Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 110–121
https://doi.org/10.54941/ahfe1008246

The Mediated Moon: Semantic Mediation Between Gesture and Generative Image

Shu-Yu Lin, Jia-Ming Day, and Der-lor Way

Taipei National University of the Arts, Taipei, Taiwan

ABSTRACT

This paper investigates how vision-language model (VLM)-based semantic mediation can reorganize the relationship between bodily input and real-time generative output in interactive art. The study develops a scaled interactive prototype from *Luminous Night*, an existing moon-inspired spatial light artwork. The system combines MediaPipe hand tracking, a locally deployed Gemma 3 model, and StreamDiffusion-based real-time image generation. A participant's hand contributes both numerical tracking data and visual input; the VLM generates four semantic keywords within an artist-defined semantic field, which then participate in conditioning generative imagery presented through palm projection and LED output. The paper uses semantic mediation to describe this intermediate transformation from bodily input to language and image, and proposes the designed gap as an artistic design proposition for a partially indeterminate interaction in which bodily input remains influential without completely specifying the result. Observations made during prototype development provide an initial account of how participants adjusted and repeated gestures while encountering changing system responses. The contribution lies in articulating semantic mediation as an interaction configuration for generative art and establishing a practical basis for examining the relationship among bodily influence, generative variation, and perceived control.

Keywords: Semantic mediation, Vision-language model, Embodied interaction, Human-AI interaction, Interactive art

BETWEEN GESTURE AND GENERATION

Interactive systems have often been organized around relatively direct relationships between action and response. A gesture triggers an effect, a movement changes a parameter, or a sensor value controls a visual property. Such mappings support legibility and control by allowing participants to anticipate how their actions will affect the system. Generative AI changes this relationship by introducing interpretation and probabilistic generation between input and output. Vision-language models (VLMs) can derive semantic associations from visual input, while diffusion-based image generation produces visual responses that remain variable even when conditioned by similar inputs.

This study explores this changing relationship through semantic mediation, an interaction structure in which bodily input passes through machine interpretation before entering a generative process. Rather than assigning a fixed visual result to each gesture, the system allows a VLM to derive semantic

Received September 21, 2026; Revised October 23, 2026; Accepted November 2, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

cues from the observed hand form. These cues, together with numerical information obtained through hand tracking, participate in shaping real-time image generation. The central research question is therefore: How can VLM-based semantic mediation reorganize the relationship between embodied input and real-time generative visual output when interaction is designed without a fixed one-to-one mapping?

Within this interaction, participant action influences the generative process while the semantic and visual results retain a degree of variation. This relation is described here as the designed gap: a condition between direct control and arbitrary generation in which bodily input remains consequential without completely determining the resulting image. Semantic interpretation and generative variation therefore become part of the interaction itself rather than functioning only as intermediate technical processes.

This inquiry is developed through *Luminous Night*, an existing spatial light artwork whose visual language emerged from the moon, reflection, and the experience of an image that can be perceived without being fully grasped. Its repeated curved structures fragment the circular form of the moon into luminous spatial elements, providing the visual and physical basis for the scaled interactive prototype developed in this study. The prototype integrates hand sensing, VLM interpretation, real-time image generation, projection, and LED output into this existing artistic language. Through this prototype, the paper examines how numerical bodily data and probabilistic semantic interpretation can jointly participate in real-time generative interaction. Its contribution lies in articulating semantic mediation as an interaction configuration for generative art and in proposing the designed gap as a conceptual means of examining the partially indeterminate relationship between bodily input and generated response. Figure 1 summarizes the relationship among embodied input, semantic mediation, generative variation, and visual response.

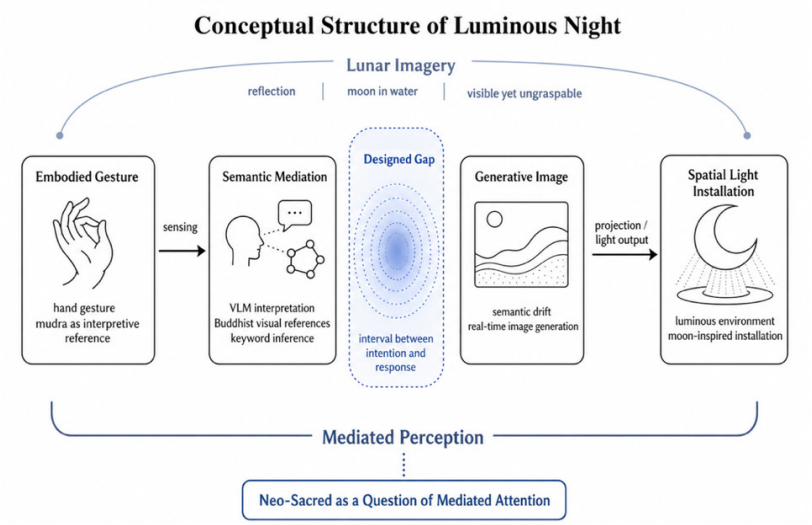


Figure 1: Conceptual structure of semantic mediation and the designed gap, relating embodied input, machine interpretation, and generative visual response.

