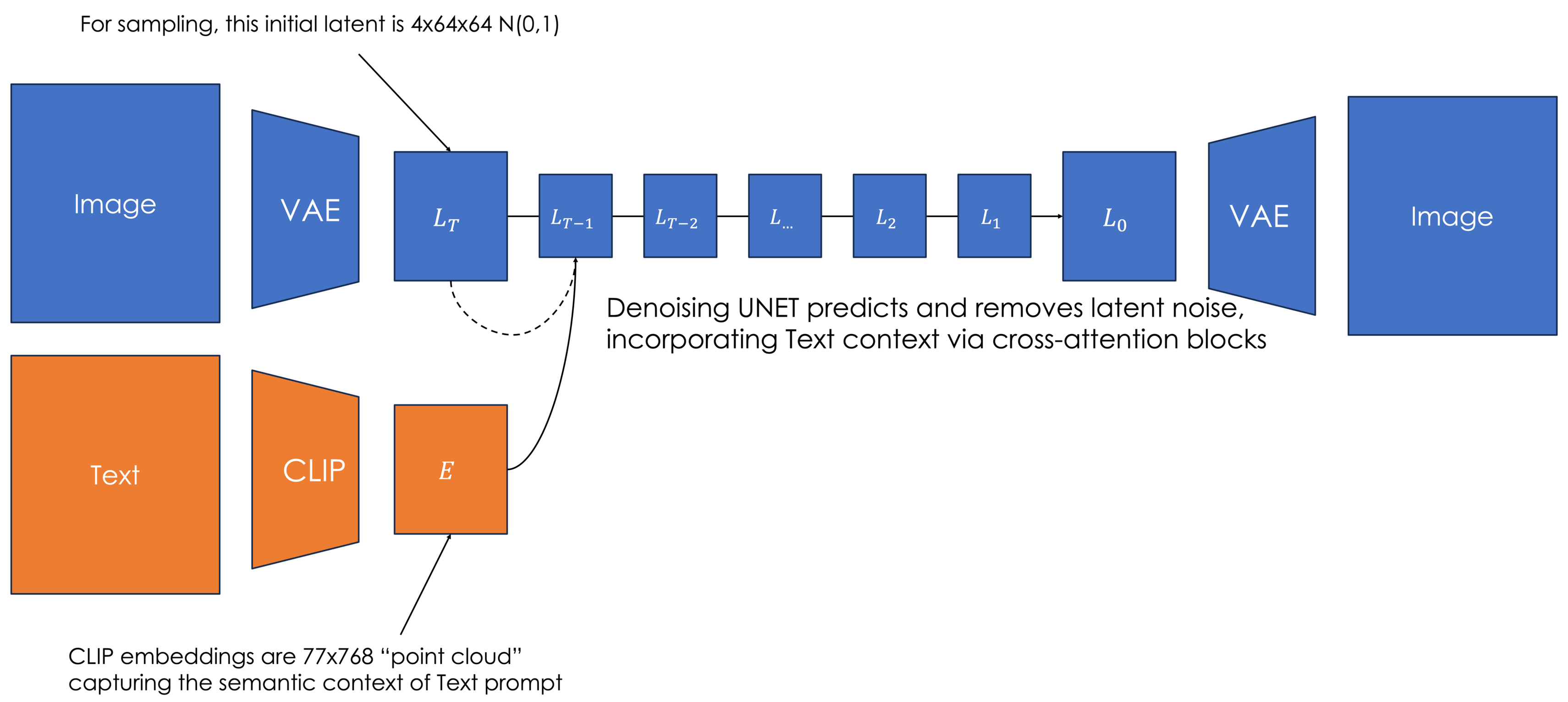
The Role of Embedding Geometry in Image Interpolation for Stable Diffusion

Nicholas Karris*†, Luke Durell*,
Javier Flores*, Tegan Emerson*‡

*Pacific Northwest National Laboratory
†University of California, San Diego
‡University of Texas at El Paso

Paper #9

STABLE DIFFUSION



Schematic of denoising UNET pipeline in Stable Diffusion [1]. The main question of interest is how the resulting image changes as a function of the CLIP embedding, assuming all other components of the pipeline (e.g., the initial noisy latent representation L_T) remain fixed.

