Figure 6 (bottom) can, at first glance, be interpreted as a simple transition from the main screen to the app list on Android. However, the context reveals a different intent: the user repeatedly attempts to swipe up to open the app list but fails to complete the gesture correctly. In response, the system triggers the animation as a Demonstration to illustrate the correct interaction for the user. Despite this contextual signal, the VLMs fail to incorporate the context into animation understanding, resulting in eight out of nine models to misclassify the animation as a Transition. This failure case indicates that current VLMs are not able to well connect perceived UI animations with contextual information such as user intent and prior interaction attempts for interpretation. As a result, animations whose meaning depends on the interaction context are prone to misclassification.

6 RQ3: Can VLMs Interpret UI Animations?

Setup. We task VLMs to generate a natural language interpretation and compare its semantic similarity to the human responses. We use GPT-5-mini as the judge model (Zheng et al., 2023). To mitigate the potential bias in LLM-as-a-judge (Chen et al., 2024; Ye et al., 2025), we randomize response orders and prompted the judge model to evaluate independently of length. We report the mean and standard deviation of the similarity scores per model. For each animation, we leverage the 10 human responses collected and evaluate model predictions either against each individual response or against a summarized version of the responses (details are listed in Appendix F). Since both approaches yield similar model rankings empirically, we report results based on the summarized responses and defer the other results to Appendix G.

	Mean (↑)	Std (↓)	Distribution
GPT-o3	**3.47**	0.91	
GPT-5	3.44	0.90	
Gemini-2.5-Pro	3.40	0.90	
GPT-5-mini	3.39	**0.82**	
Gemini-2.5-Flash	3.31	0.95	
GPT-o4-mini	3.23	1.01	
Claude-Sonnet-4	3.10	1.12	
Qwen2.5-VL-72B	2.94	1.24	
GLM-4.5V	2.71	1.47	

Table 4: RQ3 Semantic similarity scores between VLM predictions vs. the summarized human response. We report the score distribution, where the five colors from left to right correspond to scores from 0 to 5. Appendix G.1 reports results based on individual human responses. Appendix G.2 provides the pair-wise statistical test.

Answer: VLMs capture gist, but miss key details. As shown in Table 4, GPT-o3 achieves the highest average score (3.47 ± 0.91), while GLM-4.5V yields the lowest (2.71 ± 1.47). Most of these VLMs achieve an average score of 3 and above, indicating that VLMs are capable of capturing the

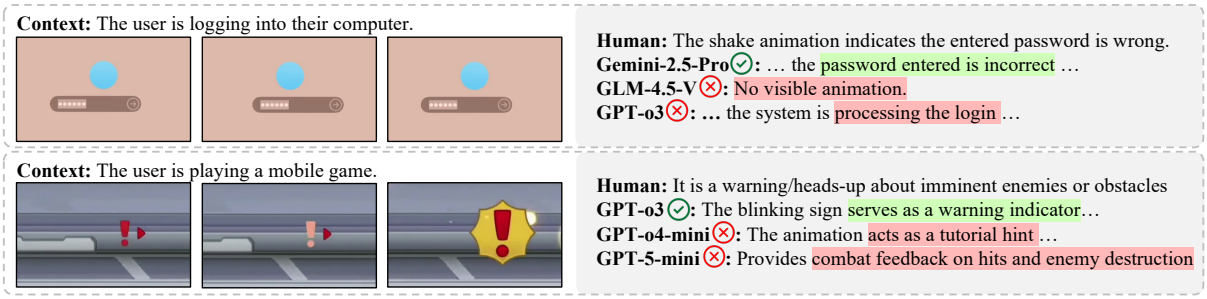


Figure 7: RQ3: Examples of Animation and the interpretations from VLMs.

gist of animation purposes according to the scoring rubric (Appendix F.2). However, these VLMs' responses often miss key details or contain subtle differences in nuance.

6.1 Error Analysis

Subtle, rapid animations pose challenges. For the example in Figure 7 (top), five out of nine models score 0, where they either do not perceive the animation at all (e.g., GLM-4.5v: "No visible animation"), or hallucinate (e.g., GPT-o3 described a "collapsing progress bar" which does not exist). This animation corresponds to a common UI pattern in which an input box briefly shakes to indicate an incorrect password. Although this shaking motion is highly recognizable to human users, it is subtle and quick. As a result, VLMs struggle to detect the motion signal, leading either to missed detections or spurious interpretations.

Small ROIs impact interpretation. Similar to RQ2 (Section 5), VLMs perform poorly on animations with a small ROI. As shown in Figure 7 (bottom), a small animated warning indicator, despite exhibiting a visually noticeable animation, receives an average score of 2. Models that fail on this example either do not perceive the animation at all or are distracted by larger static elements outside the highlighted ROI. For instance, GPT-5-mini does not mention the red exclamation mark in its reasoning and instead focused on the motion of the jetpack character. GPT-o4-mini correctly detects the exclamation mark but conflate it with surrounding visual elements (e.g., moving arrows and slider graphics), leading it to misinterpret the animation as a tutorial hint. These errors suggest that when animated elements are small, VLMs struggle to localize the relevant motion and may default to more visually salient but semantically irrelevant context. We provide more results on model performance for different categories in Appendix G.3.

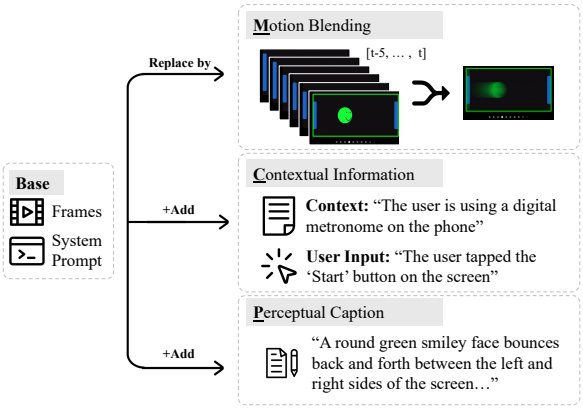


Figure 8: MCPC includes Motion Blending (blending the past six frames to capture motion), Contextual Information (interaction context and user input), and Perceptual Caption (textual descriptions of the animation).

7 Probing VLM Performance with Motion, Context, and Perceptual Cues

Setup. To identify limitations and potential improvement factors, we study how Motion, Context, and Perceptual Cues affect VLM performance. (Figure 8). For motion blending, we blend past frames into a single image with decaying transparency. This is inspired by Phosphor (Baudisch et al., 2006) that uses afterglow to show UI changes. For user context, we add contextual information such as the context of use and users' performed interactions. For perceptual caption, we provide the annotated text caption of what animations or motions are happening in the video. By combining these three factors, we re-evaluate VLM performance on purpose understanding (Section 5) and UI animation interpretation (Section 6). We test these combinations using Gemini-2.5-Flash, a lightweight model that demonstrates strong performance in earlier experiments. All other experimental setups are kept identical to Sections 5 and 6. More details about MCPC are listed in Appendix H.