ARTISTIC AND INTERACTION CONTEXT

Embodied Input Beyond Direct Mapping

Hand gestures occupy a particular position in interactive systems because they can simultaneously function as measurable bodily movements and visually interpretable forms. In conventional gesture interfaces, the hand often serves as a control surface: a swipe advances a screen, a pinch enlarges an image, or a pointing gesture selects an object. Such interfaces commonly rely on relatively stable relationships between action and response; interaction design has also examined how action and function can be coupled through feedback and feedforward (Wensveen et al., 2004). Embodied interaction research broadens this view by treating bodily action as situated rather than simply as abstract input (Dourish, 2001; Suchman, 2007), while somaesthetic interaction design further foregrounds bodily movement, perception, and felt experience in the design process (Höök, 2018). Together, these perspectives provide a basis for considering bodily input as part of an unfolding interaction rather than solely as a set of predefined commands.

Recent AI-based interaction research introduces another possibility by placing semantic interpretation between bodily input and computational response. GestureGPT investigates zero-shot free-form gesture understanding, in which language-model agents attempt to interpret gestures without requiring every possible gesture to be defined in advance as a fixed interaction class (Zeng et al., 2024). Body Prompting similarly explores bodily participation as an alternative to conventional text prompting in human-AI image generation (Oppenlaender et al., 2025). In these approaches, bodily action becomes material for machine interpretation as well as a source of computational input.

The *Luminous Night* prototype develops this possibility within an artistic interaction system. The hand contributes two forms of information: hand tracking produces numerical features that participate in the generative process, while the visual form of the hand provides input for VLM-based semantic interpretation. The same bodily action therefore enters the system as both measurable data and interpretable visual material. This dual role maintains a computational connection to bodily action while introducing an interpretive layer between gesture and visual response, allowing participant action to shape the conditions of generation without specifying a single predetermined image.

AI as Semantic Mediation

Computational mediation has long been part of interactive and new media art. Sensor values may control brightness, proximity may trigger sound, and movement may alter projected imagery; such computational operations participate in the aesthetic organization of new media (Manovich, 2001). Vision-language and generative models extend this structure by introducing processes of inference and generation between sensing and response. A VLM derives semantic associations from visual input through learned statistical relationships, allowing similar inputs to produce related but non-identical descriptions and changes in visual form to lead to different semantic associations.

In the *Luminous Night* prototype, this interpretive process is deliberately incorporated into the interaction pipeline. The observed hand form is interpreted by a VLM, which generates semantic cues that subsequently participate in real-time image generation. The semantic output emerges through the relationship among participant input, the interpretive frame established by the artist, and model inference. The term semantic mediation describes this intermediate transformation: bodily input is partially translated into language before becoming visual material again. Its significance therefore lies in the interaction structure created by this transformation rather than in the classification of gestures alone.

Comparable generative systems have begun to place computational interpretation between non-textual input and image generation. ContextCam uses LLM-based processes to interpret contextual information about the user and environment before incorporating selected context into subsequent image generation with ControlNet and Stable Diffusion (Fan et al., 2024). Although its inputs and interaction goals differ from those of *Luminous Night*, it demonstrates a related configuration in which information beyond a conventional text prompt is computationally interpreted before contributing to generative imagery. Related artistic research has explored bodily forms of AI-mediated interaction: Shadowplay situates embodied interaction within an AI-driven installation (Benjamin and Lindley, 2024); Body Oracle uses bodily posture as material for generative linguistic and symbolic transformation (Huang et al., 2025); and Precious Camouflage examines artistic relationships among human performers, computational interpretation, and AI systems (Triebus et al., 2025). Within this context, *Luminous Night* focuses specifically on the semantic transformation inserted between sensed bodily input and real-time generative imagery.

