Similarity Trajectories: Linking Sampling Process to Artifacts in Diffusion-Generated Images

Dennis Menn, Jeff Liang, Hung-Yueh Chiang, Diana Marculescu

The University of Texas at Austin
Chandra Family Department of Electrical and Computer Engineering

Abstract

Artifact detection algorithms are crucial to correcting and improving the quality of data generated by diffusion models. However, because of the variety of artifact forms, these algorithms require substantial annotated data for training, which restricts their scalability and efficiency. This paper discovers a relationship between the similarity of denoised images during the sampling process and the severity of artifacts in images generated by diffusion models. Building on this observation, we introduce the concept of *Similarity Trajectory* to characterize the sampling process and its correlation with the image artifacts presented. Using a data set with over 1,000 times fewer samples than the prior model, we trained a classifier on these trajectories to predict the presence of artifacts in images. In our human evaluation test, the classifier's choices matched those of the participants 72.8% of the time, highlighting the connection between the *Similarity Trajectory* and the occurrence of artifacts. This approach enables differentiation between high-quality and low-quality images using limited training data.

Introductions

Image generation models have recently garnered significant attention for their diverse applications in image inpainting, super-resolution, style transfer, image deblurring, etc. (Rombach et al. 2022; Chung et al. 2023; Li et al. 2022; Lugmayr et al. 2022; Kawar et al. 2022). These technologies have had a profound impact across various industries, including advertising, entertainment, and design. Despite their ability to produce impressive results, these models often suffer from a critical drawback: the generation of images with noticeable artifacts. Common problems include distorted object shapes, merging of distinct objects (Liang et al. 2024).

The presence of artifacts in the generated images highlights the critical need to detect these imperfections. Effective detection can be used to enhance the output of image generation models (Xu et al. 2023; Liang et al. 2024). Recognizing the importance of this task, recent research has focused on collecting large annotated data sets to predict human ratings (Kirstain et al. 2023; Xu et al. 2023; Wu et al. 2023). A particularly noteworthy study by Liang et al. (2024) focused on developing a data set with annotated heat maps that precisely identify where and how the image generation process fails. This line of work has proven valuable for artifact detection and provides explainable results. However, creating such a data set requires a vast amount of labeled data and meticulously curated annotations that pinpoint the specific area causing image generation to fail. This process requires considerable effort, which in turn limits the broader application of these methods.

In this paper, we adopt a different approach that focuses on the final image generated. Instead, we show that analyzing the similarity of denoised images throughout the sampling process provides valuable insight into the severity of artifacts in the final image. We believe this is because each iteration in the sampling process progressively improves the denoised image by building upon the results of the previous step. If these consecutive denoised images are highly dissimilar, this process is similar to adding two different images. This can lead to merged objects, eventually resulting in artifacts in the final image. Hence, by monitoring the similarity between denoised images during the sampling process, we can assess the severity of artifacts originating from this source.

To capture the behavior of the sampling process, we introduce the concept of a *Similarity Trajectory*. This trajectory represents the similarity between denoised images in consecutive time steps throughout the entire sampling process. Therefore, a sudden drop in similarity between denoised images likely signals the presence of artifacts in the generated image. Figure 1 illustrates this with an example: on the left, an image with pronounced artifacts, where the girl's face blends unnaturally with her hair, and on the right, a more natural-looking image. Both images are generated from the same prompt *"A student walking in front of the UT tower, with one hand holding a calculus book"*. We emphasize that a significant advantage of detecting artifacts based on the *Similarity Trajectory* is that it allows the use of a very small data set to train detection algorithms, as the trajectory provides a much more condensed representation compared to raw images. In contrast, previous methods cannot achieve this efficiency because they rely on processing raw images, which require much larger data sets.

