Which Way from B to A: The role of embedding geometry in image interpolation for Stable Diffusion

Nicholas Karris [*]
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093
nkarris@ucsd.edu

Luke Durell
Pacific Northwest National Laboratory
Richland, WA 99354
luke.durell@pnnl.gov

Javier E. Flores
Pacific Northwest National Laboratory
Richland, WA 99354
javier.flores@pnnl.gov

Tegan Emerson [†]
Pacific Northwest National Laboratory
Richland, WA 99354
tegan.emerson@pnnl.gov

Abstract

It can be shown that Stable Diffusion has a permutation-invariance property with respect to the rows of Contrastive Language-Image Pretraining (CLIP) embedding matrices. This inspired the novel observation that these embeddings can naturally be interpreted as point clouds in a Wasserstein space rather than as matrices in a Euclidean space. This perspective opens up new possibilities for understanding the geometry of embedding space. For example, when interpolating between embeddings of two distinct prompts, we propose reframing the interpolation problem as an optimal transport problem. By solving this optimal transport problem, we compute a shortest path (or geodesic) between embeddings that captures a more natural and geometrically smooth transition through the embedding space. This results in smoother and more coherent intermediate (interpolated) images when rendered by the Stable Diffusion generative model. We conduct experiments to investigate this effect, comparing the quality of interpolated images produced using optimal transport to those generated by other standard interpolation methods. The novel optimal transport–based approach presented indeed gives smoother image interpolations, suggesting that viewing the embeddings as point clouds (rather than as matrices) better reflects and leverages the geometry of the embedding space.

1 Introduction

The study of the manipulation and interpolation of the inputs to and outputs of generative diffusion-based text-to-image models like Stable Diffusion and latent diffusion models (Rombach et al., 2022; Esser et al., 2024) has attracted increased attention in recent years. These models produce novel images by denoising a random noise latent conditioned on a prompt input (Ho et al., 2020). The prompt embeddings obtained from the raw prompt text, such as those obtained by CLIP (Radford et al., 2021), are the input observed by the model and used for training and sampling. Understanding these prompt embedding spaces is desirable for a variety of creative and functional applications, including prompt optimization (Gal et al., 2022; Wang et al., 2024; Zhu et al., 2007), improved sampling diversity (Deckers et al., 2024), image and prompt inversion (Zhang et al., 2024b; Li et al.,

[*] Work conducted during a summer internship at Pacific Northwest National Laboratory.

[†] Dr. Emerson holds a joint appointment in the Department of Mathematical Sciences at the University of Texas El Paso.

1st Conference on Topology, Algebra, and Geometry in Data Science (TAG-DS 2025).

2025a), concept ablation and unlearning (Li et al., 2023; Kumari et al., 2023), and image interpolation (Wang and Golland, 2023).

Specifically, interest in image interpolation in the context of diffusion models has grown. Applications include anticipated domains such as video transitions (Zhang et al., 2025), but also extend to a variety of unexpected use-cases, such as improving echocardiography image quality (Sivaanpu et al., 2024) or classifying plant health (Lee, 2025). Methods to obtain smooth image interpolations in the context of generative diffusion models include latent-space interpolation (Yu et al., 2025; Saito and Matsubara, 2025) a mixture of latent and prompt embedding interpolation (Wang and Golland, 2023; Yang et al., 2023; Zhang et al., 2024a), and even attention interpolation (He et al., 2024). Most of these approaches (excepting He et al. (2024)) focus on starting with real images, inverting to the embedding space, interpolating between embeddings, and then generating images for the interpolated embeddings.

In parallel, there has been a growing interest in the geometry of learned latent spaces for frontier models like the work of Balestriero et al. (2025); Lee (2023); Arvanitidis et al. (2017); Sakamoto et al. (2024), and even for movement and interpolation between images in Park et al. (2023). Geometric perspectives are being leveraged extensively in the the interpretability literature Voynov and Babenko (2020); Balestriero and Baraniuk (2020). Additionally, there is research in the broader community looking at how geometric properties are connected to topics like hallucination Yeats et al. (2025); Phillips et al. (2025) and deep-fake Xie et al. (2025); Barnabò et al. (2023); Sivabalamurugan and Swapna (2024) detection. Much of this work focuses on clustering, structure, and dominant modes or directions within the learned embedding space where the representations of the data are vectors or matrices.

Rather than focus on the geometry of *latent* space as in, e.g., Park et al. (2023), we look to connect geometric perspectives on the textual *embedding* space with the image interpolation task in Stable Diffusion. This novel work leverages a permutation invariance property of the embedding-to-image generator used in Stable Diffusion to produce a new way of exploiting the geometry of embeddings in image interpolation. Rather than understanding relationships between prompts based on Euclidean or matrix-based relationships, we consider learned embeddings as unordered point-clouds. This point-cloud provides a different way to interpolate between embeddings using optimal transport. While optimal transport techniques have been incorporated into diffusion frameworks for concept optimization (Li et al., 2025b), as well as interpolation methods (Zhu et al., 2007; Yang et al., 2023), to the best of our knowledge, no studies explore the effect of using optimal transport to pair prompt embedding tokens for interpolation within a diffusion framework.