Generative Uncertainty as an Interaction Condition

Semantic interpretation introduces variation into the interaction. A VLM can generate different semantic cues from similar visual inputs, while diffusion-based image generation introduces further variation between textual conditioning and visual output. The resulting relationship between action and response therefore differs from deterministic mappings based on predefined rules. Ambiguity has been discussed in interaction design as a productive resource that can encourage interpretation rather than simply obstruct understanding. Gaver et al. (2003) argue that ambiguity can create space for users to construct relationships among actions, representations, and meanings. Sengers and Gaver (2006) further examine design approaches that allow multiple, potentially competing interpretations to coexist instead of requiring a single authoritative interpretation to be communicated by the designer.

Interaction research on diffusion-based image generation also illustrates how generative variability becomes entangled with user steering. PromptPaint (Chung and Adar, 2023) addresses the difficulty of shaping end-to-end text-to-image generation by allowing users to mix prompts and apply them across different areas and stages of the generative process. Its user study further reports that the stochasticity and complexity of generative models can make results diverge from user intentions, while iterative interaction provides

opportunities for continued exploration and steering. These findings are relevant to the present study because they show that variation in generative systems is encountered together with questions of influence, intervention, and control. Artistic research has likewise used computational and physiological signals as materials for experiential transformation: From Signal to Sensation redirects sensing toward atmospheric and affective mediation (Hedayati, 2025), while Empathic Growth uses biosignals to construct a shared perceptual environment rather than a definitive representation of internal states (Wu et al., 2025).

For generative interaction, variation becomes meaningful when participants can still perceive a relationship between their actions and the system response. A completely predictable relationship approaches direct control, while an arbitrary relationship weakens the relevance of bodily participation. The designed gap is proposed to describe the interaction space between these conditions, where bodily action contributes to the generative process while VLM interpretation and image synthesis introduce variation into the resulting response. At the present stage, the designed gap functions as an artistic design proposition whose significance depends on how effectively the system maintains this relation between participant input and generative variation. The prototype described in the following chapter provides the practical context through which this proposition is examined.

LUMINOUS NIGHT: FROM SPATIAL ARTWORK TO INTERACTIVE PROTOTYPE

Spatial and Artistic Basis

Luminous Night originated as a large-scale spatial light installation developed from the visual qualities of the moon, shown in Figure 2. Repeated curved structures fragment the circular image of the moon into a sequence of luminous arcs, distributing light across space and allowing the work to be encountered from multiple positions. Earlier exhibitions of the work established its spatial form, luminous materiality, and lunar imagery, which subsequently provided the visual basis for the interaction study presented here.



Figure 2: Previous full-scale installations of *Luminous Night,* showing the curved luminous structures that provide the spatial and visual basis for the interactive prototype.

Because of the scale of the original installation, a proportional physical prototype was later constructed as an environment for developing and testing the interactive system. The prototype retains the characteristic curved structure of *Luminous Night* while incorporating sensing, projection, real-time generative imagery, and LED output. It provides a practical setting for examining how hand input, semantic interpretation, image generation, and luminous output operate together within the spatial language derived from the original artwork.

The overall system architecture is shown in Figure 3. MediaPipe provides hand detection and numerical hand data, while the detected hand image is also submitted to the VLM for semantic interpretation. The resulting semantic cues and numerical features participate in the StreamDiffusion generation process, after which the generated imagery is presented through projection and LED mapping.

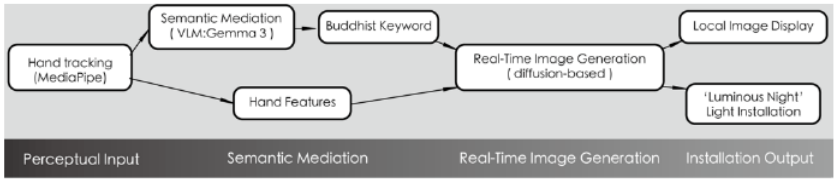


Figure 3: System architecture of the VLM-driven semantic mediation prototype, showing the relation among MediaPipe hand tracking, VLM-generated semantic keywords, StreamDiffusion image generation, projection, and LED output.

Hand Sensing and Embodied Input

The interaction begins with hand detection using MediaPipe. MediaPipe Hands uses a palm-detection stage followed by a hand-landmark model that estimates twenty-one hand landmarks (Zhang et al., 2020). In the prototype, these landmarks are tracked in real time, and the presence of a detected hand initiates the subsequent interpretation process. Hand tracking also provides numerical information describing the participant's hand position and configuration, which is passed into the real-time generative pipeline to modulate generation parameters. The present study considers these parameters at the level of the overall interaction architecture, focusing on how bodily data participates in generation.