	Purpose (RQ2)		Interp (RQ3)	
	Acc (↑)	$F1_{Macro}$ (↑)	Mean (↑)	Std (↓)
Base	0.59	0.47	3.15	1.09
+ M	0.52	0.41	3.08	1.07
+ C	0.58	0.48	3.3	0.95
+ P	0.57	0.45	3.5	0.89
+ M + P	0.53	0.4	3.48	0.86
+ M + C	0.59	0.49	3.34	0.98
+ C + P	0.55	0.46	3.48	0.77
+M+C+P	**0.61**	**0.52**	**3.52**†	**0.73**

Table 5: Effects of augmenting VLM inputs with MCPC on categorization (RQ2) and interpretation (RQ3). We adopt Gemini-2.5-Flash as the backbone model. The combined input (last row) outperforms all other combinations, demonstrating the joint effectiveness of MCPC. Appendix H.2 provides the details of the statistical test. †: significantly better than the base setup.

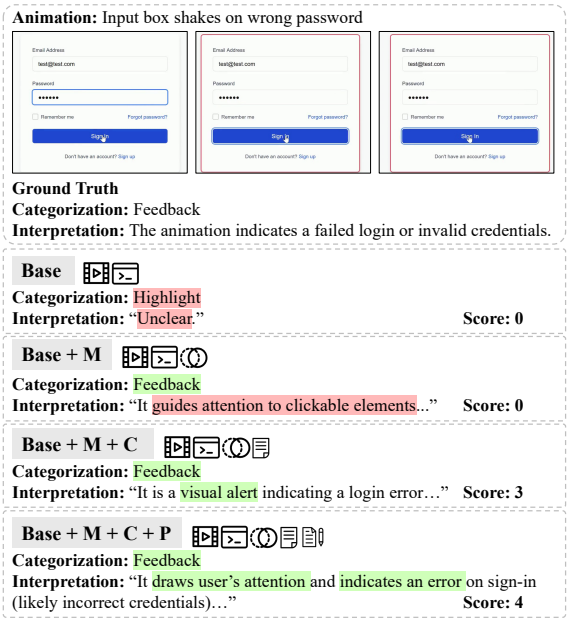


Figure 9: Improvements of incorporating MCPC: a wrong password shake is incorrectly classified as highlight in the base condition, but is correctly interpreted as an error indication with MCPC.

Results and Discussions. As shown in Table 5, combining motion, context, and perception cues leads to the best overall performance for both tasks. In Figure 9, the vanilla model fails both tasks, misclassifying the animation purpose as "Highlight" and describing the meaning as "unclear." With motion encoding alone, the model successfully classifies the purpose. Combining all three factors leads to the best performance, where the model successfully categorizes the shake and provides the most accurate interpretation compared to the other setups. This demonstrates the importance of motion, context, and perception, as well as the synergy effects across these factors in UI understanding.

8 Conclusion

In this paper, we investigate an often overlooked yet critical aspect of UI understanding – motion and animation. We construct AniMINT, a densely-annotated UI animation dataset sourced from real-world applications, and comprehensively evaluate a diverse set of state-of-the-art VLMs. We find that while most VLMs are capable of perceiving primitive motion effects, they struggle to categorize the animation purpose using the UI animation taxonomy. Also, although VLMs' interpretations often capture the gist, they frequently miss key details in their description. Furthermore, we investigate performance variations by encoding motion cues into images, adding contextual information, and supplying perception captions. They improve VLMs' performance on both the categorization and interpretation tasks, revealing the bottleneck of motion perception and the important synergy effects across perception and semantic context. We envision this work and our AniMINT dataset as a step toward interaction-aware LLM agents that operate between users and interfaces, using UI animation understanding to assist, explain, and guide user interactions involving complex animated behaviors.

9 Ethical Statement

This project was conducted in accordance with established ethical standards. All collected data were manually reviewed by the authors to ensure that no sensitive content (e.g., sexual material or violence) or potentially harmful visual stimuli (e.g., rapid flashing) were presented to annotators. Both the video data and the associated annotations were screened to prevent the inclusion of any personally identifiable information. All participants were recruited anonymously, provided informed consent, and were informed of their right to withdraw from the study at any time. The study protocol was approved by the IRB. Additional details regarding the annotation procedure are provided in Appendix B.

10 Limitations

Due to time and cost constraints, the collected animations are primarily sourced from U.S. based

applications where English is the primary language. Design practices and interaction patterns may vary across regions due to factors such as language reading direction (e.g., right-to-left vs. left-to-right) and cultural conventions (e.g., shaking to indicate confirmation) (Shen et al., 2024; Mogrovejo et al., 2024). We acknowledge that including data from a wider range of geographic and cultural contexts could introduce greater diversity into the dataset (Mihalcea et al., 2025). However, AniMINT is the first step in constructing a UI animation understanding dataset. We encourage future efforts in our community to diversify the animation sources and consider the cultural and language nuances.

Second, in the annotation process, all annotators were recruited within the United States and were English speakers, which may introduce interpretation bias in certain cases. Though this happens in many well-known NLP benchmarks (Deng et al., 2009; Bowman et al., 2015) [2], diversifying the annotation process can include more comprehensive opinions from a broader audience. Such an annotation can serve either as a training corpus that leads to a better customized model, or as an evaluation set to understand the limitations of the existing models. This is especially important as UI animation interpretation is a subjective task, therefore leading to diverse annotations (Plank, 2022; Deng et al., 2023a). For example, in stock or financial software, red often indicates an increase in some Asian countries but a decrease in the U.S. Incorporating greater cultural diversity among annotators could enrich the dataset and reveal additional insights into cross-cultural differences in how animations and visual cues are interpreted. When constructing AniMINT, we included ten annotations for each animation interpretation, hoping to cover as many cases as possible. We encourage future efforts in investigating the subjectivity in the task of UI animation understanding and extending the annotations beyond western countries.

Third, in this paper, we try our best to include a comprehensive set of VLMs in our experiments, including nine models from GPT, Gemini, and other model families. However, as the field is rapidly evolving, it is not feasible to exhaustively evaluate every available model variant. Another concern is whether to experiment with smaller models. We have conducted pre-liminary experiments in Appendix A and found that smaller models, due to their design constraints (e.g., limited context length, single-image input, etc), cannot handle the UI animation task well. Therefore, we focus primarily on the nine VLMs in Table 1. We encourage future efforts from our community to experiment with other VLMs on UI animation understanding.

[2]These early datasets typically do not report the annotator demographics. However, both datasets adopt the Amazon Mechanical Turk for annotation, which primarily consists of US workers in early stages (Ross et al., 2009; Ipeirotis, 2010; Irani, 2015)

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Raquel Avila-Munoz, Jorge Clemente-Mediavilla, Perez-Luque Perez-Luque, and Complutense University of Madrid, School of Communication. 2021. Communicative functions in human-computer interface design: A taxonomy of functional animation. *Rev. Commun. Res.*, 9:119–146.

Ronald Baecker and Ian Small. 1990. Animation at the Interface. In *The Art of Human-Computer Interface Design*. Addison-Wesley Pub. Co.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Patrick Baudisch, Desney Tan, Maxime Collomb, Dan Robbins, Ken Hinckley, Maneesh Agrawala, Shengdong Zhao, and Gonzalo Ramos. 2006. Phosphor: explaining transitions in the user interface using afterglow effects. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, UIST '06, page 169–178, New York, NY, USA. Association for Computing Machinery.