In our work, we compare the performance gained by using optimal transport for interpolation between prompt embeddings from a point-cloud view. Although there is evidence that interpolating only between prompt embeddings results in sub-par interpolations (He et al., 2024), by fixing the latent seed and focusing only on prompt embedding interpolation, we are able to isolate properties and features of the prompt embedding space which otherwise might be masked by interpolating the latents and the prompts simultaneously. Beyond simply improving interpolation methods, this research provides a potential mechanism to improve foundational understanding of the prompt embedding space and emergent behavior.

In Section 2 we present the diffusion and embedding models used, underlying cross attention permutation invariance assumptions required for the application of optimal transport, and the optimal transport methodology. Our novel approach and experimental hypotheses are shared in Section 3. Section 4 establishes our experimental design and presents our results of probing the effect of optimal transport for embedding interpolation. A discussion of the results and potential future work are provided in Section 5.

2 Preliminaries and theoretical motivation

In this section we present a high-level overview of relevant concepts. We begin with diffusion models, stable diffusion, and Contrastive Language-Image Pretraining (CLIP) in Section 2.1 followed by discussions of CLIP permutation invariance and optimal transport in Sections 2.2 and 2.3, respectively.

2.1 Diffusion models, stable diffusion, and CLIP

Diffusion models are a class of probabilistic generative models that have garnered significant attention for their ability to produce high-quality images that closely adhere to user-specified criteria, unlocking applications that range from automating artistic content creation to aiding researchers in synthetic data generation, design and modeling complex systems and processes (Corso et al., 2023; Mazé and Ahmed, 2022). At their core, diffusion models operate through two complementary processes: (*i*) a forward "diffusion" process where an original image is corrupted through the iterative addition of Gaussian noise, and (*ii*) a reverse-diffusion process that reconstructs the original image by estimating and removing the noise added at each diffusion step. More formally, in the diffusion process, the data $\mathbf{x}_0$ are perturbed at each step t through the Markov chain

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

where $\beta_t \in (0, 1)$ is a variance schedule determining the amount of noise added at step t, and $\mathcal{N}(\cdot; \mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. The reverse-diffusion process to recover $\mathbf{x}_0$ through $\mathbf{x}_T$ is modeled by a separate Markov chain

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(t)),$$

where $\mu_\theta(\mathbf{x}_t, t)$ is the predicted mean (typically learned using a neural network (U-Net)), and $\Sigma_\theta(t)$ is the variance (often fixed, but can be learned).

While several variations of diffusion models have been developed (Cao et al., 2023; Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022; Song et al., 2021, 2022), our interest lies in the widely studied *latent* diffusion model Stable Diffusion 1.5 (SD) by Stability AI (Rombach et al., 2022). Like traditional diffusion models, SD employs a forward and reverse diffusion process, but for SD, these processes operate in a latent space rather than the original data space. A pre-trained variational autoencoder (VAE) first maps input data $\mathbf{x}_0$ to a lower-dimensional latent representation $\mathbf{z}_0$ via an encoder E, which substantially improves computational efficiency of the diffusion/reverse-diffusion processes since the latent space is much lower dimensional than that of the original data. The reconstructed representation obtained through the latent reverse-diffusion process is then decoded back into the data space through the VAE decoder D.

The U-Net architecture of the reverse-diffusion process in SD differs from traditional diffusion in that it has been enhanced with attention mechanisms, and in particular cross-attention between the latent space and a textual embedding space. SD uses the Contrastive Language-Image Pretraining (CLIP) model developed by OpenAI Radford et al. (2021), which maps images and text strings to a shared embedding space in such a way that paired texts and images are nearby and unrelated pairs are farther apart. Thus, the Stable Diffusion 1.5 pipeline for producing an image corresponding to a particular prompt string is to compute a CLIP embedding of the string (padded if necessary), which yields a collection of 77 tokens vectors in $\mathbb{R}^{768}$, canonically packaged as a matrix in $\mathbb{R}^{77\times 768}$. This is the textual embedding space that is used in the cross-attention blocks in the U-Net architecture for image generation.

In summary, the Stable Diffusion pipeline has two essential "functions" around which we will center our focus: (*i*) the textual prompt embedding via CLIP and (*ii*) a reverse-diffusion processes wherein embeddings are denoised and decoded into an image influenced by the CLIP-encoded text. These two functions can be described as

- $f : \{\text{strings}\} \to \mathbb{R}^{77\times 768}$, where $s \mapsto f(s)$, the CLIP embedding of the prompt s,
- $g : \mathbb{R}^{77\times 768} \to \{\text{images}\}$, where $e \mapsto g(e)$, the image generated from embedding e

Note that the image generation function g is hiding a tremendous amount of computational complexity. In particular, it contains the entire U-Net and cross-attention architecture that underlies the "denoising" process, which involves many more initial parameters than just the CLIP embedding matrix. Throughout this paper, we will consider all other parameters of this architecture fixed (e.g., the initial "noisy" latent z_T), meaning we are only interested in how the image that results from the overall pipeline (i.e., the image that results from decoding the final denoised latent) changes as a function of the CLIP embedding input. Thus, while the underlying architecture is extremely intricate, it is helpful to consider it as a single function g that takes as input a 77×768 embedding matrix and returns an image.

2.2 Permutation invariance

We briefly detail two observations that give rise to an important permutation-invariance property of the function g above.