At the same time, the detected hand image becomes visual input for the VLM. The system therefore derives two forms of information from the participant's hand: numerical features obtained through MediaPipe and semantic cues produced through VLM interpretation. Both participate in the same generative process, combining measurable bodily data with a variable linguistic layer before image generation.

VLM as Semantic Mediator

The semantic frame of the current prototype developed from the visual language of *Luminous Night*. The title evokes brightness within the night, reflecting the work's original association with moonlight. The installation

transforms the circular image of the moon into repeated luminous arcs, while reflection and partial appearance became recurring elements in its artistic development. Motifs such as the moon reflected on water, flowers in a mirror and the moon in water (鏡花水月), and the attempt to grasp the moon's reflection in water (水中撈月) extended the work's interest in images that appear vividly yet remain intangible or beyond grasp. During the development of the interactive prototype, these associations also informed a limited set of Buddhist-related visual references used to shape the VLM's semantic field.

Within this semantic field, the detected hand image is submitted to Gemma 3, deployed locally through Ollama. Gemma 3 is a multimodal model capable of processing image input and producing text output, enabling visual information to be interpreted through language (Gemma Team, 2025). In the prototype, the model generates four semantic keywords in response to the observed hand form rather than selecting terms from a predefined keyword list. The semantic field provides a shared associative direction, while the model retains variation in the specific terms it produces. Similar hand forms can therefore generate related but non-identical semantic responses.

The resulting four keywords function as intermediate semantic conditions between the observed hand and subsequent image generation. Participant bodily input, the semantic frame developed through the artwork, and VLM inference together shape this linguistic layer before it enters StreamDiffusion. This transformation from visual bodily input into generative semantic material constitutes the semantic mediation examined in this study.

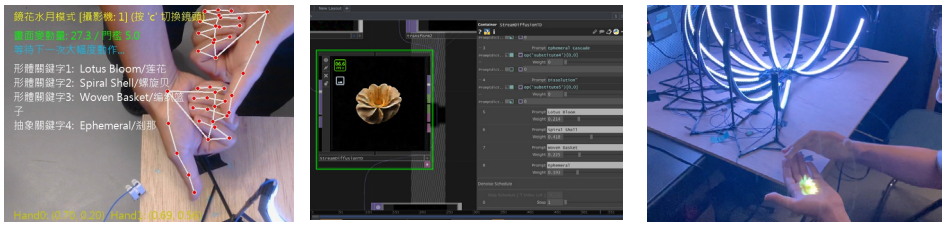
Real-Time Generative Image and Prototype Output

Diffusion-based image generation provides a flexible means of conditioning visual synthesis through information such as language. Latent diffusion models, for example, incorporate conditioning mechanisms including text through cross-attention while performing image synthesis in a latent representation (Rombach et al., 2022). StreamDiffusion extends diffusion-based generation toward streaming and real-time interactive use by restructuring the generation pipeline for higher-throughput continuous image generation (Kodaira et al., 2025).

In the *Luminous Night* prototype, the four VLM-generated keywords are used as semantic prompting input within the StreamDiffusion-based real-time image-generation pipeline. Numerical features derived from hand tracking are also supplied to the pipeline to modulate generation parameters. Semantic interpretation and numerical bodily data therefore jointly participate in shaping the visual response. Because both VLM interpretation and diffusion-based generation introduce probabilistic variation, similar bodily inputs can produce different semantic cues and visual outcomes while remaining conditioned by hand data and the semantic frame of the prototype.

Within the prototype environment, the generated imagery is projected onto the participant's palm and simultaneously mapped onto the LED structure of the scaled *Luminous Night* model. The hand therefore participates first as

an input to the generative process and subsequently as one of the surfaces through which the generated response becomes visible. Figure 4 presents the complete interaction sequence, from hand sensing and semantic keyword generation to real-time image generation and prototype output.



(a) (b) (c)

Figure 4: Interaction workflow of the scaled *Luminous Night* prototype. The three panels show: (a) hand sensing and four VLM-generated semantic keywords, (b) real-time image generation, and (c) the scaled physical prototype used for interactive system development.

DISCUSSION: SEMANTIC MEDIATION AND THE DESIGNED GAP

From Direct Mapping to Semantic Mediation

The interaction structure of the prototype differs from a conventional gesture interface in the role assigned to interpretation. In a direct mapping, the designer establishes the relation between input and output in advance: a particular gesture or sensor value corresponds to a particular response. In the *Luminous Night* prototype, bodily input enters the generative process through two related paths. MediaPipe provides numerical information that modulates generation parameters, while the visual form of the hand is interpreted by the VLM and transformed into semantic material. The visual response therefore emerges through the interaction of bodily data, machine interpretation, and image generation.

