Beyond Semantic Similarity: Open Challenges for Embedding-Based Creative Process Analysis Across AI Design Tools

SEUNG WON LEE*, Design Informatics Lab, Hanyang University, Republic of Korea and Human-Centered AI Design Institute, Hanyang University, Republic of Korea

SEMIN JIN*, Design Informatics Lab, Hanyang University, Republic of Korea and Human-Centered AI Design Institute, Hanyang University, Republic of Korea

KYUNG HOON HYUN†, Design Informatics Lab, Hanyang University, Republic of Korea and Human-Centered AI Design Institute, Hanyang University, Republic of Korea

AI-based creativity support tools (CSTs) are evaluated through domain-specific metrics, limiting cross-domain comparison of creative processes. Embedding-based protocol analysis offers a potential domain-agnostic analytical layer. However, we argue that fixed embedding similarity can misrepresent creative dynamics: it may not detect creative pivots that occur within superficially similar language, treating shifts in the problem being addressed as continued elaboration. We identify three open challenges stemming from this gap: aligning similarity measures with creative significance, segmenting and representing multimodal design traces, and evaluating agentic systems where embedding-based metrics enter the generation loop and shape agent behavior. We propose context-aware interventions using large language models as a direction for making trace analysis sensitive to session-specific creative dynamics.

Additional Key Words and Phrases: creative activity traces, creativity support tools, design process analysis, linkography, AI-mediated creativity

ACM Reference Format:

Seung Won Lee, Semin Jin, and Kyung Hoon Hyun. 2026. Beyond Semantic Similarity: Open Challenges for Embedding-Based Creative Process Analysis Across AI Design Tools. 1, 1 (February 2026), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

AI-based creativity support tools (CSTs) are evaluated through domain-specific metrics, making findings difficult to compare across systems. This paper asks how such cross-domain comparison might be achieved, and what obstacles remain.

We argue that cross-domain comparison of creative processes requires shifting from content-level evaluation—judging the quality or characteristics of final design outputs—to process-level structural analysis that examines how

*Both authors contributed equally to this work.
†Corresponding author.

Authors' Contact Information: Seung Won Lee, lswgood0901@gmail.com, Design Informatics Lab, Hanyang University, Seoul, Republic of Korea and Human-Centered AI Design Institute, Hanyang University, Seoul, Republic of Korea; Semin Jin, tpals97@gmail.com, Design Informatics Lab, Hanyang University, Seoul, Republic of Korea and Human-Centered AI Design Institute, Hanyang University, Seoul, Republic of Korea; Kyung Hoon Hyun, hoonhello@gmail.com, Design Informatics Lab, Hanyang University, Seoul, Republic of Korea and Human-Centered AI Design Institute, Hanyang University, Seoul, Republic of Korea.

Manuscript submitted to ACM

creative exploration unfolds over time: how designers branch into new directions, revisit earlier ideas, and develop concepts through iteration. Embedding-based trace analysis offers a promising approach. It takes the sequence of discrete design actions that a designer performs during a design process and projects them into a shared representational space using neural embedding models. From these representations, one can construct a fuzzy linkograph [9], a directed graph that maps semantic relationships between design moves, and derive metrics such as entropy and link density that characterize the process independently of domain-specific content. This reasoning finds support in recent work on recommendation evaluation: Ishii et al. [4] showed that conventional similarity-based metrics become unreliable across systems because they conflate item familiarity with meaningful exploration, and that an information-theoretic formulation enables more robust cross-system comparison. Creative process evaluation faces an analogous challenge: fixed embedding models measure surface-level semantic overlap between moves, so they cannot distinguish genuine conceptual continuity from creative pivots that happen to share similar vocabulary—the kind of subtle but significant shifts that often drive a design process forward.

However, realizing this vision is far from straightforward. This paper identifies three open challenges at the intersection of representation and evaluation: (1) whether the similarity structures captured by current embedding models align with what matters creatively, (2) how such analysis can accommodate multimodal design traces, and (3) how to evaluate creative processes in increasingly agentic AI systems where both humans and AI contribute to the creative trajectory.

