Latent Diffusion-Based Cross-Model Learning for Spatial Transcriptomics in Human Skin

Vedhsai Thiriveedi*,
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

ABSTRACT

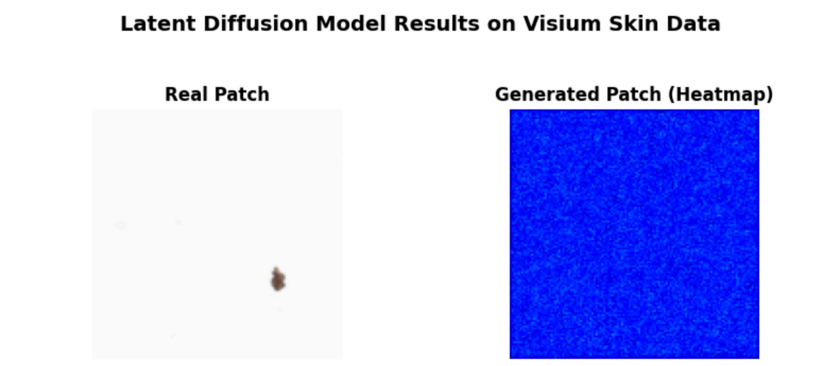
- Spatial transcriptomics enables gene expression mapping directly onto tissue sections, but integrating molecular and morphological features remains a challenge. We developed a novel framework combining Contrastive Language–Image Pretraining (CLIP) with a Latent Diffusion Model (LDM) to align histology patches with transcriptomic profiles in human skin. Using 10x Genomics Visium data, tissue sections were tiled into 256 × 256 patches and linked to the top 50 expressed genes per spot. CLIP embeddings provided a shared latent space for image and gene features, and the LDM was trained to reconstruct histology conditioned on transcriptomic input.
- Our model achieved strong embedding alignment, with cosine similarity of 0.933, LPIPS of 1.095, and SSIM of 0.102, demonstrating preservation of transcriptomic structure despite limited visual fidelity. Heatmap visualization highlighted transcriptomic signal capture within reconstructed patches. These findings provide the first proof-of-concept for latent diffusion in skin spatial transcriptomics, supporting future efforts to improve reconstruction quality, enable cluster-level classification, and integrate pathologist-verified labels for clinical use.

INTRODUCTION

- Spatial transcriptomics (ST) is transforming tissue biology by linking gene expression to intact histological architecture. In skin, ST can uncover spatial heterogeneity among keratinocytes, fibroblasts, melanocytes, and immune populations. However, most current analyses focus only on gene clustering or dimensionality reduction, providing limited integration with histological morphology.
- Deep learning frameworks offer opportunities to unify molecular and visual features. Convolutional neural networks capture local morphology but struggle with high-dimensional gene expression patterns. Latent Diffusion Models (LDMs), when paired with multimodal encoders such as CLIP, can enable cross-modal translation between transcriptomic and histological domains. This work applies LDMs to human skin ST, evaluating the feasibility of cross-modal reconstruction.

RESULTS

- The CLIP + Latent Diffusion framework achieved strong cross-modal alignment on the Visium skin dataset. Quantitatively, reconstructions reached a cosine similarity of 0.933, an LPIPS of 1.095, and an SSIM of 0.102, demonstrating preservation of transcriptomic structure despite limited visual fidelity.
- Qualitative assessment showed that reconstructed patches captured global tissue texture but appeared noisy at the cellular level. To improve interpretability, outputs were visualized as heatmaps, which highlighted transcriptomic signal capture even when direct patch reconstructions lacked clarity.
- The composite figure illustrates this result, showing a real histology patch alongside its diffusion-based reconstruction. While fine morphology remains limited, the strong embedding alignment confirms that latent diffusion can successfully link transcriptomic profiles with histological features. These findings provide a proof-of-concept foundation for applying generative models to spatial transcriptomics in human skin.



Metric	Value
Cosine Similarity	0.933
LPIPS	1.095
SSIM	0.102

Reconstruction captures transcriptomic signal (high embedding alignment, cosine = 0.933) though visual fidelity remains limited.

METHOD

Dataset: Human skin spatial transcriptomics (10x Genomics Visium). Includes high-resolution histology image, tissue spot coordinates, and filtered gene expression matrix.

Patch-to-spot mapping: Histology tiled into 256 × 256 patches. Each patch linked with overlapping Visium spots and their top 50 expressed genes.

Embedding: CLIP encoders embedded histology and gene vectors into a joint latent space.

Model: Latent Diffusion Model trained on paired embeddings. Reconstructions generated with transcriptomic conditioning.

- Evaluation: Cosine similarity, Learned Perceptual Image Patch Similarity (LPIPS), and Structural Similarity Index (SSIM).

CONCLUSION

This study demonstrates the first application of latent diffusion models to human skin spatial transcriptomics. Results confirm strong embedding alignment between histology and transcriptomic profiles, validating the feasibility of cross-modal learning. While early reconstructions remain noisy, future improvements will include:

- Connected component analysis to isolate real cell clusters.
- Classification of clinically relevant subtypes (keratinocytes, fibroblasts, immune).
- 3D/4D latent visualizations of tissue heterogeneity.
- Together, these advances will position latent diffusion as a powerful framework for integrating molecular and morphological insights in dermatopathology.

REFERENCES

References Link:
https://docs.google.com/document/d/1TYUzN0DfHAQ1mhjpJRLdbc3XAZfpoRC9utN7BOtBu0A/edit?usp=sharing