This arrangement makes the VLM more than an additional input mechanism. Its role is to insert a semantic transformation between sensing and generation. The participant does not formulate a text prompt, while the model produces linguistic material from the observed hand within an artist-defined semantic field. Semantic mediation therefore describes a shift in where interaction is designed: the artist specifies the conditions through which interpretation takes place, while part of the resulting relation between gesture and image is produced during interaction.

Participant input continues to shape the process through both numerical hand data and VLM interpretation, while the VLM and diffusion model introduce probabilistic transformations. The resulting interaction is consequently organized around a partially variable relationship between bodily action and generated response rather than a fixed gesture vocabulary.

The Designed Gap: Influence Without Full Determination

The designed gap describes the relation produced by this configuration. Its central characteristic is that participant action influences the generative response without completely specifying it. A fixed gesture-to-image mapping minimizes this interval by making the result predictable, whereas a response with little perceptible connection to bodily input weakens the relevance of participation. The designed gap occupies the space between these conditions, where participants can affect the system while encountering results that remain partially indeterminate.

This indeterminacy occurs at more than one stage. The VLM can produce different semantic cues from similar hand forms, and StreamDiffusion introduces further variation when translating those cues into imagery. The resulting gap is therefore primarily semantic and generative. Temporal latency represents a different system characteristic: a delayed response can still be completely deterministic, while a rapid response can remain semantically indeterminate. Distinguishing these conditions helps locate the designed gap in the transformation between bodily input, semantic inference, and generated output rather than in response time alone.

The distinction between meaningful variation and arbitrary randomness is equally important. Generative variation becomes relevant to interaction when participant input continues to constrain the range of possible responses. In the current prototype, this relation is shaped through the combination of hand-tracking data and a bounded semantic field for VLM interpretation. Systematic comparison of repeated inputs, generated keywords, and resulting images would allow the degree and structure of this variation to be characterized more precisely in subsequent development.

Ambiguity therefore functions here as a designed condition that leaves room for interpretation. Gaver et al. (2003) describe ambiguity as a resource through which users can construct relationships among actions, representations, and meanings. In *Luminous Night*, the designed gap extends this principle into a generative AI context by locating ambiguity within the changing relation among bodily action, machine interpretation, and generated imagery.

Situated Interaction in the Prototype

Observations made during prototype development provide an initial view of how this interaction structure is encountered in practice. Participants frequently repeated or modified their hand configurations while observing changes in the generated image. Adjustments in hand angle, distance, and finger configuration often followed one another as participants compared successive responses and explored how the system was interpreting their actions.

These behaviors indicate an ongoing process of testing and interpretation during interaction. A generated image that appeared closely related to a previous response could encourage repetition, while an unexpected result could prompt a different hand configuration or another attempt. Such actions

can reflect several overlapping processes, including exploration, calibration, curiosity, uncertainty, and attempts to understand the relationship between bodily action and generative output. They provide situated observations of how participants negotiate an interaction whose mapping is only partially predictable.

At this stage, these observations provide an initial context for examining semantic mediation and the designed gap. More systematic records of repeated gesture–keyword–image sequences, together with participant accounts from a full-scale exhibition setting, would make it possible to investigate more precisely when generative variation remains perceptibly connected to bodily action and when that relationship begins to weaken.

CONCLUSION

This paper presented a VLM-driven interactive prototype developed from the spatial light artwork *Luminous Night*. The system combines MediaPipe hand tracking, VLM-generated semantic keywords, StreamDiffusion-based real-time image generation, palm projection, and LED output. Within this configuration, the participant's hand contributes both numerical information and visually interpretable material to the generative process.

The study uses semantic mediation to describe the transformation through which bodily input passes through machine-generated language before becoming visual material. This configuration reorganizes the relationship between gesture and output as a process jointly shaped by participant action, artist-defined semantic conditions, VLM inference, and diffusion-based generation. The designed gap further identifies the partially indeterminate relationship produced by this process: bodily input remains influential while the resulting semantic and visual responses retain variation.

Through the scaled prototype, the study establishes a practical basis for examining this form of generative interaction. Further development can extend this foundation through systematic records of input–semantic–output variation and through observations of interaction when the system is deployed within the full-scale spatial installation. These investigations can provide a more precise account of when generative variation remains perceptibly connected to bodily action and how participants negotiate the relationship between influence and indeterminacy.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant No. 114-2420-H-119-001.