The main contributions of our paper are summarized below:

- We discovered that the similarity of denoised images

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

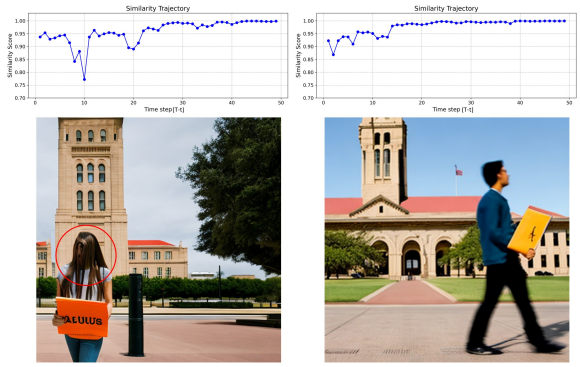


Figure 1: Comparison of images exhibiting strong artifacts (left) versus a more natural appearance, alongside their corresponding *Similarity Trajectories*. The image on the left displays pronounced artifacts, particularly in the circled area where the subject's face blends unnaturally with her hair. This is reflected in the *Similarity Trajectory*, which is more erratic and shows a significant drop. In contrast, the right image appears more natural, with a smoother *Similarity Trajectory* that exhibits consistency.

throughout the sampling process is correlated with the presence of artifacts in generated images.

- We introduce the concept of the *Similarity Trajectory* to characterize the sampling process's behavior and demonstrate its correlation with artifacts in images. This finding allows us detect image artifacts utilizing only a minimal number of training samples.
- We show that the average *Similarity Trajectory* is related to the model's performance in generating images.

In the rest of the paper, we will first review related work in artifact detection. We then explain how the *Similarity Trajectory* is obtained and why it is linked to the presence of image artifacts. Next, we describe our detailed experimental setup used to verify that the *Similarity Trajectory* is statistically different for images with and without artifacts. And we demonstrate how to train a classifier to select images based on the trajectory. Finally, we present experimental results that support our claim that the *Similarity Trajectory* is related to the presence of image artifacts.

Related Works

Over the years, image generation technology has undergone significant advancements (Kingma and Welling 2022; Goodfellow et al. 2014; Ho, Jain, and Abbeel 2020; Rombach et al. 2022). One of the first foundational approaches is the Variational Autoencoder (VAE), which learns to map images to and from a latent space (Rombach et al. 2022). This capability enables VAEs to generate images by randomly sampling the latent space.

In recent years, Diffusion Models have achieved state-of-the-art performance in generative tasks (Dhariwal and Nichol 2021). In this framework, images are generated by multiple iterations of the denoising step (Ho, Jain, and Abbeel 2020). The iterative nature of the diffusion model enables it to produce high-quality data, but at the same time requires substantial computational demands (Ho, Jain, and Abbeel 2020). To address this, the concept of Latent Diffusion Models (LDM) was introduced Rombach et al. (2022). This approach, combined with Auto Encoder, generates images in a lower-dimension latent space, thereby significantly reducing computational costs. Consequently, this development has made the concept of large-scale text-to-image models feasible in reality (Rombach et al. 2022; Podell et al. 2024).

Despite their astonishing performance, the images generated by diffusion models still suffer from artifacts. To address this issue, recent research has focused on detecting these flawed images and fixing the artifacts (Wu et al. 2023; Kirstain et al. 2023; Liang et al. 2024; Xu et al. 2023). A common approach involves labeling the generated images according to human preference and training a model to predict the human preference (Wu et al. 2023; Xu et al. 2023; Kirstain et al. 2023). For example, the "Pick-a-Pic" paper creates a data set of more than 500,000 pairs of images, each pair consisting of two images with a human preferred choice (Kirstain et al. 2023). This data set is then used to fine-tune a CLIP-H model to predict human preferences. Similarly, "ImageReward" collects a data set by ranking and rating images on the basis of their quality. This data set is then used to train a model, which is subsequently used to fine-tune the diffusion model, enhancing the image quality generated (Xu et al. 2023).

Another approach involves providing more detailed information to the model beyond just human preference scores. This is done by annotating specific areas of the image where artifacts occur (Zhang et al. 2023a; Liang et al. 2024). In Liang et al. (2024), the author extends this method by labeling not only artifact regions but also areas with misaligned keywords. This additional information helps the model learn more effectively by offering a deeper understanding of the malformed part.

Our work differs from previous studies that judge the quality on the basis of the final generated images. We discover that the sampling process itself provides fruitful information on whether an image contains artifacts. To our knowledge, this is the first work to reveal this relationship. A direct benefit of uncovering this relationship is that the training data required for classification is reduced by orders of magnitude compared to previous methods. This significant reduction is achieved because characterizing the sampling process is considerably easier than explaining the entire images, as we will elaborate on shortly.