The first is a general property about the ubiquitous attention operation. Given two matrices of arbitrary dimension, X and $X^{'}$, a common formulation of the attention operation (Vaswani et al., 2017) is

$$A(Q, K, V) = \text{softmax}_{row}\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{1}$$

where the input matrices are a query ($Q = XW_Q$), key ($K = X^{'}W_K$), and value ($V = X^{'}W_V$) matrix constructed as the product of X or $X^{'}$ and a matrix of learned weights. Equation (1) describes self-attention when $X = X'$ and cross-attention when $X \neq X'$. It is a known property that cross-attention is invariant under joint permutation of the rows of K and V, and hence also of X'. At a high level, this is because softmax operates row-wise, and permuting the columns simply permutes the probabilities in the same way; for more detail, we defer to (Fleuret, 2021; Ji et al., 2019).

The second critical observation is that the U-Net backbone of Stable Diffusion 1.5 is constructed as a series of self-attention blocks on the image latent and cross-attention blocks between the image latent and the CLIP embedding matrix (Rombach et al., 2022). In the cross-attention blocks, the image latent z corresponds to X and the CLIP embedding e corresponds to $X^{'}$. That is, the function g defined in 2.1, as a function of the CLIP embeddings e, is nothing but a series of cross-attention blocks with $X' = e$. Together with the fact that cross-attention is permutation-invariant on the rows of X', we conclude that the function g is permutation-invariant on the rows of e.

This means that the *order* of the 77 token vectors is not relevant, only the tokens themselves as points in $\mathbb{R}^{768}$, which suggests that one can view these embeddings not as *matrices* but instead as *point clouds*. That is, we can view g as a function $g : \mathcal{P}_{77}^{\text{unif}}(\mathbb{R}^{768}) \to \{\text{images}\}$, where $\mathcal{P}_{N}^{\text{unif}}(\mathbb{R}^{d})$ denotes the set of uniform measures on N (possibly nondistinct) discrete points in $\mathbb{R}^d$ (i.e., measures of the form $\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$, where $x_i \in \mathbb{R}^d$). For notational clarity, when viewing embeddings as matrices, we will denote them with Roman letters (e.g., e_0 or e_1), and when viewing them as point clouds, we will use Greek letters (e.g., μ_0 or μ_1). Note that there is a many-to-one equivalence between matrices and point clouds, where a matrix e and point cloud μ are equivalent if the rows of e are exactly the points in μ.

2.3 Optimal transport and Wasserstein space

This new point-cloud interpretation of the embedding space comes with a new natural distance metric: the Wasserstein distance. For two point clouds $\mu = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ and $\nu = \frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}$ in $\mathcal{P}_{N}^{\text{unif}}(\mathbb{R}^{d})$, we define the Wasserstein distance between μ and ν to be

$$W_2(\mu, \nu) := \min_{\sigma \in S_N}\left(\sum_{i=1}^{N}\|x_i - y_{\sigma(i)}\|_2^2\right)^{1/2}, \tag{2}$$

where the optimization is over all permutations $\sigma \in S_N$, the symmetric group on N elements.[3] A map T that satisfies $T(x_i) = y_{\sigma^*(i)}$ for some optimal permutation σ^* is called an "optimal transport map" from μ to ν and is denoted T_μ^ν. When the support of μ is N *distinct* points (which in practice is almost always true for our CLIP embeddings), then at least one such optimal transport map exists (Proposition 2.1 in Peyré and Cuturi (2020)). Though not necessarily unique, the notation T_μ^ν will refer to a particular choice of an optimal map, which we will call "the" optimal transport map.

While we lose the Euclidean space structure that comes with the matrix perspective, the point-cloud perspective comes with an analogous mathematical structure of a (formal) Riemannian manifold

[3] In general, one typically needs to define the Wasserstein distance between discrete measures using the Kantorovich formulation, which allows for the possibility that the optimal "plan" may not be induced by a "map". However, Proposition 2.1 in Peyré and Cuturi (2020) guarantees that, in the case of uniform discrete measures on the same number of points, there exists a permutation that is optimal. This is the only case of interest in this paper, so we make the definition of Wasserstein distance using permutations, rather than more general plans.

(Ambrosio et al., 2008; Otto, 2001) Specifically, W_2 is a metric on $\mathcal{P}_N^{\text{unif}}(\mathbb{R}^d)$, and $\mathcal{P}_N^{\text{unif}}(\mathbb{R}^d)$ is a subset of the "Wasserstein manifold" of measures with finite second moment, which itself comes with some useful geometric properties. As we begin to explore the landscape of embedding space with this new perspective, the key geometric property of interest to this paper is the nice relationship between barycenters and geodesics. Specifically, if $T_{\mu_0}^{\mu_1}$ is an optimal transport map between $\mu_0, \mu_1 \in \mathcal{P}_N^{\text{unif}}(\mathbb{R}^d)$, then there is an explicit expression for a "constant-speed geodesic" $\mu : [0,1] \to \mathcal{P}_N^{\text{unif}}(\mathbb{R}^d)$ from μ_0 to μ_1 given by $t \mapsto \mu_t$ with

$$\mu_t = ((1-t)\,\mathrm{id} + tT_{\mu_0}^{\mu_1})_{\#}\mu_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{(1-t)x_i + ty_{\sigma^*(i)}}, \tag{3}$$

where the notation $T_{\#}\mu$ denotes the pushforward of μ by the map $T : \mathbb{R}^d \to \mathbb{R}^d$ (Ambrosio et al., 2008). In particular, for $t \in [0,1]$, μ_t is the weighted Wasserstein barycenter; i.e., μ_t satisfies

$$\mu_t = \underset{\mu \in \mathcal{P}_N^{\text{unif}}(\mathbb{R}^d)}{\operatorname{argmin}} \left[(1-t)W_2(\mu_0, \mu)^2 + tW_2(\mu_1, \mu)^2\right]. \tag{4}$$