REFERENCES

Benjamin, J.J. and Lindley, J. (2024). "Shadowplay: An Embodied AI Art Installation." Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '24), Article 93, 3 p. https://doi.org/10.1145/3623509.3635318

Chung, J.J.Y. and Adar, E. (2023). "PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions." Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), Article 6, pp. 1–17. https://doi.org/10.1145/3586183.3606777

Dourish, P. (2001). Where the Action Is: The Foundations of Embodied Interaction. MIT Press, Cambridge, MA.

Fan, X., Wu, Z., Yu, C., Rao, F., Shi, W. and Tu, T. (2024). "ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation." Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Article 157, pp. 1–17. https://doi.org/10.1145/3613904.3642129

Gaver, W.W., Beaver, J. and Benford, S. (2003). "Ambiguity as a Resource for Design." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03), pp. 233–240. https://doi.org/10.1145/642611.642653

Gemma Team. (2025). Gemma 3 Technical Report. arXiv:2503.19786.

Hedayati, M. (2025). "From Signal to Sensation: A Critical Practice of Affective Mediation." Conference on Animation and Interactive Art (Expanded 2025), pp. 109–116. https://doi.org/10.1145/3749893.3749898

Höök, K. (2018). Designing with the Body: Somaesthetic Interaction Design. MIT Press, Cambridge, MA.

Huang, D., Huang, K., Wang, R. and Hu, B.A. (2025). "Body Oracle: Exploring Somatic Hieroglyphs for Collective Bodily Awareness." Conference on Animation and Interactive Art (Expanded 2025), pp. 117–124. https://doi.org/10.1145/3749893.3749912

Kodaira, A., Xu, C., Hazama, T., Yoshimoto, T., Ohno, K., Mitsuhori, S., Sugano, S., Cho, H., Liu, Z., Tomizuka, M. and Keutzer, K. (2025). "StreamDiffusion: A Pipeline-Level Solution for Real-Time Interactive Generation." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12371–12380. https://doi.org/10.1109/ICCV51701.2025.01150

Manovich, L. (2001). The Language of New Media. MIT Press, Cambridge, MA.

Oppenlaender, J., Johnston, H., Silvennoinen, J.M. and Barranha, H. (2025). "Artworks Reimagined: Exploring Human-AI Co-Creation through Body Prompting." Proceedings of the ACM on Human-Computer Interaction, 9(4), Article EICS012, pp. 1–34. https://doi.org/10.1145/3734189

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

Sengers, P. and Gaver, W. (2006). "Staying Open to Interpretation: Engaging Multiple Meanings in Design and Evaluation." Proceedings of the 6th Conference on Designing Interactive Systems (DIS '06), pp. 99–108. https://doi.org/10.1145/1142405.1142422

Suchman, L. (2007). Human-Machine Reconfigurations: Plans and Situated Actions. 2nd ed. Cambridge University Press, Cambridge.

Triebus, C., Druzetic-Vogel, I. and Geiger, C. (2025). "Precious Camouflage: Artistic Research for the Creation of a Dance Piece with Humans and AI." Conference on Animation and Interactive Art (Expanded 2025), pp. 188–198. https://doi.org/10.1145/3749893.3749964

Wensveen, S.A.G., Djajadiningrat, J.P. and Overbeeke, C.J. (2004). "Interaction Frogger: A Design Framework to Couple Action and Function through Feedback and Feedforward." Proceedings of the 5th Conference on Designing Interactive Systems (DIS '04), pp. 177–184. https://doi.org/10.1145/1013115.1013140

Wu, Z.-W., Yang, M., Fei, D.-L., Shi, J.-Q. and McCormack, J. (2025). "Empathic Growth: Designing Human-Plant Interaction through Artistic Exploration." Conference on Animation and Interactive Art (Expanded 2025), pp. 213–221. https://doi.org/10.1145/3749893.3749973

Zeng, X., Wang, X., Zhang, T., Yu, C., Zhao, S. and Chen, Y. (2024). "GestureGPT: Toward Zero-Shot Free-Form Hand Gesture Understanding with Large Language Model Agents." Proceedings of the ACM on Human-Computer Interaction, 8(ISS), Article 545, pp. 462–499. https://doi.org/10.1145/3698145

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L. and Grundmann, M. (2020). "MediaPipe Hands: On-device Real-time Hand Tracking." arXiv preprint arXiv:2006.10214.