M. Betrancourt and B. Tversky. 2000. Effect of computer animation on users' performance: a review / (effet de l'animation sur les performances des utilisateurs: une sythèse). *Le Travail Humain*, 63(4):311. Last updated - 2013-05-03.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Bay-Wei Chang and David Ungar. 1993. Animation: from cartoons to the user interface. In *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology*, UIST '93, page 45–55, New York, NY, USA. Association for Computing Machinery.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, Tianshuo Zhou, Yue Yu, Chujie Gao, Qihui Zhang, Yi Gui, Zhen Li, Yao Wan, Pan Zhou, Jianfeng Gao, and Lichao Sun. 2025. Gui-world: A video benchmark and dataset for multimodal gui-oriented understanding. *Preprint*, arXiv:2406.10819.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Jessie YC Chen and Jennifer E Thropp. 2007. Review of low frame rate effects on human performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):1063–1076.

Fanny Chevalier, Nathalie Henry Riche, Catherine Plaisant, Amira Chalbi, and Christophe Hurter. 2016. Animations 25 years later: New roles and opportunities. AVI '16, page 280–287, New York, NY, USA. Association for Computing Machinery.

J.W. Davis. 2001. Hierarchical motion history images for recognizing human motion. In *Proceedings IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46.

Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, 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*, UIST '17, page 845–854, New York, NY, USA. Association for Computing Machinery.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023a. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023b. Mind2web: Towards a generalist agent for the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Charles-Eric Dessart, Vivian Genaro Motti, and Jean Vanderdonckt. 2011. Showing user interface adaptivity by animated transitions. In *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '11, page 95–104, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Jeffrey Heer and George Robertson. 2007. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247.

Rebecca Ussai Henderson. 2015. The principles of ux choreography. Accessed 2025-06-30.

Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk.

Lilly Irani. 2015. Difference and dependence among digital workers: The case of amazon mechanical turk. *South Atlantic Quarterly*, 114(1):225–234.

Yunseok Jang, Yeda Song, Sungryull Sohn, Lajanugen Logeswaran, Tiange Luo, Dong-Ki Kim, Kyunghoon Bae, and Honglak Lee. 2025. Scalable Video-to-Dataset Generation for Cross-Platform Mobile Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jie Lei, Tamara Berg, and Mohit Bansal. 2023. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, Toronto, Canada. Association for Computational Linguistics.

Chenliang Li, He Chen, Ming Yan, Weizhou Shen, Haiyang Xu, Zhikai Wu, Zhicheng Zhang, Wenmeng Zhou, Yingda Chen, Chen Cheng, Hongzhu Shi, Ji Zhang, Fei Huang, and Jingren Zhou. 2023a. ModelScope-agent: Building your customizable agent system with open-source large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 566–578, Singapore. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Daniel Liddle. 2016. Emerging guidelines for communicating with animation in mobile user interfaces. In *Proceedings of the 34th ACM International Conference on the Design of Communication*, SIGDOC '16, New York, NY, USA. Association for Computing Machinery.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2024. LLaVA-plus: Learning to use tools for creating multimodal agents.

Eva Mackamul, Fanny Chevalier, Géry Casiez, and Sylvain Malacria. 2025. Does adding visual signifiers in animated transitions improve interaction discoverability? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Brandon Man, Ghadi Nehme, Md Ferdous Alam, and Faez Ahmed. 2025. Videocad: A dataset and model for learning long-horizon 3d cad ui interactions from video. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Benedikt Merz, Alexandre N. Tuch, and Klaus Opwis. 2016. Perceived user experience of animated transitions in mobile user interfaces. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 3152–3158, New York, NY, USA. Association for Computing Machinery.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

David Novick, Joseph Rhodes, and Wervyn Wert. 2011. The communicative functions of animation in user interfaces. In *Proceedings of the 29th ACM International Conference on Design of Communication*, SIGDOC '11, page 1–8, New York, NY, USA. Association for Computing Machinery.

OpenAI. Introducing operator. Accessed 2025-07-01.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. 2023. Android in the wild: A large-scale dataset for android device control. *Preprint*, arXiv:2307.10088.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Joel Ross, Andrew Zaldivar, Lilly Irani, and Bill Tomlinson. 2009. Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep*, 49.

Céline Schlienger, Stéphane Conversy, Stéphane Chatty, Magali Anquetil, and Christophe Mertz. 2007. Improving users' comprehension of changes with animation and sound: An empirical assessment. In *Human-Computer Interaction – INTERACT 2007*, pages 207–220, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina Toutanova. 2023. From pixels to UI actions: Learning to follow instructions via graphical user interfaces. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Bruce H. Thomas and Paul Calder. 2001. Applying cartoon animation techniques to graphical user interfaces. *ACM Trans. Comput.-Hum. Interact.*, 8(3):198–222.

Marcus Trapp and René Yasmin. 2013. Addressing animated transitions already in mobile app storyboards. In *Design, User Experience, and Usability. Web, Mobile, and Product Design*, pages 723–732, Berlin, Heidelberg. Springer Berlin Heidelberg.

Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. 2002. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Jason Wu, Yi-Hao Peng, Xin Yue Amanda Li, Amanda Swearngin, Jeffrey P Bigham, and Jeffrey Nichols. 2024. Uiclip: A data-driven model for assessing user interface design. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. WebWalker: Benchmarking LLMs in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10290–10305, Vienna, Austria. Association for Computational Linguistics.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yanda Li, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2025. Appagent: Multimodal agents as smartphone users. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Dehai Zhao, Zhenchang Xing, Qinghua Lu, Xiwei Xu, and Liming Zhu. 2025. *SeeAction: Towards Reverse Engineering How-What-Where of HCI Actions from Screencasts for UI Automation*, page 463–475. IEEE Press.

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

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024b. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Model Selection Rationale

The task of UI animation understanding presents unique challenges compared to typical video understanding tasks. Unlike standard video understanding where sparse frames could be sufficient (Lei et al., 2023), perceiving UI animations requires dense frame extraction to capture fine-grained motion. This requires models to have comparatively larger context lengths to accommodate longer sequences of frames. Additionally, inferring the underlying purpose of an animation requires complex reasoning over interface elements, motion patterns, and context. Therefore, we prioritize larger models with advanced reasoning capabilities and selected state-of-the-art commercial and open-source multimodal large language models listed in Table 1.

In addition, we examined smaller VLMs with model sizes ranging from 7B to 14B and observed several limitations that hinder reliable evaluation. As a result, we exclude these smaller models from our primary analysis. In particular, models with restricted context lengths, such as Qwen2.5-VL 7B Instruct (32k tokens) (Bai et al., 2025), struggle to accommodate a sufficient volume of motion frames for animation interpretation. Other models, such as Llama 3.2 11B Vision Instruct (Grattafiori et al., 2024), are primarily for image understanding and cannot be fed with multiple images. As illustrated in Figure 10, preliminary tests with smaller models like Pixtral-12B (Agrawal et al., 2024) and Qwen-2.5-VL-7B (Bai et al., 2025) reveal the failure to perceive animation and its temporal changes, resulting in incorrect categorization and interpretation. These preliminary results suggest that existing smaller models are not yet capable of UI animation understanding. Therefore, we focus primarily on the VLMs listed in Table 1.

All model inference in this work was performed through OpenRouter `https://openrouter.ai/`. Model settings were left as default.