This barycenter definition is exactly analogous to a "weighted average" in Euclidean space, which means these geodesics between point clouds on the Wasserstein manifold are exactly analogous to straight lines — they trace out the shortest path between two points in the space. Thus, in essence, this says that the OT-inspired way to interpolate two N-point clouds μ_0, μ_1 is to pair off points in μ_0 with points in μ_1 using the optimal transport "coupling" (i.e., the pairing of x_i with $T_{\mu_0}^{\mu_1}(x_i) = y_{\sigma^*(i)}$) and then to linearly interpolate between each pair simultaneously. Despite losing the nice structure of the vector space of matrices, the point-cloud interpretation of the embedding space still exhibits a natural way to interpolate between embeddings. Exploiting the point-cloud interpretation of the learned embedding space provides the basis for our novel, geometrically-informed image interpolation technique presented in Section 3

3 Geometry-informed Image Interpolation Approach

In this work we explore how the permutation-invariance property described in Section 2.2 can inform novel image interpolation techniques by operating on the embeddings. Specifically, the discussion above suggests that the natural way to interpolate between embeddings (viewed as point clouds) is to use (3) with the optimal coupling σ^*. This gives the shortest path through the embedding space between the two prompts, meaning that any other method of interpolating the embeddings gives a longer (or at least not shorter) path through Wasserstein space. For example, there is a natural way to interpolate between embeddings e_0, e_1 viewed as matrices, which is to linearly interpolate by setting $e_t = (1-t)e_0 + te_1$. By viewing e_t now as a point cloud, this method traces out a path through Wasserstein space also described by (3), except instead of using the optimal coupling σ^*, it uses the coupling induced by the order of the rows in the CLIP matrices, which we call the "CLIP coupling." In fact, the length of this path is exactly the cost of the associated coupling (consequence of Theorem 8.3.1 in Ambrosio et al. (2008)), and for the CLIP coupling, that cost is the standard 2-norm of the difference of corresponding rows. Thus, not only do we know that the CLIP interpolating path is longer (since the optimal coupling cost is at least as small as the CLIP coupling cost), we know precisely how much longer.

One way to assess whether our point-cloud perspective on embedding space is "more natural" than the matrix perspective is to investigate how these path lengths through embedding space relate to notions of similarity in image space. Specifically, for a particular interpolation path through embedding space, there is an associated path through image space obtained by generating the image that corresponds to each interpolated embedding along the embedding path. For any embedding-interpolation path, the start and end embeddings are fixed, and this means that the associated image path connects the images corresponding to the start and end embeddings. If one embedding-interpolation path is "better" than another, its associated image-interpolation path should be smoother, more natural-looking, and contain images which are more similar. In short, the path through image space should be "shorter." Thus, if considering the CLIP embeddings as point clouds really is a more faithful geometric interpretation of embedding space, we hypothesize that:

1. The geodesic path through embedding space is shorter for prompts which are more similar (i.e., embeddings are closer in Wasserstein distance when the prompts are more similar), and the associated image paths have lower PPL scores,

2. suboptimal couplings give relatively worse embedding interpolations for prompts which are more similar (because suboptimal couplings are relatively more costly when embeddings are close in Wasserstein distance)
3. for a fixed pair of prompt embeddings, a shorter interpolating path between them gives a better associated image interpolation, and
4. this effect increases for prompts which are more similar because the relative path-length increase is less severe.

To properly test these hypotheses, we need a suitable notion of path length in image space. Rather than use a pixel-wise metric, which promotes structurally smooth but unrealistic image interpolations, we seek an image-similarity score that promotes smooth image transitions while retaining realism and context shifts. To this end, we use Perceptual Path Length (PPL) Karras et al. (2019) as a surrogate for path length. PPL is defined as the average of image similarity scores between consecutive equally-spaced images along a path, with the intuition that equally spaced points are closer if the overall path is shorter. The standard image similarity score used for PPL is Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018), which is a standard method for judging image similarity.

Thus, our high-level approach is as follows, as illustrated in Figure 1. Two embedding matrices are obtained from two prompt pairs using CLIP. The embedding matrices are treated as point clouds, and three interpolation couplings (OT, CLIP, and a random coupling) are used to define a path between the embedding point clouds. The interpolated path in embedding space is sampled along a grid, and resultant embedding point clouds are obtained. Both the original embeddings and the interpolated embeddings are used to produce a corresponding trajectory of images in image space. For each grid sample along the interpolated path in embedding space, we produce a corresponding image with a diffusion model (Stable Diffusion), and the resultant images are compared using PPL, based on the LPIPS scores between the k^{th} and $(k+1)^{th}$ image denoted ℓ_k. In Section 4, we present experimental results that use this pipeline to explore the validity of the hypotheses listed above.