CLIP EMBEDDINGS

CLIP takes a text prompt and uses self-attention to generate 77 token vectors. Each token is a vector in $\mathbb{R}^{768}$ and the 77 tokens are packaged into a 77x768 matrix. To interpolate between two prompts, one approach is to linearly interpolate between the embedding matrices.

PERMUTATION INVARIANCE:

The cross-attention blocks in the Stable Diffusion denoising UNET are invariant under permutation of the CLIP token vectors. Cross-attention between two matrices X and $\mathrm{X'}$ is given by

$$\mathrm{A(Q,K,V)} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{D}}}\right)\mathrm{V}$$

where $\mathrm{Q = XW_Q}$ is a query matrix, $\mathrm{K = X'W_K}$ is a key matrix, $\mathrm{V = X'W_V}$ is a value matrix, and D is the dimension of the keys. Since softmax operates row-wise, permuting the columns of K and V does not change the result, which means cross-attention is invariant under permutations of the rows of $\mathrm{X'}$. In Stable Diffusion, $\mathrm{X'}$ corresponds to the CLIP embeddings, and so the output image of the above pipeline is invariant under permutations of the rows of the CLIP matrix. Thus:

CLIP embeddings behave more like "point clouds" and less like "matrices"

COUPLINGS:

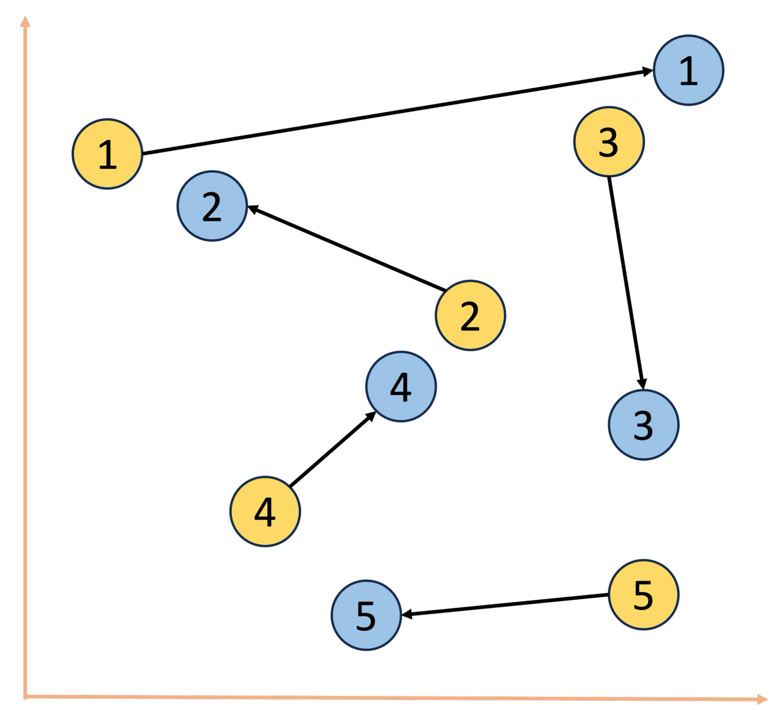
As point clouds, there are many ways to "pair off" tokens before linearly interpolating. The CLIP matrix induces a coupling by pairing off corresponding rows, but other couplings may give more "natural" interpolations.