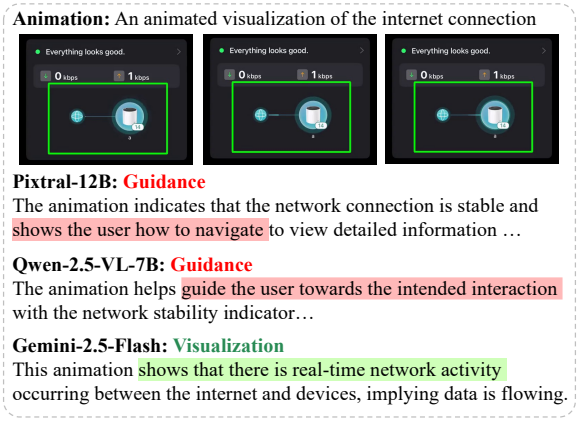


Figure 10: An example where small models failed the UI animation understanding task, while the advanced model (Gemini-2.5-Flash) succeeded. The generated interpretation from these small models suggests that these models are not yet capable of robust animation perception and interpretation.

B Annotation Details

B.1 Annotation Setup

We recruited 300 unique participants from Prolific, each of whom annotated a set of 10 videos through a short survey as illustrated in Figure 11, resulting in a total of 3000 responses. Participants were compensated $3 per 10 responses. The study is IRB approved, and all participants provided informed consent prior to participation.

The annotation task was hosted on Qualtrics in the form of a short survey, with each session consisting of 10 videos. To start, participants were given a tutorial of the labeling interface, task details, annotation best practices and requirements, and example high-quality annotations. Participants were specifically instructed to focus on the animation within the green bounding box to minimize distractions from other concurrent animations. Participants can modify their answers or revisit the tutorial materials at any time during the session.

B.2 Ethical Considerations

To protect participants from exposure to sensitive or potentially harmful content, all videos were manually reviewed by the research team prior to annotation. This verification process ensured that the dataset contained no sensitive material (e.g., sexual content, violence) or potentially harmful visual stimuli (e.g., rapid flashing). Participants were informed of their right to withdraw from the study at any time and were provided with contact information for the research team to address any concerns.

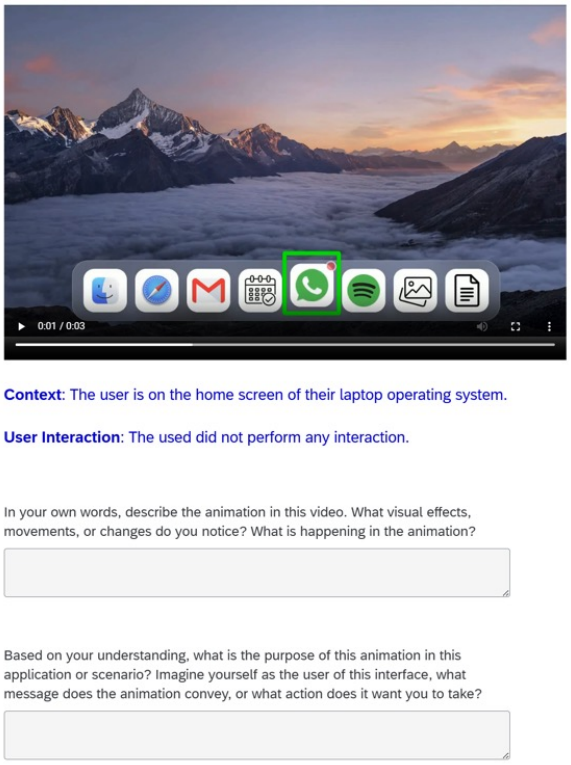


Figure 11: An example of the labeling interface, where the annotator can play the animation video with the green bounding box highlighting the animated region, see the context and user interaction details, and provide their interpretations.

B.3 Annotator Demographics

All participants were recruited from within the United States and reported English as their primary language. To increase annotation quality, we recruited participants who have finished at least 1000 tasks on Prolific before, and has an approval rate of 100%. All annotators were 18 years of age or older, with a mean age of 44.26 years (SD = 13.46). The gender distribution was 158 (52.8%) female, 140 (46.8%) male, and 1 (0.3%) participant who preferred not to disclose their gender.

B.4 Annotation Filtering

To preserve the authenticity of human interpretation, we applied minimal filtering, excluding only empty or inappropriate responses. As a result, the dataset retains brief annotations and explicit expressions of uncertainty (e.g., “I don’t know”). This decision ensures that the data captures the inherent ambiguity of the animations. For example, if a visual stimulus is confusing to human annotators, we expect a robust VLM to reflect similar uncertainty to achieve true alignment.

B.5 Animation Purpose Categorization

Animation purposes were annotated by three domain experts. All experts were provided with detailed definitions of each category, along with three example animations per category, to establish a shared understanding of the distinctions between classes. The original annotations exhibited an inter-annotator agreement of Krippendorff's $\alpha = 0.78$, indicating substantial and reliable agreement despite some disagreements. Final labels were determined by majority voting across annotators. In cases where no majority was reached (i.e., all three annotators assigned different labels), the instances were discussed in a follow-up adjudication session to reach consensus. All experts were compensated at $20/hr.

B.6 Intended Use of the Dataset

The dataset created in this work is intended for research use, such as evaluation and benchmarking. The dataset is constructed from publicly available content and does not include sensitive information or personally identifiable data.

B.7 Dataset Documentation

- Dataset size: 300
- Data Coverage: Video recordings of public mobile apps, operating systems, and websites.
- Video Language: English
- Annotation Language: English
- Annotator Region: United States
- Annotator Age: average 44.26 (20-80).
- Annotator Gender: 52.8% female, 46.8% male, 0.3% undisclosed.

B.8 Informed Consent

Below is the informed consent for data annotation:

You are invited to participate in a research study about evaluating machine learning model's understanding of animations used in user interfaces (UI), such as mobile apps, desktop software, and web interfaces. Specifically, the project investigates whether these models can perceive, interpret, and understand user interfaces the same way as humans do. To answer this question, researchers will evaluate human understanding of various UI examples, and then compare the results with machine learning model's responses to the same questions and see to what extent two sets of answers align with each other. If you agree to be part of the research study, you will be asked to watch recordings of UI animations and provide your interpretations of them. You will annotate 10 examples. You are not required to finish all examples, and can end the study anytime. We will primarily collect data through your responses in the questionnaire. We will protect the confidentiality of your research records by storing data on a secure server. We do not collect your identifiable information (e.g., your name, email). There is no direct personal benefit from being in this study. The risks and discomfort associated with participation in this study are minimal. Compensation: You will receive $3 for finishing annotating 10 samples. Participating in this study is completely voluntary. Even if you decide to participate now, you may change your mind and stop at any time. You may choose not to watch any UI recordings, interact with the labeling interface, answer any survey question, or continue with the study for any reason. If you have questions about this research study, please contact (anonymous) If you agree to participate, please proceed to the study below.

C RQ1 Evaluation Setup

We created a 3-second clip at 60 fps for each animation effect and used these clips in the RQ1 evaluation. Each prompt was repeated ten times, with answer choices randomly permuted to mitigate potential biases due to option ordering (Zheng et al., 2024b). For each run, the same randomized ordering was used across all models to ensure fair and consistent comparisons.

```
You are given a sequence of frames,
uniformly sampled at 10 frames per second
from a video of an animation.

Task:
Identify which single animation type best
matches the video you observe.

Options:

A. Move (object moves in any direction)
B. Rotate (object rotates along any axis)
C. Size (object changes sizes along any
   axis)
D. Color (object changes in hue, saturation,
   or brightness)
E. Fade (object change in
   transparency/opacity)
F. Blur (object change in sharpness or
   clarity)
G. Morph (object transformation from one
   shape/form to another)

Output format:
First line: the single letter (A to G) that
corresponds to the animation type. Second
line: an explanation of why this animation
type matches the video.
```

D RQ2 Evaluation Setup

D.1 Definition of Animation Purposes

Our animation categorization and definitions are derived from prior literature on UI animation taxonomy (Liddle, 2016; Betrancourt and Tversky, 2000; Chevalier et al., 2016), including:

Transition (Transit.): Animations that support layout changes.
Example: A flame animation in a privacy browser burning away tabs to transition to a new session.

