# Investigating Implicit Support for Image Generation Processes

ANONYMOUS AUTHOR(S)

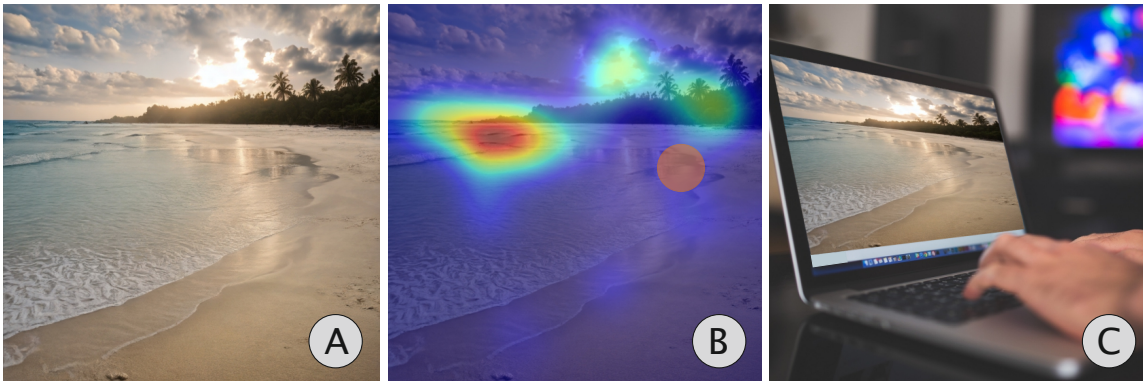


Fig. 1. In our paper, we explore how implicit input can be used to support image generation processes that classically just use text prompts (A). We discuss what information can be extracted using implicit input and how it can be integrated into image generation processes. Our investigation focuses on (B) gaze behavior and (C) keystroke dynamics.

Users of Generative Artificial Intelligence models often struggle to generate a desired image due to difficulties in expressing complex visual concepts. Current approaches to solving this problem, like adding conditional control, require users to give explicit input, which can be tedious. This paper explores how implicit input can be used to support image generation processes. We focus on two exemplary implicit input modalities: gaze behavior and keystroke dynamics. In a preliminary evaluation, we investigated the correlation between gaze behavior and user annotations, showing that users looked longer at areas they wanted to regenerate. Further, we assessed what information can be extracted from keystroke dynamics to be used as an additional input for image generation models.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Generative Artificial Intelligence, Eye Tracking, Image Generation

## 1 Introduction

Generative Artificial Intelligence (GenAI) empowers users to effortlessly produce sophisticated outputs, such as images, from basic inputs, often textual descriptions. By that, users can create images without requiring advanced skills in image editing [23]. Models like Stable Diffusion [7] or FLUX [15] have risen in popularity due to their applications across various fields such as entertainment, advertising, and art [2]. These models primarily leverage natural language input but face ongoing challenges related to the ambiguity of text. Despite the advancements of the models, errors, particularly those resulting from misinterpretation and ambiguous prompts, continue to occur [29]. Additionally, users often initially lack a complete vision of the outcome and, instead, iteratively refine the image. Therefore, current one-shot generation methods may fail to produce the intended results, leaving user goals unmet [33]. Due to this, there is a growing need for supplementary methods allowing users to express intent beyond textual descriptions to maintain control over the generative process [18].

There are two main strategies to enhance user control and the quality of output generated by these models. The first involves additional conditioning beyond text, such as reference images. Reference images assist users in conveying

visual ideas since they carry implicit attributes that are difficult to articulate through text alone [33]. Yet, this still does not solve the problem of users not knowing what they want to generate initially, leading to an iterative exploration and refinement through numerous generation cycles [11]. The second strategy, generative inpainting, focuses on speeding up the iterative refinement by enabling users to regenerate sections of the output while retaining others selectively [32]. This involves marking regions for revision using mouse and text input [17]. Although both strategies rely mainly on explicit input from users, this requirement can be tedious and places an additional cognitive load on users, who must decide what changes are necessary [21].

