Toward Machine-Interpretable Spatial Organization Patterns: Learning from Creativity Traces

MARYAM REZAIE and SHEELAGH CARPENDALE, Simon Fraser University, Canada

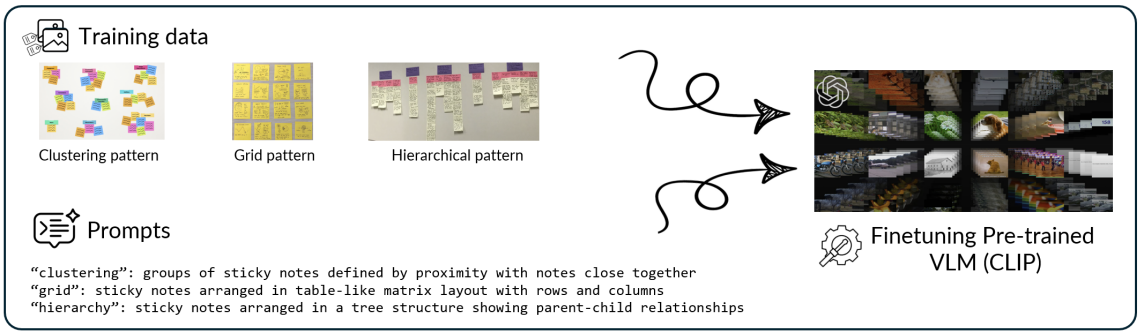

Fig. 1. Overview of our approach. Labeled image dataset exhibiting three spatial organization patterns (clustering, grid, hierarchy) are paired with descriptive prompts and used to fine-tune a pre-trained vision–language model (CLIP). The fine-tuned model learns to classify spatial organization structures from images.

Many creative processes, such as ideation, leave behind residues that often depict spatial organization structures that reflect how people organize ideas. While such structures are meaningful to humans, they are difficult for AI systems to interpret or preserve when transitioning to digital environments. As a step toward enabling machine understanding of spatial aspects of creativity traces, we fine-tune a pre-trained vision-language model (CLIP) to detect and differentiate 3 common spatial organization patterns: clustering, grid and hierarchical pattern. We see this work as an initial step toward enabling computational models to recognize and preserve spatial reasoning structures in hybrid physical–digital creativity support systems.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Additional Key Words and Phrases: Spatial Organization, layout recognition, Finetuning, CLIP

ACM Reference Format:
Maryam Rezaie and Sheelagh Carpendale. 2026. Toward Machine-Interpretable Spatial Organization Patterns: Learning from Creativity Traces. 1, 1 (February 2026), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' Contact Information: Maryam Rezaie, maryam_rezaie@sfu.ca; Sheelagh Carpendale, sheelagh@sfu.ca, Simon Fraser University, Burnaby, BC, Canada.

Manuscript submitted to ACM

1 Introduction

Creative processes such as ideation frequently leave behind artifacts [3]. These artifacts such as sticky notes are often arranged in spatial configurations that represent externalized cognitive structures. For example, ideas may be clustered, sequenced, or organized hierarchically, often accompanied by annotations such as arrows indicating relationships [7]. These spatial arrangements reflect how people structure and communicate ideas during creative work.

However, while humans readily perceive meaning in these spatial organizations and use spatial organization to externalize and communicated ideas, machines currently lack an understanding of spatial organization within creativity artifacts. Digital creativity support systems often treat spatial layout as arbitrary positioning data rather than as semantically meaningful structure [6]. When physical whiteboards are digitized, or when ideation artifacts, such as sticky note board, transitions to digital canvases, the spatial organization are frequently flattened into only coordinates missing the structural relationships between the items (e.g clustering, hierarchy). Currently, creativity enabling tools such as Miro [4], can detect sticky notes and their coordinates and analyze text content, but they struggle to interpret how ideas are spatially arranged.

Previous works, identified recurring spatial organization patterns in physical ideation settings, including clustering, Lines, Matrix, hierarchical, circular, etc. [2, 7]. Building on this empirical foundation, this position paper proposes a computational step toward enabling machine interpretation of such structures. This proposed approach, explores how vision-language models can be adapted to recognize structured spatial configurations of sticky notes as creativity activity traces. Specifically, we fine-tune a CLIP-based vision-language model, to enable detecting spatial organization pattern of images (Figure 1). By enabling machine recognition of ideation layouts, digital whiteboards and hybrid physical–digital creativity support tools could begin to preserve, reflect, and augment spatial reasoning rather than merely store it. This work therefore initiates a method for advancing machine literacy in spatial creativity residues, opening new possibilities for AI-assistance in creative processes.

2 Methodology

Our goal was to automatically classify the spatial organization patterns of images of arranged sticky notes on digital or physical boards. We focused on three common patterns for this work: clustering (multiple distinct groups), grid (table-like arrangements), and hierarchical (tree structures with parent-child relationships).