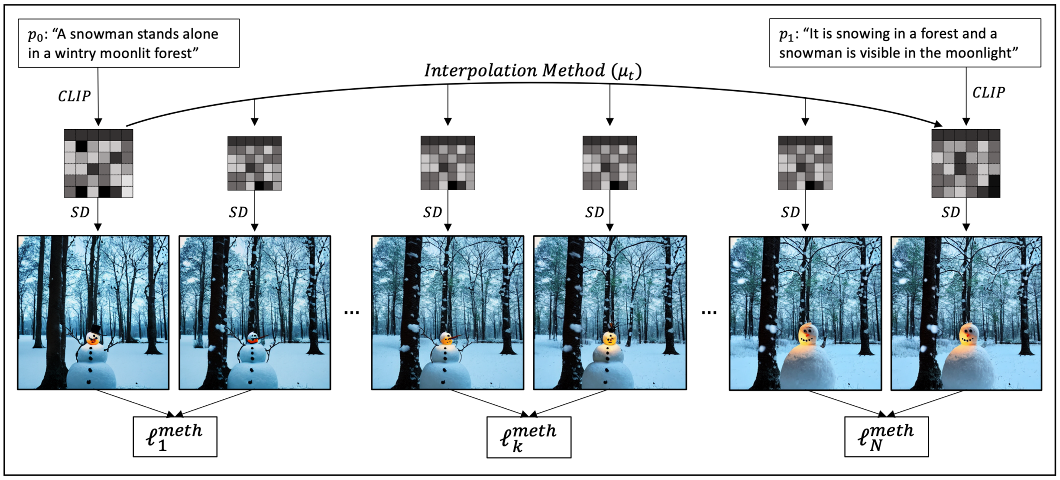
p_0: "A snowman stands alone in a wintry moonlit forest"
p_1: "It is snowing in a forest and a snowman is visible in the moonlight"
Interpolation Method (μ_t)
CLIP
CLIP
SD
SD
SD
SD
SD
SD
...
...
ℓ_1^{meth}
ℓ_k^{meth}
ℓ_N^{meth}

Figure 1: Embedding interpolation workflow where SD indicates use of the Stable Diffusion model to produce the associated image.

4 Experimental Design and Results

4.1 Experimental setup

To evaluate the impact of capturing the geometric structure of CLIP embeddings in image interpolation, we consider many pairs of prompts (see below for more details about how the prompt pairs were selected). For each pair, we follow the pipeline illustrated in Figure 1 for each of three couplings — OT, CLIP, and a random coupling (to create the random coupling, we fix the first row of the CLIP matrices — this row is always the same for any prompt — and randomly pair the last 76 rows in each CLIP matrix). We compute the cost of each coupling and the resulting PPL score for each coupling's image interpolation (we use the AlexNet architecture when computing LPIPS scores for PPL). We note that LPIPS gives lower scores to images that are more similar, and so a smaller PPL score means that the given path through image space is “shorter”. Since the OT interpolation method produces a geodesic μ_t^{OT} through embedding space, and each of the other two interpolations μ_t^{clip} and μ_t^{rand} are longer paths, we hypothesize that the LPIPS/PPL scores for the OT method ℓ_k^{ot} will be lower on average than the other methods.

4.2 Dataset

To properly test the efficacy of the three interpolation methods, we must produce interpolated images from each method for many pairs. The structure of our hypothesis requires a dataset with prompt pairs that carry a known similarity score. To create such a dataset, we leverage the Crisscrossed Captions dataset (Parekh et al., 2021), which contains captioned images from the the MS-COCO dataset (Lin et al., 2015) that have been manually scored according to their similarity on a scale from 0 to 5, with 5 being most similar. We assume that the similarity score between two images can be extended as a similarity score the associated pair of captions. We discretize the range of similarity scores into bins of width 0.5 and randomly select 1,000 pairs from each similarity group (excluding any pairs where the two captions were identical). These captions are then used as prompts in the SD model with a fixed seed. The results shown reflect the experiment described in Section 4.1 applied to 10,000 caption/prompt pairs. The Crisscrossed Captions dataset has a custom, open source license at https://github.com/google-research-datasets/Crisscrossed-Captions/blob/master/LICENSE. The MS-COCO dataset has a Creative Commons Attribution 4.0 License, and the SD model has a CreativeML Open RAIL-M License. CPU and GPU workers were used on an internal cluster and the total compute cost for the experiments was $1,094$ CPU hours and $2,172$ GPU hours. Preliminary and failed experiments accounted for an additional approximately $7,000$ CPU hours and 575 GPU hours of compute resources.