Demonstration (Demo.): Animations that reveal or explain the behavior, functionality, or structure of the interface and its elements.
Example: An animation of the Face ID setup demo.

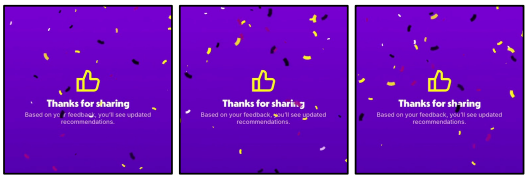
Guidance (Guide): Animations that guide the user towards an intended interaction.
Example: An animated arrow guiding the user to swipe up to capture the Pokemon.

Feedback: Animations that provide visual responses to user interactions.
Example: A ripple animation appears when two iPhones are near each other for proximity AirDrop.

Visualization (Vis.): Animations that represent system status, data, or other information.
Example: An animated bottle icon gradually filling up to visualize the loading process.

Highlight: Animations that emphasize specific content or draw the user's attention to key elements.
Example: A pulsing ripple animation highlighting the menu button in the corner.

Aesthetic: Animations that enhance the visual appeal, create an emotional impact, or improve user experiences.
Example: Animated confetti falling from the top.



D.2 Evaluation Prompt

```
You are a UI animation expert. You will
analyze an ordered sequence of frames
sampled uniformly at 10 fps from a
user-interface (UI) animation. Within each
video, a green box will appear when the
animation starts, and disappear when the
animation ends. Please primarily focus on
the animation happening within the green box
when you answer the questions. Please see
all the frames, and answer the following
questions about the UI animation in this
video.
You will be given the following information
as Inputs
- frames: a sequence of images captured at
10 fps. A green box will appear to identify
the region of animation.
- context: brief description of the
situation (e.g., app, user goal)
- input: description of any user interaction
right before or during the animation (tap,
swipe, talk, etc.), or no input was actively
performed.
Data for this video
context: {context}
input: {input}
Question: What is the primary purpose of
this UI animation? Describe your rationale
and explain how the animation effect
supports that purpose. Single-answer
question. Select only one option.
Options:

A. Transition: Animations that support
   layout changes.
```

Figure 12: Accuracy of the majority vote of predictions from nine models. The y-axis denotes ground-truth labels, and the x-axis denotes majority-vote predictions across models. Videos without a majority (fewer than five agreeing models) are labeled as "Abstain." Transition, Demonstration, Guidance, Feedback, and Visualization show the strongest performance and consistency, whereas Highlight and Aesthetic exhibit the weakest.

```
B. Demonstration: Animations that reveal or
   explain the behavior, functionality, or
   structure of the interface and its
   elements.
C. Guidance: Animations that guide the user
   towards an intended interaction
D. Feedback: Animations that provide visual
   responses to user interactions.
E. Visualization: Animations that represent
   system status, data, or other
   information.
F. Highlight: Animations that emphasize
   specific content or draw the user's
   attention to key elements.
G. Aesthetic: Animations that enhance the
   visual appeal, create an emotional
   impact, or improve user experiences.

For the selected category, write a sentence
describing your rationale and explain how
the animation effect supports that purpose

Output format:
Write exactly one line for the selected
category and its explanation/description.
For example: <Letter> – <PurposeName>: <Your
rationale>
```

E RQ2 Additional Results

E.1 Error Patterns and Category Confusions

Figure 13 shows the confusion matrix of individual models, and Figure 12 shows the confusion matrix of the majority-voted predictions. For majority-vote predictions across models, when there are fewer than 5 models agreeing on the same answer, the prediction will be labeled as "abstain". This is to reflect an overview of how VLMs in general performs on the categorization task. A closer inspection of misclassified cases reveals distinct error patterns. Animations in the Highlight category frequently fail to reach a majority consensus (i.e., Abstain; 13 out of 38 cases) or are misclassified as Guidance (11 out of 38). Similarly, Aesthetic animations are most often misclassified as Feedback (5 out of 14) or Highlight (5 out of 14). In addition, Figure 12 reveals several bidirectional confusion pairs, including Transition and Feedback (T as F: 13, F as T: 6), Feedback and Visualization (F as V: 3, V as F: 13), and Demonstration and Guidance (D as G: 4, G as D: 3). These confusions occur primarily between conceptually adjacent categories, suggesting that models struggle to capture fine-grained distinctions between closely related animation purposes, both visually and conceptually. Additionally, the systematically lower performance for Highlight and Aesthetic categories suggests that models are less effective at recognizing animations that serve subtle affective or cognitive purposes, indicating that the conceptual categories may be "memorized" than cognitively "perceived" by VLMs. This shows that a cognitive gap still exist between VLMs and human users in understanding these subtle interface cues.

E.2 Performance on Video Models

We conduct additional evaluations using Video LLaMa, LLaVA-Video, Qwen-2.5-VL, and Gemini-2.5-pro, where we use videos directly as input. The results are listed below:

Model	Acc	MacF1
gemini-2.5-pro	0.63	0.55
qwen2.5-vl-72b-instruct	0.42	0.38
videollama3-7b-local	0.21	0.22
llava-video-7b-local	0.19	0.20

Table 6: Model performance with video input

E.3 Statistical Significance Test

Test selection. We use McNemar's test (McNemar, 1947) to compare system variants, as our evaluation is paired and yields binary correctness outcomes for each video instance. Specifically, each system configuration produces a categorical prediction for the same set of videos, which we evaluate against the human-annotated ground truth to determine whether the prediction is correct or incorrect. McNemar's test assesses whether two classifiers