Implicit input provides an alternative approach by intuitively detecting users' intentions through their behaviors, eliminating the need for explicit specification [26]. Doing so can significantly streamline interactions and potentially enhance the user experience within GenAI systems. However, challenges remain in accurately interpreting these behavioral patterns and integrating them effectively into the image generation process.

This paper explores the potential of implicit input to enhance image generation processes by focusing on two exemplary modalities: gaze behavior and keystroke dynamics. Our preliminary evaluation of gaze behavior reveals a correlation between gaze behavior and user annotations, demonstrating that users tend to fixate longer on areas they want to regenerate. This insight suggests that gaze behavior can be effectively used to implicitly communicate user intent during image generation. Additionally, we investigate the potential of keystroke dynamics to serve as an alternate input mechanism, offering information that can further refine image generation models. By using these implicit inputs, users can communicate their intentions more naturally and seamlessly, potentially enhancing both the efficiency and the user experience of image generation systems. This approach offers insights for reducing the need for explicit input, thus streamlining the interaction process and reduce cognitive load.

## 2 Gaze Behavior for Inpainting

In a preliminary evaluation, we investigated the potential of using gaze behavior to enhance image generation processes. Gaze behavior is usually captured using eye-tracking devices with cameras to measure corneal reflections, enabling the calculation of gaze points on screens or in virtual environments [4]. Gaze points can be used explicitly, such as controlling a mouse for individuals with motor impairments or operating drawing applications [6, 10]. However, this can be challenging due to the difficulty of consciously controlling gaze [12]. Alternatively, eye tracking can serve as an implicit input modality, where gaze behavior is passively analyzed to adapt UI content or trigger functions without active user control [1, 5, 35]. This includes predicting user intentions for assistive robots, automating image cropping, or manipulating images based on gaze saliency [9, 13, 28].

Previous investigations regarding the gaze behavior for Artificial Intelligence (AI)-generated images have demonstrated that gaze is a crucial indicator in determining whether images are AI-generated [3]. Moreover, prior research has confirmed that user intentions can be inferred from gaze patterns [1]. However, to effectively utilize this knowledge to improve AI-generated images, it is essential first to evaluate the connection between users' gaze behaviors and their intentions to change or preserve generated images.

Thus, in our preliminary user study, we aim to address the following research question: What can users' gaze behavior tell us about generated images and the users' intentions to preserve or change the content?

We assume that users' gaze focuses on specific regions that should either be regenerated or remain unchanged, supported by earlier studies linking gaze behavior with user intentions [22]. We hypothesize that users fixate on areas they wish to regenerate due to an implicit negative bias towards AI-generated images, particularly when certain

parts appear unnatural [8, 34]. Additionally, users tend to search implicitly for irregularities in AI-generated images, maintaining their focus on these inconsistencies [3].

To evaluate our hypothesis, we conducted a preliminary within-subject lab study with 16 participants (median age 27). Participants used a custom Python-based UI to view 768 × 768-pixel images and annotate areas to keep or regenerate marked in green and red, respectively. For image generation, we used the Stable Diffusion XL-Turbo [25] network. Gaze behavior was tracked at 250 Hz using an EyeLogic LogicOne[1] eye tracker, which filtered out blinks.

We extracted the fixations from our gaze data using a velocity-based algorithm and overlapped the fixation area with the annotated areas. We calculated an Aligned Rank Transform (ART) [30] before continuing with analysis of variance (ANOVA) and post-hoc paired t-tests with Bonferroni-Holm correction. The ANOVA test revealed a significant effect of annotation type on the region of interest (ROI)s, $F(2, 45) = 8.467, p < 0.001, \eta_p^2 = 0.273$. The posthoc tests showed that users looked significantly more on the regenerate area ($M = 0.235$, $SD = 0.072$) compared to the keep area ($M = 0.133$, $SD = 0.077$), $t(15) = 4.73, p < 0.001, Cohen's\ d = 1.18$, and the area that was not annotated ($M = 0.15$, $SD = 0.068$), $t(15) = 5.86, p < 0.001, Cohen's\ d = 1.47$. Qualitative feedback indicated that participants particularly fixated on image elements perceived as unnatural, such as human figures or iconic scenarios with irregular geometries. Thus, our results confirm our hypothesis.