We collected a dataset of board images showing sticky notes organized into one of three spatial pattern categories: clustering, grid, and hierarchical, excluding other configurations for this initial investigation. Following the spatial descriptions in [7], the *Clustering* pattern consists of multiple separate groups or clusters of sticky notes, where groups are separated by space, visual boundaries (such as drawn outlines or curves), or labels. *Grid* pattern represents a structured table-like organization with sticky notes arranged in rows and columns, often with visible grid lines or table frames. *Hierarchical* pattern shows a tree structure with clear parent-child relationships, typically indicated by arrows, lines, color coding, or size differences across at least three levels. Images were collected from various sources and manually annotated by categorizing each whiteboard image into one of the three pattern types. The dataset was split into training and validation sets, with the validation set containing 8 images, per pattern category (24 images total). The training set contained 75 images across all three categories.

Initial Approach: Zero-Shot Classification We began with using CLIP [5], a pre-trained vision-language model that learns to associate images with text descriptions, to classify whiteboard images without any training. We provided the model with natural language descriptions of each pattern; for example, "sticky notes arranged in a structured

grid with rows and columns" and asked it to identify which description best matched each image. This initial attempt yielded near-random performance (roughly 33% accuracy for a 3-class problem) which means CLIP model struggled to distinguish between patterns.

Iterative Prompt Refinement As first attempt to improve classification accuracy, we carefully refined how we described each pattern. Through iterative experimentation, we discovered that the model's performance was highly sensitive to the specific language used in our prompts. For example, we initially described clustering as "groups defined by proximity" but found the model often confused this with grids where notes happened to be grouped. We revised our prompts to explicitly emphasize the key distinguishing feature: "multiple separate groups or clusters, not a single unified grid." Similarly, for grids, we added descriptions that highlighted visual cues like "drawn table or grid template in the background" to capture both perfectly aligned grids and hand-drawn table structures. We also used multiple descriptions per pattern (5-6 prompts each) rather than a single description, allowing the model to match against whichever phrasing best captured the visual features of each specific image. This prompt engineering improved performance to approximately 50-60% accuracy: better than random, but still insufficient for practical use.

Contrastive Fine-Tuning To bridge the remaining gap, we fine-tuned the vision-language model on our specific dataset using contrastive learning [1], where the model learns to bring together images and text descriptions that match (positive pairs) while pushing apart images and descriptions that do not match (negative pairs). For each training image, the model learned to maximize its confidence in the correct pattern's descriptions (positive examples) while minimizing confidence in other patterns' descriptions (negative examples). This approach preserved the model's ability to reason about natural language descriptions while specializing it to the specific visual characteristics of whiteboard spatial patterns.

The fine-tuned CLIP model achieved approximately 80% accuracy (19 out of 24 validation images correct), a substantial improvement over the zero-shot baseline. Performance was balanced across classes: clustering 7/8 (87.5%), grid 6/8 (75%), and hierarchical 6/8 (75%). The model shows high confidence on correct predictions, with many probabilities near 1.0, indicating clear pattern recognition.

The fine-tuning process successfully improved recognition across all three pattern types, suggesting that contrastive learning with domain-specific prompts helps the model distinguish spatial patterns. The remaining errors reveal systematic challenges: some clustering images are misclassified as hierarchical, likely when clusters have clear internal structure or vertical arrangements that resemble tree-like organization. Some grid images are misclassified as clustering when the grid structure is less regular or when spacing creates apparent groupings. These errors suggest the model relies on visual cues that can be ambiguous at pattern boundaries, which is expected given the small dataset size and the inherent overlap between some spatial organizations.

3 Conclusion and Future Work

This work represents an initial step toward enabling machine recognition of spatial creativity residues. We focused on three spatial organization patterns: clustering, grid, and hierarchical structures as a constrained proof of concept. The fine-tuned model substantially outperformed zero-shot classification, although the dataset is limited. Expanding the range of spatial organization patterns (e.g., circular) and increasing dataset diversity are important next steps. Further improvements in model accuracy, robustness to ambiguous layouts will be necessary before such systems can reliably operate in real-world creativity support contexts. Beyond performance metrics, this work opens a broader research direction: if computational models can recognize spatial organization in creative artifacts, how might creativity support tools leverage this understanding to more effectively support ideation and other creative activities?

We position this contribution not as a complete solution, but as an initiation: a demonstration that spatial ideation structures are computationally learnable and worthy of further investigation. By advancing machine literacy in spatial organization, we aim to encourage deeper integration of spatial reasoning into the analysis and design of creative activity trace systems.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[2] Aron D. Fischel and Kim Halskov. 2020. Chapter 9 - A framework for sticky note information management. In *Sticky Creativity*, Bo T. Christensen, Kim Halskov, and Clemens N. Klokmose (Eds.). Academic Press, 199–230. doi:10.1016/B978-0-12-816566-9.00009-4

[3] David Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454.

[4] Miro. 2024. Miro. https://miro.com Online; accessed 19 February 2026.

[5] 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*. PmLR, 8748–8763.

[6] Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2020. Evaluating creativity support tools in HCI research. In *Proceedings of the 2020 ACM designing interactive systems conference*. 457–476.

[7] Maryam Rezaie, Samuel Huron, Parnian Taghipour, Lien Quach, Victor Cheung, and Sheelagh Carpendale. 2025. Studying Visual Evidence from Using Physical Space to Think. *Proc. ACM Hum.-Comput. Interact.* 9, 8, Article ISS011 (Nov. 2025), 23 pages. doi:10.1145/3773068

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009