4.3 Results

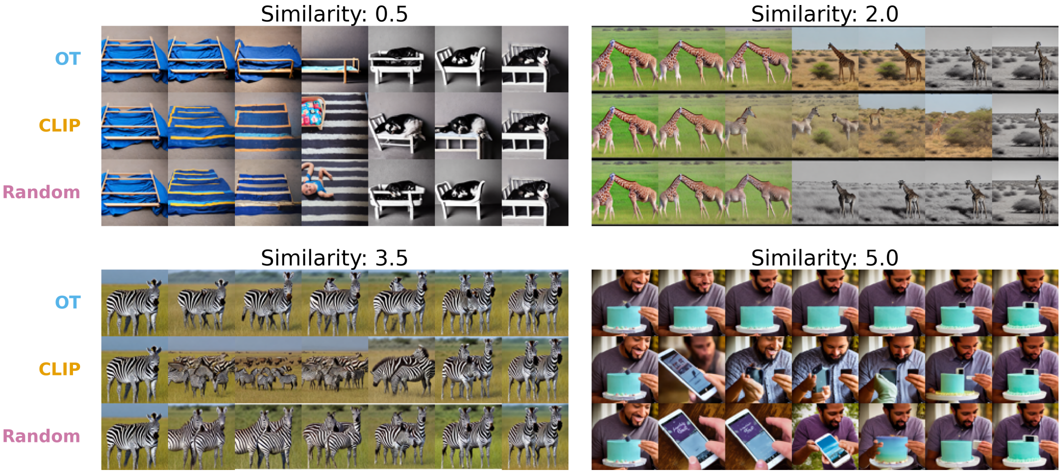

Figure 2: Image interpolations for each method across four selected prompt pairs of varying similarity.

To help build the reader’s intuition, we start by showing the qualitative impact of the geometric relationships being considered. Figure 2 shows a selection of interpolated images, where for each of the four panels, the sub-images on the far left and far right correspond to the images produced for each prompt pair, i.e., the endpoints of our path in image space. For each panel, the rows of sub-figures are ordered from top to bottom as OT, CLIP, and random. Examination of the image trajectories shows cases where the CLIP and/or random method appears to hallucinate objects in the interpolated trajectory. The prompt pairs shown were selected based on having quantitatively larger differences in the associated path lengths for illustration purposes. However, as will be shown subsequently, the OT path is found to perform comparably or above the CLIP or random couplings across the board.

For a quantitative evaluation, we compute the perceptual path length (PPL) (as defined above in Section 4.1) (Karras et al., 2019) for each interpolation method (OT, CLIP, random) as a measurement of transition smoothness between the trajectory of interpolated images for all 10,000 pairs of prompts. Smaller PPL values are desirable and correspond to higher smoothness across the interpolated image trajectories. In the subsequent plots, we report PPL and coupling costs over the collection of all 1000 prompt pairs of each similarity level.

The boxplots of PPL scores and coupling costs are plotted for each method and similarity group in Figure 3. To test our hypothesis, we assess the statistical significance of both the difference in median PPL and the difference in median coupling cost between the OT interpolation results and the CLIP or random couplings, respectively within each similarity group. Significance of p-values for testing the difference in median PPL within each interpolation method and similarity group are reported in Table 1. As hypothesized, the impact of embedding path is more significant for more similar pairs of prompts. All tests for the difference in median coupling cost between optimal transport and each other interpolation method are highly significantly different, $p < 0.0001$, for all similarity groups. Additionally, the path length decreases as a function of similarity and the significance of an optimal coupling grows as the similarity increases as hypothesized based on geometric intuition. The PPL scores and coupling costs were determined to be not normally distributed, so for both sets of tests we employed a Wilcoxon signed-rank test to test the null hypothesis that the median of the difference is zero. In addition to the typical assumptions of independence and randomness, the Wilcoxon signed-rank test assumes continuous data distributed symmetrically about the median (Ramachandran and Tsokos, 2021).

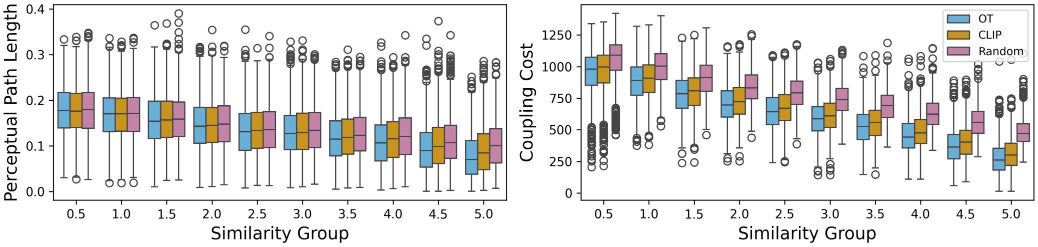

Figure 3: PPL (left) and coupling cost (right) by interpolation method and similarity group.

Table 1: Significance of p-values for the Wilcoxon test comparing the median of the PPL scores. ($p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$)

	Similarity Group									
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
OT vs. CLIP PPL	-	-	-	-	∗∗∗	∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗
OT vs. Random PPL	∗	-	∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗

5 Conclusions and Future Directions

In this work we observe that interpolating using the OT coupling in general results in shorter image paths than using the linear and random couplings, and that the improvement is more substantial for more similar prompts. Both of these observations are consistent with the hypothesis that path length through Wasserstein space is reflected by "path length" through image space. We also notice that in many cases where OT drastically outperforms the other two methods, it does so because the other methods produce interpolated images which are relatively very different than either end image. This suggests that embedding space has a "convexity" property with respect to Wasserstein distance that it does not have when the embedding matrices are handled as matrix objects. Concretely, we find given two embeddings of a particular class, an embedding "between" them in Wasserstein space is more likely to be in the same class than an embedding "between" them in matrix space. Our experimental results suggest that the point-cloud perspective indeed does a better job of capturing the "geometry" of embedding space in ways that reflect more desirable properties in image interpolation.