These results align with previous studies showing that users tend to fixate more on the noise present in images [24]. We theorize that this attention to noise can also extend to flaws in images, whether generated or natural, making users focus on these specific areas. Furthermore, we propose that users inherently notice imperfections in images, even without immediately discerning the specific anomalies [3].

## 3 Keystroke Dynamics Features for Image Generation

A second modality besides gaze behavior that can provide implicit information is keystroke dynamics. Keystroke dynamics focuses on analyzing features like typing speed and dwell times, i.e., the time between consecutive keys for example to identify and authenticate users [20]. Additionally, current research has shown that information like the demographics or emotions of users can be successfully extracted from keystroke dynamics [14, 31].

For demographic prediction, keystroke dynamics can provide information about a user's characteristics, such as gender and age. These insights are derived from several features: typing speed and rhythm variations, dwell time and flight time, and specific typing patterns involving punctuation and function keys that may correlate with gender and language proficiency. Additionally, the frequency of typing errors and correction methods can further provide demographic information [16].

Emotion prediction using keystroke dynamics also presents promising potential for customizing GenAI outputs. Emotional states can manifest as changes in typing speed and rhythm, impacting the variability of typing patterns. Although pressure data is typically gathered from specialized keyboards, it can be inferred from dwell and flight time variations, offering insights into different emotional states. Typing errors and pauses often increase with negative emotions or cognitive stress, providing additional emotional cues. Research has demonstrated that various emotions, such as fear, are more recognizable due to distinctive typing patterns, and machine learning models can achieve classification accuracies exceeding 90% for certain emotions[19].

In the context of image generation, leveraging keystroke dynamics for demographic and emotion prediction allows for creating more personalized and relevant outputs. For instance, understanding a user's emotional state can guide the

[1]EyeLogic. LogicOne. https://www.eyelogicsolutions.com/logicone/ (last visited on February 17, 2025)

generation of visuals that either reflect or counterbalance those emotions, enhancing user engagement and potentially reducing negative emotions by generating content to trigger positive emotions. Additionally, recognizing demographic information enables the tailoring of image content to resonate culturally or contextually with users, improving user experience. However, there is also a risk of reinforcing biases of GenAI models by integrating information about the demographics of the users [27]. Thus, the integration has to be designed carefully to prevent this.

Although challenges remain, particularly in achieving consistent predictions across diverse populations, the integration of keystroke dynamics into generative AI showcases significant potential for advancing interactive and adaptive systems. Ongoing research will be essential in refining these models and expanding their applicability to broader contexts and user demographics.

## 4 Conclusion

In conclusion, this paper has explored the potential of implicit input modalities, specifically gaze behavior and keystroke dynamics, to enhance image generation processes within image generation systems. Our findings highlight that users' gaze behavior can effectively identify areas that should be regenerated, which can be useful for image inpainting. Similarly, keystroke dynamics provide meaningful data related to user demographics and emotional states, which can be leveraged to personalize and refine generative outputs.

While our preliminary evaluations demonstrate the feasibility of integrating these implicit inputs, challenges remain in optimizing their accuracy and effectiveness. Technical limitations in current eye-tracking and keystroke analysis technologies, as well as the complexities of accurately interpreting behavioral data, pose significant challenges. However, as these technologies evolve, they seem promising for reducing reliance on explicit user input, thereby streamlining interactions and minimizing cognitive load.

Future research could focus on expanding the scope of implicit input modalities, including the integration of other physiological signals like respiration rate or EEG, to further enrich user interaction with image generation systems. Additionally, exploring multimodal approaches that combine various implicit cues could lead to more robust and intuitive generative processes. Ultimately, using implicit input has the potential to transform how users engage with AI-driven creative tools, making the experience more natural, efficient, and responsive to individual needs.