OPTIMAL TRANSPORT

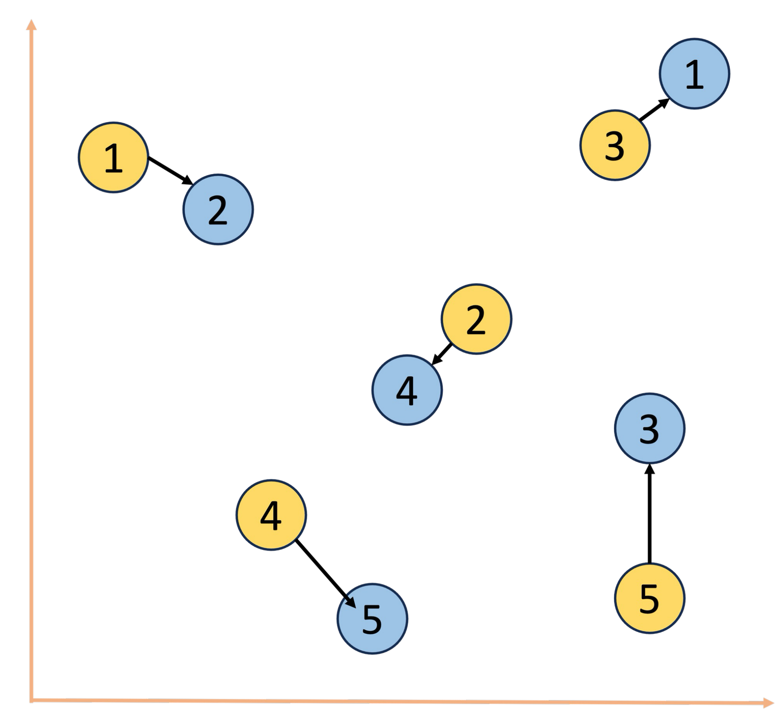
Given two point clouds $\mu = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ and $\nu = \frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}$, i.e., uniform discrete measures supported on the same number of points, we define the Wasserstein distance between μ and ν to be

$$\mathrm{W}_2(\mu,\nu) = \min_{\sigma\in S_N}\left(\sum_{i=1}^{N}\left\|x_i - y_{\sigma(i)}\right\|^2\right)^{\frac{1}{2}}$$

where the optimization is over all permutations $\sigma \in S_N$, the symmetric group on N elements. A map T that satisfies $\mathrm{T}(x_i) = y_{\sigma*(i)}$ for some optimal coupling $\sigma*$ is called an "optimal transport map" and is denoted T_μ^ν. When the support of μ is N *distinct* points, then at least one such optimal transport map exists [2].

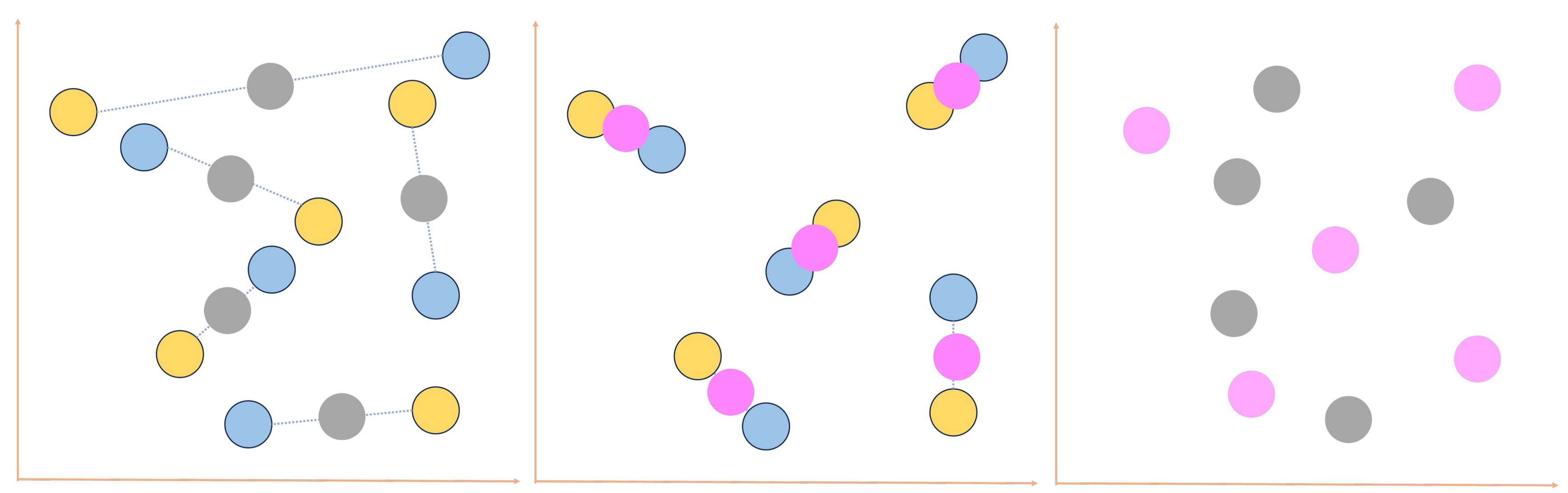


(suboptimal coupling from CLIP)