In particular, it is perhaps surprising that the random couplings, while noticeably worse than the optimal couplings, still result in reasonably good image interpolations. Additionally, anecdotal experiments suggest that even maximally-costly couplings (i.e., couplings that maximize the objective quantity in (2)) still give image interpolations which are not catastrophic. In terms of the geometry of the Wasserstein manifold, these interpolating paths using different couplings are all “straight” (even though they are not all “geodesics”). This suggests some redundancy in the embeddings’ geometry and raises a natural question: How far can we push the interpolating paths before they result in catastrophic image behavior? Investigating this more general question could help shed even more light on the geometry of embedding-space and how it interacts with the “geometry” of image-space.

There are also more general questions one could ask about interpolating between more than two embeddings. The notion of a Wasserstein barycenter is perfectly well-defined when there are more than two measures (the definition is exactly analogous to (4)), which means one could compute the Wasserstein barycenter of many embeddings and observe the corresponding image outputs. Investigating these multi-way interpolations would provide much deeper insight into the “manifold” structure of the embedding-space, which is difficult to fully investigate with only one-parameter curves. While computing Wasserstein barycenters of multiple measures is substantially more expensive than computing an optimal transport map (there is no closed-form equation like (3)), one could use the techniques of linearized optimal transport Linwu et al. (2025); Moosmüller and Cloninger (2021); Cloninger et al. (2023) to approximate the Wasserstein barycenter much more computationally efficiently.

Deeper exploration of these questions could uncover properties and conditions under which prompt interpolation results in thematically consistent and smooth images, giving rise not only to improved methods for image interpolation that rely on prompt interpolation (Wang and Golland, 2023), but also informing scenarios where prompt interpolation or manipulation is an auxiliary step, such as variant refinement (Deckers et al., 2024). Another area for future work is in developing additional metrics to capture pair-wise changes in images. The path length surrogate considered in this work leverages a standard image similarity score, LPIPS, but that score does not explicitly capture or penalize for contextual and content shifts between image pairs. Additional work in this area would enable more direct connections with applications like hallucination detection and associated mitigation strategies.

Finally, the scope of this work focused solely on Stable Diffusion 1.5, whose architecture differs from the latest Stable Diffusion model (3.5) available at the time of writing. While architectural innovations in the latest models present barriers to direct extension of the proposed approach, the point cloud perspective may still better preserve underlying geometry and desired invariances than performing operations like token concatenations used in Stable Diffusion 3.5. Therefore, an important future direction would be extending and evaluating the point cloud framing for shared, multi-modal token spaces and transformer-based denoising architectures like those of current state-of-the-art models. However, although it is not the current “state-of-the-art” model, the fact that it can be run locally means there is a substantial benefit to investigating Stable Diffusion 1.5 specifically. A better understanding of the geometry underlying this model could provide insights that could be thoroughly explored without the need for heavy resources, thus providing a theoretical framework for a large class of geometric experiments that can still be run on an accessible testbed.

Acknowledgments and Disclosure of Funding

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.

References

Ambrosio, L., Gigli, N., and Savare, G. (2008). *Gradient Flows*. Birkhäuser, Basel.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. (2017). Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*.

Balestriero, R. and Baraniuk, R. G. (2020). Mad max: Affine spline insights into deep learning. *Proceedings of the IEEE*, 109(5):704–727.

Balestriero, R., Humayun, A. I., and Baraniuk, R. G. (2025). On the geometry of deep learning. *NOTICES OF THE AMERICAN MATHEMATICAL SOCIETY*, 72(4).

Barnabò, G., Siciliano, F., Castillo, C., Leonardi, S., Nakov, P., Da San Martino, G., and Silvestri, F. (2023). Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33:100244.

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., and Li, S. Z. (2023). A Survey on Generative Diffusion Model. arXiv:2209.02646 [cs].

Cloninger, A., Hamm, K., Khurana, V., and Moosmüller, C. (2023). Linearized Wasserstein dimensionality reduction with approximation guarantees. arXiv:2302.07373 [cs].

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. (2023). DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv:2210.01776 [q-bio].

Deckers, N., Peters, J., and Potthast, M. (2024). Manipulating Embeddings of Stable Diffusion Prompts. volume 8, pages 7636–7644. ISSN: 1045-0823.

Dhariwal, P. and Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs].

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., and Rombach, R. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 12606–12633, Vienna, Austria. JMLR.org.

Fleuret, F. (2021). Deep Learning Attention Mechanisms. Technical report, University of Geneva, Geneva, Switzerland.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-or, D. (2022). An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion.

He, Q., Wang, J., Liu, Z., and Yao, A. (2024). AID: Attention Interpolation of Text-to-Image Diffusion. *Advances in Neural Information Processing Systems*, 37:97766–97799.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Ji, S., Xie, Y., and Gao, H. (2019). A Mathematical View of Attention Models in Deep Learning. Technical report, Texas A&M University, College Station, TX, USA.

Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:1812.04948 [cs].

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. (2023). Ablating Concepts in Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22634–22645. ISSN: 2380-7504.

Lee, M. (2023). The geometry of feature space in deep learning models: A holistic perspective and comprehensive review. *Mathematics*, 11(10):2375.