2 From Siloed Evaluations to Process-Level Analysis

Current CST evaluations lack a principled basis for comparing process structures across tools. Subjective instruments such as the Creativity Support Index (CSI) [1] and NASA-TLX [3]—employed, for instance, by GenQuery [10] and InkSpire [7] to assess creative support in visual search and product design, respectively—capture perceived experience. *Manual process analysis*, as in Lee et al.'s [6] linkographic comparison of sketch- and prompt-guided 3D modeling, can reveal modality-specific process dynamics but requires significant human effort and resists standardization. *Domain-specific quantitative metrics*, such as FontCraft's [11] convergence measures for optimization-driven font design, are tightly coupled to particular workflows. None of these approaches addresses this gap.

Fuzzy-linkography [9] offers a potential path forward. It automates linkograph construction [2] by computing semantic similarity between sequential design moves using embedding models, deriving established metrics such as entropy, link density, and critical moves [5]. The key property is that it operates on traces of design activity rather than on domain-specific outputs: any CST producing recordable design moves—prompts, sketches, selections, or parameter adjustments—could in principle be analyzed through this lens. However, whether such analysis yields valid cross-tool comparisons depends on several assumptions that remain untested.

3 Open Challenges and Research Agenda

Realizing this potential requires addressing several open questions that we propose as a research agenda for the community.

3.1 Semantic Similarity vs. Creative Significance

General-purpose embeddings measure semantic overlap between utterances, but in design processes, creative significance often resides in shifts that occur within superficially similar language. Consider a designer working on a micro-apartment project who starts with *"stackable chair modules for compact storage"*—solving the problem of limited floor space when furniture is not in use. The designer then proposes *"stackable wall modules for reconfigurable room*

layouts," realizing that the same principle could let residents transform a studio into separate sleeping and working areas on demand. An embedding model would score these moves as highly similar—both share *"stackable"* and *"modules"*—but the second move is a creative pivot: the designer has reframed a furniture storage strategy as a solution to an entirely different problem, spatial adaptability in constrained housing. Under embedding-based trace analysis, such moves would be linked as continuous elaboration of a single idea, when in fact they mark a shift in the problem being addressed. The resulting link structure would understate the exploratory breadth of the session, and derived metrics such as entropy would reflect apparent conceptual continuity rather than the actual dynamics of ideation.

Large language models offer several possible intervention points: preprocessing raw interaction logs to segment design moves around meaningful conceptual shifts; assisting link formation by judging whether two moves share creative relevance given the session's task framing and preceding trajectory; or reinterpreting multimodal traces by extracting design intent from sketches or images that fixed embeddings would reduce to visual similarity alone. Whether such context-aware interventions produce link structures that better align with expert judgment remains an empirically testable question, and we see their investigation as a concrete next step toward closing the gap between system-computed and creativity-relevant measures.

3.2 Multimodal Trace Integration

Most CSTs involve multimodal interaction, yet current automated trace analysis focuses primarily on text. While multimodal embedding models such as CLIP [8] could extend fuzzy-linkography to visual modalities, computing similarity between images does not straightforwardly translate to measuring creative process dynamics. Two images might be visually similar yet represent different design strategies, or visually dissimilar yet represent a coherent creative evolution—a rough early sketch and a refined final rendering may share little visual similarity while representing continuous development of the same concept.

Moreover, the meaningful unit of analysis—what constitutes a "design move"—becomes more ambiguous in multimodal contexts. In text-based interaction, a prompt-response pair provides a natural segmentation boundary. But in InkSpire's [7] sketch-and-generate cycles, a single design move might span a sequence of pen strokes, a generation request, and a selective incorporation of AI suggestions. In FontCraft's [11] preference-based optimization, each feedback iteration involves an implicit comparison across multiple candidates—is this one move or several? Without principled segmentation methods, the resulting linkographic structures may reflect arbitrary analysis choices rather than meaningful process dynamics. Developing domain-informed but generalizable segmentation heuristics is a prerequisite for extending embedding-based analysis beyond text.