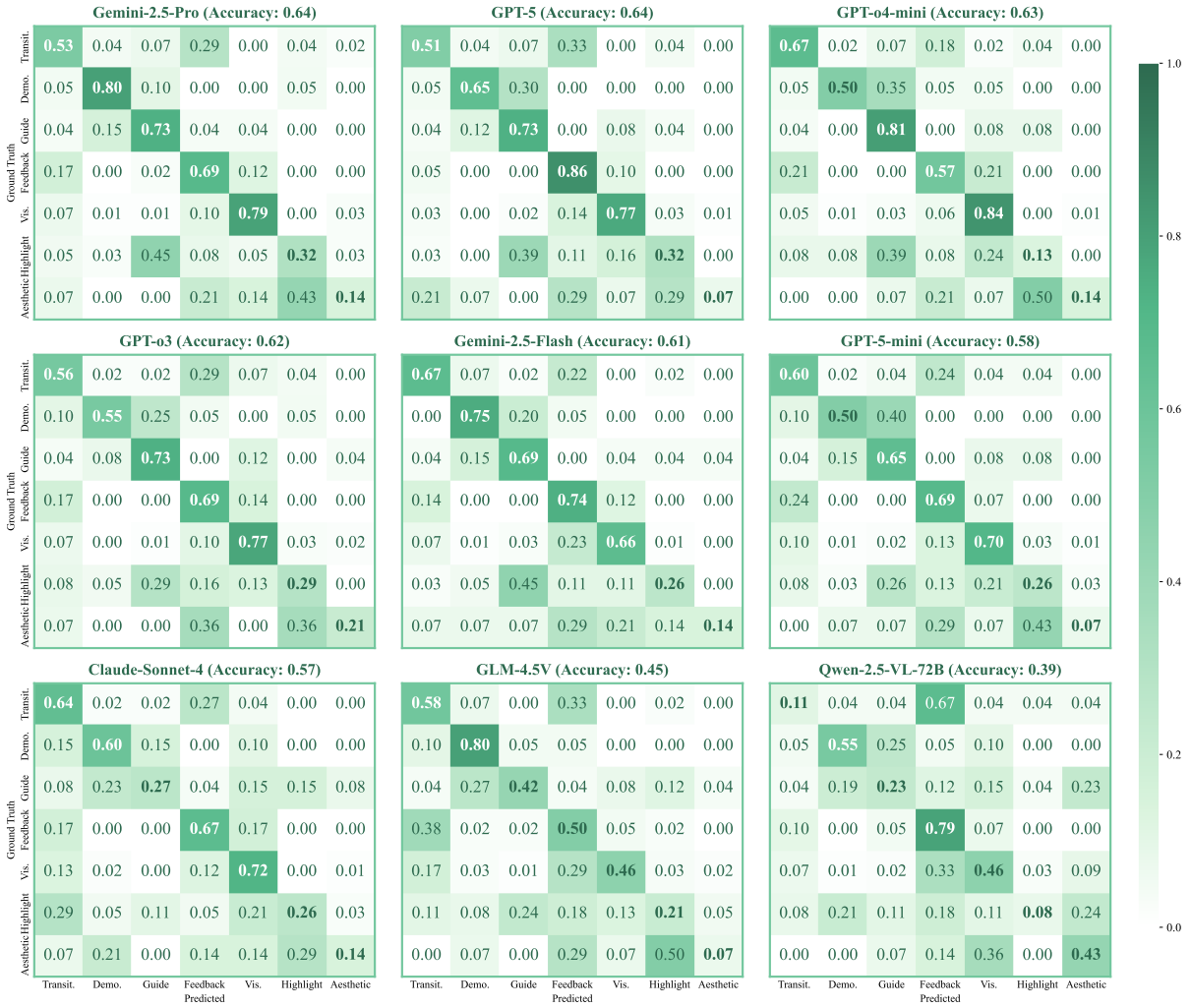


Figure 13: Confusion matrix for each model.

differ significantly when tested on the same examples, while accounting for the dependency between paired observations.

Test results. We conducted McNemar's test between each pair of VLMs in Table 3 to investigate whether their performance is statistically different. Table 7 shows the result. We highlight that more than half of the pairs yield a difference that is statistically significant with $p < 0.05$ or $p < 0.1$.

F RQ3 Evaluation Setup

F.1 Comparison Methods

As shown in Figure 14, we conducted two types of comparisons. The first approach (Figure 14 left) compares the VLM interpretation with each of the 10 human interpretations for a specific video, resulting in 10 similarity scores per video. The second approach (Figure 14 right) uses GPT-5 to summarize all 10 interpretations into a single response and compares the VLM interpretation with the summarized response, resulting in one score. These two methods offer perspectives at different levels of granularity. While the individual comparisons are susceptible to variations in annotation quality (e.g., short or unclear responses), they capture the distribution of direct similarities. Conversely, the summarized approach reflects alignment with the overall human understanding and focuses on high-level concepts, though it may lose some specific details found in individual responses. Empirically, we found that both approaches yield similar model rankings. Therefore, we report the results of the summarized responses in the main text and provide the individual comparison results in Appendix G.1 for additional context.

F.2 Evaluation Prompt

```
Please act as an impartial judge and compare
two short texts (Text A and Text B) that
describe the purpose/interpretation of the
```

Model 1	Model 2	p value	p < 0.05	p < 0.1
Gemini-2.5-Pro	Qwen2.5-VL-72B	4.25e-13	True	True
GPT-5	Qwen2.5-VL-72B	1.20e-12	True	True
GPT-o3	Qwen2.5-VL-72B	5.56e-11	True	True
GPT-o4-mini	Qwen2.5-VL-72B	7.47e-11	True	True
Gemini-2.5-Flash	Qwen2.5-VL-72B	5.68e-10	True	True
GPT-5-mini	Qwen2.5-VL-72B	2.87e-09	True	True
Gemini-2.5-Pro	GLM-4.5V	2.08e-08	True	True
GLM-4.5V	GPT-5	3.22e-08	True	True
Claude-sonnet-4	Qwen2.5-VL-72B	7.68e-08	True	True
GLM-4.5V	GPT-o4-mini	2.88e-07	True	True
GLM-4.5V	GPT-o3	3.72e-07	True	True
Gemini-2.5-Flash	GLM-4.5V	4.75e-07	True	True
GLM-4.5V	GPT-5-mini	1.30e-05	True	True
Claude-sonnet-4	GLM-4.5V	2.24e-04	True	True
Claude-sonnet-4	GPT-5	2.57e-02	True	True
Claude-sonnet-4	Gemini-2.5-Pro	2.63e-02	True	True
GPT-5	GPT-5-mini	3.31e-02	True	True
Claude-sonnet-4	GPT-o4-mini	3.56e-02	True	True
Gemini-2.5-Pro	GPT-5-mini	4.74e-02	True	True
GLM-4.5V	Qwen2.5-VL-72B	6.71e-02	False	True
Claude-sonnet-4	GPT-o3	8.05e-02	False	True
GPT-5-mini	GPT-o4-mini	9.80e-02	False	True
GPT-5-mini	GPT-o3	1.48e-01	False	False
Claude-sonnet-4	Gemini-2.5-Flash	2.42e-01	False	False
Gemini-2.5-Flash	GPT-5	2.60e-01	False	False
Gemini-2.5-Flash	Gemini-2.5-Pro	2.66e-01	False	False
Gemini-2.5-Flash	GPT-5-mini	4.64e-01	False	False
Gemini-2.5-Flash	GPT-o4-mini	4.89e-01	False	False
GPT-5	GPT-o3	5.33e-01	False	False
Gemini-2.5-Pro	GPT-o3	5.56e-01	False	False
Gemini-2.5-Flash	GPT-o3	6.25e-01	False	False
Claude-sonnet-4	GPT-5-mini	7.16e-01	False	False
Gemini-2.5-Pro	GPT-o4-mini	7.24e-01	False	False
GPT-5	GPT-o4-mini	7.98e-01	False	False
GPT-o3	GPT-o4-mini	8.90e-01	False	False
Gemini-2.5-Pro	GPT-5	1.00	False	False

Table 7: Pair-wise statistical significance test for results reported in Table 3. We highlight that more than half of the pairs yield a difference that is statistically significant with $p < 0.05$ or $p < 0.1$.

```
same UI animation. Decide their semantic
equivalence and coverage, considering:

  • Topics and actions, entities, and roles
  • Key attributes: numbers, units,
    dates/times, polarity/negation
  • Causal/temporal relations and
    constraints

Scoring (choose exactly one numeric score):

5: Paraphrase/equivalent meaning — Fully
   equivalent or one fully contains the
   other with no contradictions. No missing
   key facts.
4: Nearly equivalent; minor nuance
   differences — Main points identical,
   only subtle wording or emphasis
   differences.
3: Same gist; missing/extra key detail(s) —
   Core idea matches but some important
   details missing, added, or slightly
   inconsistent.
2: Some overlap; key differences — Partial
   overlap in main topic but significant
   differences in specifics or
   interpretation.
1: Same topic only — Related to same
   general subject but different focus,
   purpose, or approach.
0: Unrelated or contradictory — Completely
   unrelated topics or directly
   contradictory statements.
```