Lee, Y. (2025). Enhancing plant health classification via diffusion model-based data augmentation. *Multimedia Systems*, 31(2):143.

Li, M., Zhang, G., Wang, Z., Ma, S., Pan, S., Cartwright, R., and Zhai, J. (2025a). EDITOR: Effective and Interpretable Prompt Inversion for Text-to-Image Diffusion Models. arXiv:2506.03067 [cs] version: 1.

Li, S., Wang, Z., Luo, Z., and Lei, N. (2025b). An optimal transport-guided diffusion framework with mitigating mode mixture. *Neurocomputing*, 616:128910.

Li, S., Weijer, J. v. d., Hu, T., Khan, F., Hou, Q., Wang, Y., and Yang, J. (2023). Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Image Diffusion Models.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs].

Linwu, J., Khurana, V., Karris, N., and Cloninger, A. (2025). Linearized Optimal Transport pyLOT Library: A Toolkit for Machine Learning on Point Clouds. arXiv:2502.03439 [stat].

Mazé, F. and Ahmed, F. (2022). Diffusion Models Beat GANs on Topology Optimization. arXiv:2208.09591 [cs].

Moosmüller, C. and Cloninger, A. (2021). Linear Optimal Transport Embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. arXiv:2008.09165 [stat].

Otto, F. (2001). The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174.

Parekh, Z., Baldridge, J., Cer, D., Waters, A., and Yang, Y. (2021). Crisscrossed Captions: Extended Intramodal and Intermodal Semantic Similarity Judgments for MS-COCO. arXiv:2004.15020 [cs].

Park, Y.-H., Kwon, M., Choi, J., Jo, J., and Uh, Y. (2023). Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142.

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

Phillips, E., Wu, S., Molaei, S., Belgrave, D., Thakur, A., and Clifton, D. (2025). Geometric uncertainty for detecting and correcting hallucinations in llms. *arXiv preprint arXiv:2509.13813*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

Ramachandran, K. M. and Tsokos, C. P. (2021). Chapter 12 - Nonparametric Statistics. In Ramachandran, K. M. and Tsokos, C. P., editors, *Mathematical Statistics with Applications in R (Third Edition)*, pages 491–530. Academic Press.

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 (CVPR)*, pages 10674–10685. ISSN: 2575-7075.

Saito, S. and Matsubara, T. (2025). Image Interpolation with Score-based Riemannian Metrics of Diffusion Models. arXiv:2504.20288 [cs].

Sakamoto, K., Sakamoto, R., Tanabe, M., Akagawa, M., Hayashi, Y., Yaguchi, M., Suzuki, M., and Matsuo, Y. (2024). The geometry of diffusion models: Tubular neighbourhoods and singularities. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*.

Sivaanpu, A., Noga, M., Becher, H., Punithakumar, K., and Le, L. H. (2024). Denoising Echocardiography with an Improved Diffusion Model. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. ISSN: 2694-0604.

Sivabalamurugan, M. and Swapna, T. (2024). Deepfake detection and classification using local surface geometrical features. In *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, pages 1–6. IEEE.

Song, J., Meng, C., and Ermon, S. (2022). Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs].

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. arXiv:2011.13456 [cs].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Voynov, A. and Babenko, A. (2020). Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR.

Wang, C. and Golland, P. (2023). Interpolating between Images with Diffusion Models.

Wang, R., Liu, T., Hsieh, C.-J., and Gong, B. (2024). On Discrete Prompt Optimization for Diffusion Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 50992–51011. PMLR. ISSN: 2640-3498.

Xie, H., He, H., Fu, B., and Sanchez, V. (2025). Grdt: Towards robust deepfake detection using geometric representation distribution and texture. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 734–744.

Yang, Z., Yu, Z., Xu, Z., Singh, J., Zhang, J., Campbell, D., Tu, P., and Hartley, R. (2023). IMPUS: Image Morphing with Perceptually-Uniform Sampling Using Diffusion Models.

Yeats, E., Buckheit, J., Scullen, S. M., Kennedy, B., Truong, L., Brown, D., Kay, B., Joslyn, C., Emerson, T., and Henry, M. J. (2025). What do geometric hallucination detection metrics actually measure? In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Yu, Q., Singh, J., Yang, Z., Tu, P. H., Zhang, J., Li, H., Hartley, R., and Campbell, D. (2025). Probability Density Geodesics in Image Diffusion Latent Space. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27989–27998. ISSN: 2575-7075.

Zhang, K., Zhou, Y., Xu, X., Dai, B., and Pan, X. (2024a). DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7912–7921. ISSN: 2575-7075.

Zhang, R., Chen, Y., Liu, Y., Wang, W., Wen, X., and Wang, H. (2025). TVG: A Training-Free Transition Video Generation Method With Diffusion Models. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):7471–7484.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924 [cs].

Zhang, X., Wei, X.-Y., Wu, J., Zhang, T., Zhang, Z., Lei, Z., and Li, Q. (2024b). Compositional inversion for stable diffusion models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, volume 38 of *AAAI'24/IAAI'24/EAAI'24*, pages 7350–7358. AAAI Press.

Zhu, L., Yang, Y., Haker, S., and Tannenbaum, A. (2007). An Image Morphing Technique Based on Optimal Mass Preserving Mapping. *IEEE Transactions on Image Processing*, 16(6):1481–1495.