## References

[1] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, and Julian Mennenöh. 2012. Increasing the user's attention on the web: using implicit interaction based on gaze behavior to tailor content. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12)*. Association for Computing Machinery, New York, NY, USA, 544–553. doi:10.1145/2399016.2399099

[2] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montr\'{e}al QC Canada, 396–410. doi:10.1145/3600211.3604722

[3] Giuseppe Cartella, Vittorio Cuculo, Marcella Cornia, and Rita Cucchiara. 2024. Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. *IEEE Signal Processing Letters* 31 (2024), 820–824. doi:10.1109/LSP.2024.3375288 Conference Name: IEEE Signal Processing Letters.

[4] Benjamin T. Carter and Steven G. Luke. 2020. Best practices in eye tracking research. *International Journal of Psychophysiology* 155 (Sept. 2020), 49–62. doi:10.1016/j.ijpsycho.2020.05.010

[5] Shiwei Cheng and Ying Liu. 2012. Eye-tracking based adaptive user interface: implicit human-computer interaction for preference indication. *Journal on Multimodal User Interfaces* 5, 1 (March 2012), 77–84. doi:10.1007/s12193-011-0064-6

[6] Aditya Dave and C. Aishwarya Lekshmi. 2017. Eye-ball tracking system for motor-free control of mouse pointer. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. 1043–1047. doi:10.1109/WiSPNET.2017.8299921

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. doi:10.48550/arXiv.2403.03206 arXiv:2403.03206 [cs].

[8] Parul Gupta, Komal Chugh, Abhinav Dhall, and Ramanathan Subramanian. 2020. The eyes know it: FakeET- An Eye-tracking Database to Understand Deepfake Perception. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 519–527. doi:10.1145/3382507.3418857

[9] Nora Horanyi, Yuqi Hou, Ales Leonardis, and Hyung Jin Chang. 2023. G-DAIC: A Gaze Initialized Framework for Description and Aesthetic-Based Image Cropping. *Proceedings of the ACM on Human-Computer Interaction* 7, ETRA (May 2023), 1–19. doi:10.1145/3591132

[10] Lida Huang, Thomas Westin, Mirjam Palosaari Eladhari, Sindri Magnússon, and Hao Chen. 2023. Eyes can draw: A high-fidelity free-eye drawing method with unimodal gaze control. *International Journal of Human-Computer Studies* 170 (Feb. 2023), 102966. doi:10.1016/j.ijhcs.2022.102966

[11] Rong Huang, Haichuan Lin, Chuanzhang Chen, Kang Zhang, and Wei Zeng. 2024. PlantoGraphy: Incorporating Iterative Design Process into Generative Artificial Intelligence for Landscape Rendering. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642824

[12] Robert J. K. Jacob. 1990. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90*. ACM Press, Seattle, Washington, United States, 11–18. doi:10.1145/97243.97246

[13] Fatemeh Koochaki and Laleh Najafizadeh. 2018. Predicting Intention Through Eye Gaze Patterns. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, Cleveland, Ohio, USA, 1–4. doi:10.1109/BIOCAS.2018.8584665 ISSN: 2163-4025.

[14] Puneet Kumar and Balasubramanian Raman. 2022. A BERT based dual-channel explainable text emotion recognition system. *Neural Networks* 150 (June 2022), 392–407. doi:10.1016/j.neunet.2022.03.017

[15] Black Forest Labs. 2023. FLUX. https://github.com/black-forest-labs/flux.

[16] Guoqiang Li, Parisa Rezaee Borj, Loic Bergeron, and Patrick Bours. 2019. Exploring keystroke dynamics and stylometry features for gender prediction on chat data. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 1049–1054.