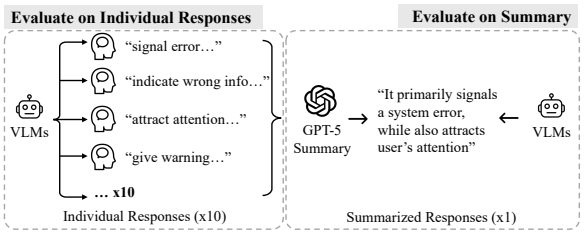


Figure 14: Illustration of semantic similarity computed against individual annotations ($N = 10$) versus a consolidated summary ($N = 1$).

```
Output Format: Return STRICT JSON (no code
fences) with schema:

{"score": 5 | 4 | 3 | 2 | 1 | 0, "reason": "..."}

Be concise and objective. Avoid any position
biases and ensure that the order in which
the responses were presented does not
influence your decision. Do not allow the
length of the responses to influence your
evaluation. Be as objective as possible.

Text A: {text_a}

Text B: {text_b}
```

G RQ3 Additional Results

G.1 Individually Compared Results

Table 9 shows the similarity score statistics if individually compared with human annotation. Comparing with individual (Table 9) or summarized (Table 4) response both yield similar model rankings.

G.2 Statistical Significance Test

Test selection. To compare scores of the model-generated UI animation interpretation, we use the Wilcoxon signed-rank test (Wilcoxon, 1945). In this setting, the scores for the same set of video instances yield paired but non-normally distributed observations. The Wilcoxon signed-rank test makes no assumptions about score normality, respects the paired structure of the evaluation, and tests whether the median difference between two systems' scores is zero. We therefore adopt the Wilcoxon signed-rank test to assess pairwise performance differences in Table 8.

Test results. We conducted the Wilcoxon signed-rank test between each pair of VLMs in Table 4. We highlight that most pairs yield a difference that is statistically significant with $p < 0.05$ or $p < 0.1$.

G.3 Discussions

Feedback presents the most challenges to VLMs. Category-wise, Feedback animations exhibited the

Model 1	Model 2	p value	p < 0.05	p < 0.1
GLM-4.5V	GPT-o3	1.56e-13	True	True
GLM-4.5V	GPT-5	3.49e-13	True	True
GLM-4.5V	GPT-5-mini	1.42e-11	True	True
Gemini-2.5-Pro	GLM-4.5V	8.33e-11	True	True
GPT-o3	Qwen2.5-VL-72B	2.15e-10	True	True
GPT-5-mini	Qwen2.5-VL-72B	2.75e-09	True	True
Gemini-2.5-Flash	GLM-4.5V	1.26e-08	True	True
GPT-5	Qwen2.5-VL-72B	3.98e-08	True	True
GLM-4.5V	GPT-o4-mini	1.00e-07	True	True
Gemini-2.5-Pro	Qwen2.5-VL-72B	3.30e-07	True	True
Claude-Sonnet-4	GPT-o3	1.30e-06	True	True
Claude-Sonnet-4	GPT-5	7.10e-06	True	True
Gemini-2.5-Flash	Qwen2.5-VL-72B	3.49e-05	True	True
Claude-Sonnet-4	GPT-5-mini	5.08e-05	True	True
GPT-o3	GPT-o4-mini	7.90e-05	True	True
Claude-Sonnet-4	GLM-4.5V	1.23e-04	True	True
Claude-Sonnet-4	Gemini-2.5-Pro	1.48e-04	True	True
GPT-o4-mini	Qwen2.5-VL-72B	2.84e-04	True	True
GPT-5	GPT-o4-mini	9.31e-04	True	True
Claude-Sonnet-4	Gemini-2.5-Flash	3.46e-03	True	True
GPT-5-mini	GPT-o4-mini	4.52e-03	True	True
Gemini-2.5-Flash	GPT-o3	1.50e-02	True	True
Gemini-2.5-Pro	GPT-o4-mini	1.78e-02	True	True
GLM-4.5V	Qwen2.5-VL-72B	4.38e-02	True	True
Claude-Sonnet-4	Qwen2.5-VL-72B	6.93e-02	False	True
GPT-5-mini	GPT-o3	8.59e-02	False	True
Gemini-2.5-Flash	GPT-5	8.85e-02	False	True
Claude-Sonnet-4	GPT-o4-mini	1.07e-01	False	False
Gemini-2.5-Flash	Gemini-2.5-Pro	1.71e-01	False	False
Gemini-2.5-Flash	GPT-5-mini	1.79e-01	False	False
Gemini-2.5-Pro	GPT-o3	1.87e-01	False	False
Gemini-2.5-Flash	GPT-o4-mini	2.49e-01	False	False
GPT-5	GPT-5-mini	3.97e-01	False	False
Gemini-2.5-Pro	GPT-5	4.68e-01	False	False
GPT-5	GPT-o3	6.23e-01	False	False
Gemini-2.5-Pro	GPT-5-mini	7.99e-01	False	False

Table 8: Pair-wise statistical significance test for results reported in Table 4. We highlight that most pairs yield a difference that is statistically significant with $p < 0.05$ or $p < 0.1$.

highest rate of unrelated responses, with 66.7% receiving one or more unrelated predictions. In contrast, the corresponding rates were substantially lower for other categories: 11.1% for Transition, 20.0% for Demonstration, 23.1% for Guidance, 22.6% for Visualization, 26.3% for Highlight, and 21.4% for Aesthetic.

We thus investigate whether these VLMs struggle with the Feedback category due to conceptual understanding or perceptual limitations. For example shown in Figure 7 (top), five models produced responses along the lines of "the system is giving feedback that it is verifying the password", which shows that although it can still correctly categorize the high-level animation purpose as Feedback, it fails to recognize the shake itself, or recognize the actual meaning of the shake. This highlights a limitation of VLMs in detecting rapid or small-scale movements such as shaking, which in turn prevents accurate interpretation of feedback animations.

	Mean (↑)	Std (↓)	Distribution
GPT-5-mini	2.76	1.10	
GPT-5	2.76	1.14	
Gemini-2.5-Flash	2.74	1.19	
GPT-o3	2.73	1.14	
Gemini-2.5-Pro	2.71	1.17	
GPT-o4-mini	2.65	1.14	
Claude-Sonnet-4	2.53	1.26	
Qwen2.5-VL-72B-Instruct	2.48	1.21	
GLM-4.5V	2.29	1.27	

Table 9: Statistics for semantic similarity scores. We calculate the score by comparing the model prediction with each individual human's response and report the average score. Similar to Table 4, we report the score distribution, where the five colors from left to right correspond to scores from 0 to 5.

This limitation also explains the generally weaker performance of VLMs on Feedback animations. Compared to other categories, feedback animations are often shorter in duration and involve less pronounced graphical change, making them especially reliant on detailed motion cues. The frequent misrecognition of shaking movements suggests that VLMs may face challenges in extracting frame-to-frame changes, motion dynamics, or perceiving visual changes as a whole. Addressing these challenges could be an important avenue for future work, both in evaluating perceptual sensitivity and in developing techniques to improve VLM's perception ability. Despite these limitations, VLMs demonstrated certain amount of overlap with human interpretations in most categories, and when they successfully perceived the animation, they were generally able to reason about its purpose within the context. These findings suggest that while perceptual challenges continue to hinder performance in certain cases, especially subtle or motion-dependent animations, VLMs have potential to capture, and align with, human interpretations of UI animations.