3.3 Evaluating Creative Processes in Agentic AI Systems

As CSTs evolve toward agentic architectures—where AI agents autonomously generate alternatives and steer creative direction—embedding-based metrics take on a different role. Agents must evaluate their own design trajectories in real time to decide what to propose next, when to shift direction, and when to continue elaborating. If these evaluations rely on embedding similarity, the limitations described above no longer sit at the analysis stage—they enter the generation loop itself, shaping what the agent produces.

This compounds the evaluation problem. An agent's generation policy—its diversity settings, sampling strategy, and steering heuristics—will leave signatures in any process trace. An agent configured to maximize output variety may produce traces with high link entropy that reflect its policy rather than genuine creative development. Without methods to disentangle such artifacts from the designer's creative dynamics, process metrics cannot reliably assess whether an

agentic CST is a productive collaborator or one that merely produces varied output. Developing such methods might require comparative studies across different levels of agent autonomy on matched design tasks—an empirical agenda the community has yet to undertake.

4 Conclusion

We have argued that embedding-based trace analysis offers a promising direction for cross-domain comparison of creative processes in AI-based CSTs, but that its utility depends on closing the gap between what fixed embedding similarity captures and what constitutes creative significance for human designers. This gap manifests in three dimensions: in the failure to detect creative pivots that occur within superficially similar language, in the added ambiguity of segmenting and representing multimodal design activity, and in agentic systems where embedding-based evaluation enters the generation loop and shapes the agent's creative behavior itself. We propose that large language models, leveraged as context-aware intermediaries in the analytical pipeline, offer a viable path toward making trace analysis sensitive to session-specific creative dynamics, and that comparative studies across different levels of AI agency could provide the empirical grounding this agenda requires.

Acknowledgments

This work was supported by the Industrial Technology Innovation Program(RS-2025-02317326, Development of AI-Driven Design Generation Technology Based on Designer Intent) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

[1] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools Through the Creativity Support Index. In *ACM Transactions on Computer-Human Interaction*, Vol. 21. ACM, 1–25.

[2] Gabriela Goldschmidt. 2014. *Linkography: Unfolding the Design Process.* MIT Press, Cambridge, MA.

[3] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (1988), 139–183.

[4] Tatsuya Ishii, Tianxiang Yang, and Masayuki Goto. 2026. BIG-PU: An Evaluation Metric for Exploration Based on Preference Elicitation in Recommender Systems. *Expert Systems with Applications* 308 (2026), 131077. doi:10.1016/j.eswa.2025.131077

[5] Jeff W.T. Kan and John S. Gero. 2007. Quantitative Analysis of Design Protocols: A Comparison of Entropy Measures and the Shannon Entropy of the Design Process. In *Proceedings of the International Conference on Engineering Design.*

[6] Seung Won Lee, Tae Hee Jo, Semin Jin, Jiin Choi, Kyungwon Yun, Sergio Bromberg, Seonghoon Ban, and Kyung Hoon Hyun. 2024. The Impact of Sketch-guided vs. Prompt-guided 3D Generative AIs on the Design Exploration Process. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM. doi:10.1145/3613904.3642218

[7] David Chuan-En Lin, Hyeonsu B. Kang, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K. Hong. 2025. InkSpire: Supporting Design Exploration with Generative AI through Analogical Sketching. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM. doi:10.1145/3706598.3713397

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning.*

[9] Amy Smith, Barrett R. Anderson, Jasmine Tan Otto, Isaac Karth, Yuqian Sun, John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2025. Fuzzy Linkography: Automatic Graphical Summarization of Creative Activity Traces. In *Proceedings of the 2025 Conference on Creativity and Cognition.* ACM. doi:10.1145/3698061.3726915

[10] Kihoon Son, DaEun Choi, Tae Soo Kim, Young-Ho Kim, and Juho Kim. 2024. GenQuery: Supporting Expressive Visual Search with Generative Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM. doi:10.1145/3613904.3642847

[11] Yuki Tatsukawa, I-Chao Shen, Mustafa Doga Dogan, Anran Qi, Yuki Koyama, Ariel Shamir, and Takeo Igarashi. 2025. FontCraft: Multimodal Font Design Using Interactive Bayesian Optimization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM. doi:10.1145/3706598.3713863