[17] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 10748–10758. doi:10.1109/CVPR52688.2022.01049

[18] Carina Liebers, Niklas Pfützenreuter, Jonas Auda, Uwe Gruenefeld, and Stefan Schneegass. 2024. "Computer, Generate!" – Investigating User-Controlled Generation of Immersive Virtual Environments. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 213–227. doi:10.3233/FAIA240196

[19] Aicha Maalej, Ilhem Kallel, and Javier Jesus Sanchez Medina. 2022. Investigating Keystroke Dynamics and Their Relevance for Real-Time Emotion Recognition. *SSRN Electronic Journal* (2022). doi:10.2139/ssrn.4250964

[20] Fabian Monrose and Aviel Rubin. 1997. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM conf. on Comp. and comm. security, Apr 1-4*. 48–56.

[21] Diego Navarro and Veronica Sundstedt. 2017. Simplifying game mechanics: gaze as an implicit interaction method. In *SIGGRAPH Asia 2017 Technical Briefs*. ACM, Bangkok Thailand, 1–4. doi:10.1145/3145749.3149446

[22] Joshua Newn, Ronal Singh, Eduardo Velloso, and Frank Vetere. 2019. Combining implicit gaze and AI for real-time intention projection. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 324–327. doi:10.1145/3341162.3343786

[23] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. Bermano, E. Chan, T. Dekel, A. Holynski, A. Kanazawa, C.k. Liu, L. Liu, B. Mildenhall, M. Nießner, B. Ommer, C. Theobalt, P. Wonka, and G. Wetzstein. 2024. State of the Art on Diffusion Models for Visual Computing. *Computer Graphics Forum* 43, 2 (2024), e15063. doi:10.1111/cgf.15063 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.15063.

[24] Florian Röhrbein, Peter Goddard, Michael Schneider, Georgina James, and Kun Guo. 2015. How does image noise affect actual and predicted human gaze allocation in assessing image quality? *Vision Research* 112 (July 2015), 11–25. doi:10.1016/j.visres.2015.03.029

[25] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial Diffusion Distillation. *arXiv preprint arXiv:2311.17042* 1, 1 (2023), 1–10.

[26] Barış Serim and Giulio Jacucci. 2019. Explicating "Implicit Interaction": An Examination of the Concept and Challenges for Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–16. doi:10.1145/3290605.3300647

[27] Ramya Srinivasan and Kanji Uchino. 2021. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 41–51.

[28] Maximilian Söchting and Matthias Trapp. 2022. Tracking Eye Movement for Controlling Real-Time Image-Abstraction Techniques. In *Computer Vision, Imaging and Computer Graphics Theory and Applications*, Kadi Bouatouch, A. Augusto de Sousa, Manuela Chessa, Alexis Paljic, Andreas Kerren, Christophe Hurter, Giovanni Maria Farinella, Petia Radeva, and Jose Braz (Eds.). Springer International Publishing, Cham, 103–123. doi:10.1007/978-3-030-94893-1_5

[29] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3613904.3642803

[30] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[31] Liying Yang and Sheng-Feng Qin. 2021. A Review of Emotion Recognition Methods From Keystroke, Mouse, and Touchscreen Dynamics. *IEEE Access* 9 (2021), 162197–162213. doi:10.1109/ACCESS.2021.3132233

[32] Shiyuan Yang, Xiaodong Chen, and Jing Liao. 2023. Uni-paint: A Unified Framework for Multimodal Image Inpainting with Pretrained Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 3190–3199. doi:10.1145/3581783.3612200

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Paris, France, 3836–3847. https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html

[34] Yizhen Zhou and Hideaki Kawabata. 2023. Eyes can tell: Assessment of implicit attitudes toward AI art. *i-Perception* 14, 5 (Sept. 2023), 20416695231209846. doi:10.1177/20416695231209846 Publisher: SAGE Publications.

[35] Rongrong Zhu, Chi Cheng, Yunpeng Song, and ZhongMin Cai. 2024. Modeling Attentive Interaction Behavior for Web Content Identification in Exploratory Information Seeking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (Nov. 2024), 181:1–181:28. doi:10.1145/3699750