(optimal transport coupling)

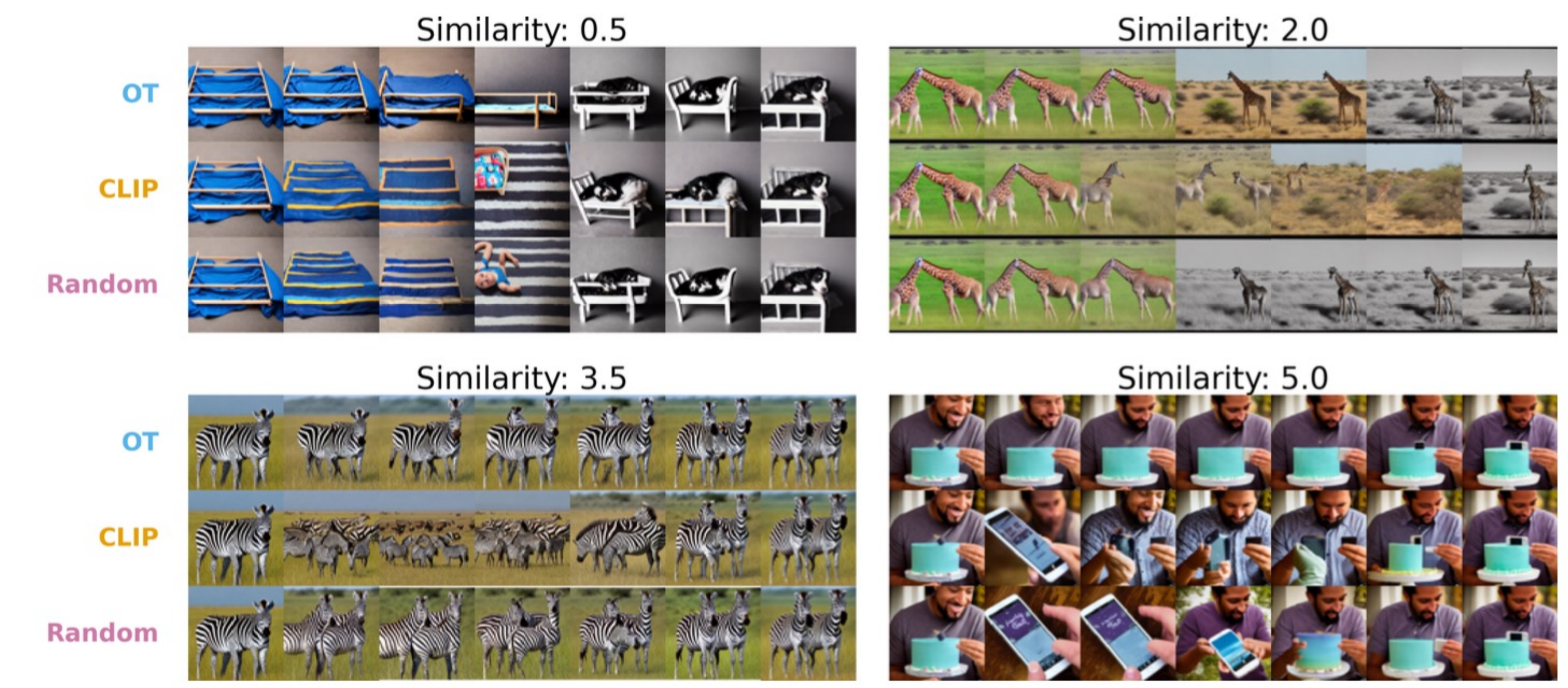
EMBEDDING INTERPOLATIONS



Interpolating via different couplings gives different "intermediate" embeddings. Better couplings result in shorter interpolating "paths" through Wasserstein space – they better preserve the "geometry" of the embedding point clouds.