G.4 Performance on Video Models

We conducte additional evaluations using Video LLaMa, LLaVA-Video, Qwen-2.5-VL, and Gemini-2.5-pro, where we use videos directly as input. The results are listed in Table 12.

H Additional Details for *MCPC*

H.1 MCPC Setup Details

Motion. To explicitly capture temporal dynamics, we generate a simplified recency-weighted blended

Setting 1	Setting 2	p value	p < 0.05	p < 0.1
-	C	0.02	True	True
-	CP	0.09	False	True
-	M	0.46	False	False
-	MC	0.22	False	False
-	MCP	1.00	False	False
-	MP	0.06	False	True
-	P	0.24	False	False
C	CP	0.47	False	False
C	M	0.00	True	True
C	MC	0.20	False	False
C	MCP	0.01	True	True
C	MP	0.00	True	True
CP	M	0.01	True	True
CP	MC	0.64	False	False
CP	MCP	0.03	True	True
CP	MP	0.00	True	True
M	MC	0.05	True	True
M	MCP	0.45	False	False
M	MP	0.32	False	False
MC	MCP	0.17	False	False
MP	MC	0.00	True	True
MP	MCP	0.09	False	True
P	C	0.00	True	True
P	CP	0.01	True	True
P	M	0.78	False	False
P	MC	0.02	True	True
P	MCP	0.27	False	False
P	MP	0.64	False	False

Table 10: Pair-wise statistical significance test for purpose categorization (RQ2) results reported in Table 5. "-" indicates the vanilla model (the base setting in Table 5.

Setting 1	Setting 2	p value	p < 0.05	p < 0.1
-	C	3.43e-02	True	True
-	CP	1.38e-05	True	True
-	M	2.98e-01	False	False
-	MC	1.26e-02	True	True
-	MCP	4.59e-07	True	True
-	MP	3.08e-05	True	True
-	P	1.81e-06	True	True
C	CP	4.37e-03	True	True
C	M	1.31e-03	True	True
C	MC	6.15e-01	False	False
C	MCP	1.97e-03	True	True
C	MP	6.96e-03	True	True
CP	M	5.75e-08	True	True
CP	MC	2.30e-02	True	True
CP	MCP	3.86e-01	False	False
CP	MP	9.17e-01	False	False
M	MC	7.44e-04	True	True
M	MCP	1.05e-09	True	True
M	MP	1.02e-07	True	True
MC	MCP	6.28e-03	True	True
MP	MC	1.93e-02	True	True
MP	MCP	5.05e-01	False	False
P	C	2.65e-03	True	True
P	CP	5.84e-01	False	False
P	M	2.20e-08	True	True
P	MC	1.55e-02	True	True
P	MCP	9.93e-01	False	False
P	MP	7.90e-01	False	False

Table 11: Pair-wise statistical significance test for UI animation interpretation (RQ3) results reported in Table 5. "-" indicates the vanilla model (the base setting in Table 5.

motion image inspired by Motion History Image (Davis, 2001), which integrates changes across multiple frames into a single static representation. This is used as a unified technique to encode motion for models that does and does not have native temporal processing capabilities. The blended image is computed as:

$$B = \frac{1-\gamma}{1-\gamma^N} \sum_{k=1}^{N} \gamma^{N-k} F_k$$

where B is the blended motion image, $F_k \in \mathbb{R}^{H\times W\times C}$ is the k-th frame (indexed $k = 1$ oldest $\rightarrow k = N$ newest), N is the number of frames, and γ is the exponential decay factor set as 0.85 giving higher weight to recent frames (operations are elementwise over pixels/channels). This representation visualizes temporal changes, such as trajectories, transitions, and rotations. In our implementation, we create blended images at 10 fps, where each blended image blends the 6 most recent frames sampled at 60 fps. Example outcomes are illustrated in Figure 15.

Model	Mean	Var	Score distribution (0–5)
gemini-2.5-pro (I)	2.66	1.57	
gemini-2.5-pro (S)	3.28	1.01	
qwen2.5-vl-72b-instruct (I)	2.44	1.53	
qwen2.5-vl-72b-instruct (S)	3.02	1.55	
llava-video-7b-local (I)	2.06	1.39	
llava-video-7b-local (S)	2.29	1.45	
videollama3-7b-local (I)	1.50	1.26	
videollama3-7b-local (S)	1.47	1.32	

Table 12: Video input performance. (I): individually compared. (S): compared to summary.

Context. We evaluate the impact of contextual information by appending textual context description and the user interaction description to the model's prompt. While this information was included by default in prior evaluations, we explicitly varied this factor here to quantify its impact on performance.

Perceptual Caption. Perceptual captions are human-annotated textual descriptions of the animation effects, which function as "alt text" for the visual dynamics. This setup tests the hypothesis that if a model struggles with raw motion perception, providing an explicit textual description of the movement will bridge the perception gap and improve reasoning performance.

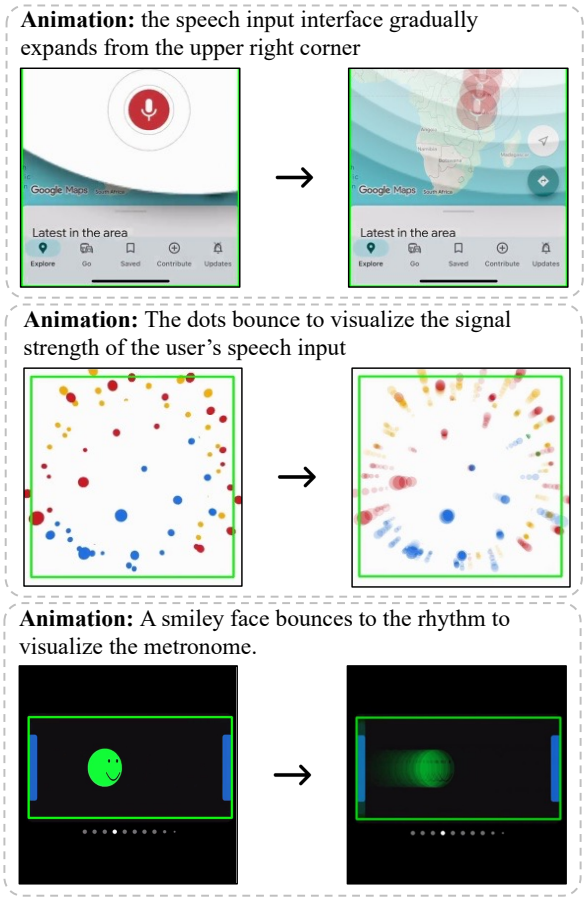


Figure 15: Examples of regular frames vs. motion blended images where motion blended images show the movement patterns in the past few frames.

H.2 Statistical Significance Test

Following Appendix E.3 and G.2, we adopt the McNemar's test and the Wilcoxon signed-rank test for the purpose categorization (RQ2) and UI animation interpretation (RQ3), respectively. Tables 10 and 11 report the results, respectively. For purpose categorization (RQ2), the improvement introduced by *MCP* is not statistically significant. In contrast, for interpretation (RQ3), *MCP* yields a statistically significant improvement.