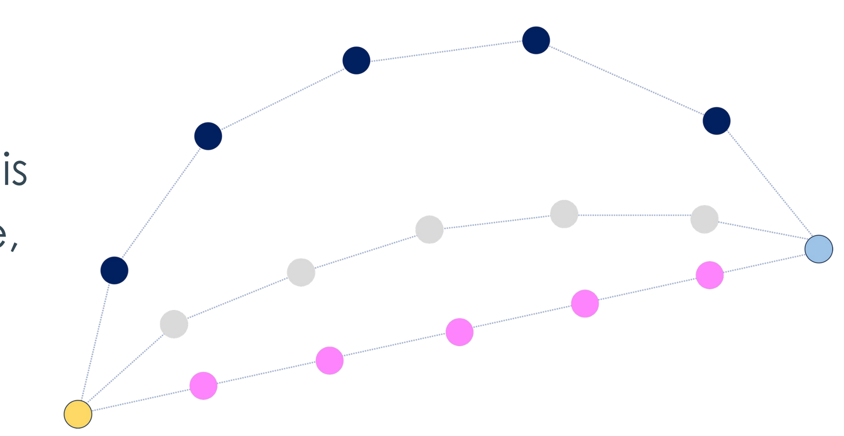
QUESTION: Do shorter paths through Wasserstein space result in "shorter paths" through "image space"?

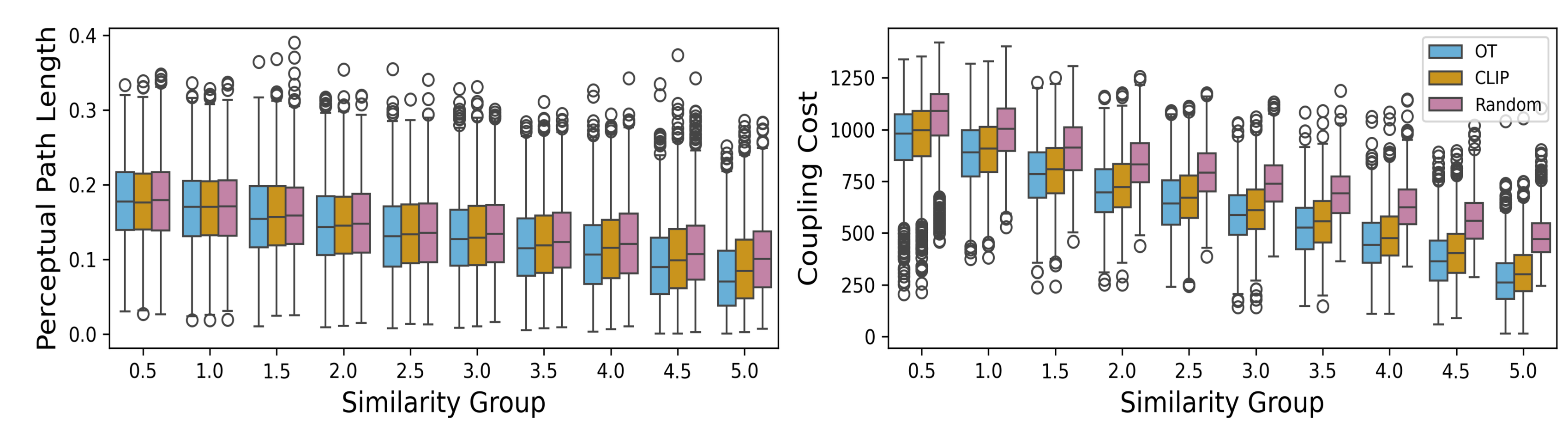
IMAGE INTERPOLATIONS



To quantify the "length" of a "path" through image space, we use Perceptual Path Length [3], which is the average LPIPS [4] score between consecutive images.

Along shorter paths, the average distance between consecutive points is smaller. LPIPS is not actually a distance, but it is a reasonable measure of image similarity.





The optimal transport couplings result in (statistically significant) smoother image interpolations. We see a greater relative improvement from the optimal coupling for more similar prompts. These both suggest that the "correct" interpretation of CLIP embeddings is as point clouds, or distributions.

REFERENCES / ACKNOWLEDGEMENTS

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models.* In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[2] Peyré, G. and Cuturi, M. (2020). Computational Optimal Transport. arXiv:1803.00567

[3] Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[4] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

This work was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy under contract DE-AC05-76RL01830.