Similarity Trajectory

At a high level, the *Similarity Trajectory* represents the similarity between denoised images in consecutive time steps during the sampling process of a diffusion model. In the following, we will present the mathematical definition of the *Similarity Trajectory* and then explain the intuition behind its importance.

We first define a similarity metric $d(\cdot, \cdot)$ that quantifies how similar two images are: the more similar the images, the

higher the value of d. In this paper, we define d as one minus the output score of the DreamSim model (Fu et al. 2023) when given two images. This subtraction is applied to make sure that our similarity metric increases as the similarity between two images increases, matching our expectations, and keeping its value within the range of [0,1]. We define the *Similarity Trajectory* as a time series $\{z_t\}_{t=T-1}^{1}$, as follows:

$$\{ z_t \}_{t=T-1}^{1}, \text{where } z_t = d\left(x_0^{(t)},\ x_0^{(t-1)}\right). \tag{1}$$

z_t represents the similarity score between the denoised images x_0 predicted in consecutive time steps t and $t-1$. The time index t adheres to the convention used in previous research, where it indicates the sequential steps of the forward diffusion process.

The *Similarity Trajectory* can be computed for different diffusion model training frameworks, provided that the model is trained to predict denoised images from noise corruption. This is because the *Similarity Trajectory* essentially consists of a sequence of denoised images at varying noise levels during the sampling process. However, the exact mathematical form of $x_0^{(t)}$ may vary depending on the training approach used. In the following, we provide the specific mathematical formulas for $x_0^{(t)}$ in the context of Stable Diffusion 2 (SD2) using the DDIM sampler and the EDM2 framework (Karras et al. 2024).

Denoised Image from SD2 with DDIM Sampler In Equation 2, x_t represents the noisy image, i.e., input to the diffusion model. $\epsilon_\theta^{(t)}(\cdot)$ denotes the predicted noise of the U-net in Stable Diffusion 2 at the time step t. The term α_t denotes the cumulative product of $(1-\beta_t)$ corresponding to the time step t. Finally, $x_0^{(t-1)}$ is the denoised image in the next iteration.

$$x_0^{(t-1)} = \frac{x_t - \sqrt{1-\alpha_t}\cdot\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \tag{2}$$

Note that we adopt the notation from (Song, Meng, and Ermon 2021) for Equation 2.

Denoised Image from EDM2 with Heun Sampler In Equation 3, $x_0^{(t-1)}$ represents the denoised image at time step $t-1$, where x_t is the current noisy image at time step t. σ_t is the predetermined noise level at the time step t. The terms n_t and n_t' are the predicted noise, normalized to a standard deviation of 1, during the first and second orders of the Heun method, respectively.

$$x_0^{(t-1)} = x_t - \frac{1}{2}\sigma_t \cdot (n_t + n_t') \tag{3}$$

Linking *Similarity Trajectories* to Image Artifacts

We observe that a sharp decline in the *Similarity Trajectory* suggests the presence of artifacts in the generated image. This decline reflects a misalignment between the denoised images in consecutive time steps. We believe that such misalignment can lead to artifacts because, during the sampling process, each iteration involves subtracting a small amount of predicted noise from the noisy image (latent). This process essentially removes a section of the noisy image (latent) and reintroduces the denoised image (latent). When the denoised images differ significantly between steps, the noise subtraction blends two different denoised images (latent) together. As the diffusion model forms the contours of the objects, this blending can merge the contours of objects predicted in the two different time steps, distorting the original shapes. Consequently, incorrect contours may cause the diffusion model to generate odd shapes and merged objects in the final image. Figure 2 illustrates an example of how artifacts can form. The images shown are the denoised images x_0 at various time steps $T-t$. At time step 11, two men are standing together. However, in the subsequent time step, a different prediction leads to the overlapping of distinct objects, such as a horse and one of the men's bodies, which eventually result in artifacts in the final time step. It is important to note that artifacts can originate from multiple sources. Apart from the merging of different objects, artifacts may also occur when the prediction of the same object changes shape across consecutive time steps. Moreover, the formation of artifacts is frequently unclear and elusive, likely because the denoising occurs within the latent space instead of the direct image space, coupled with the artifacts arising at earlier time steps in the sampling process.

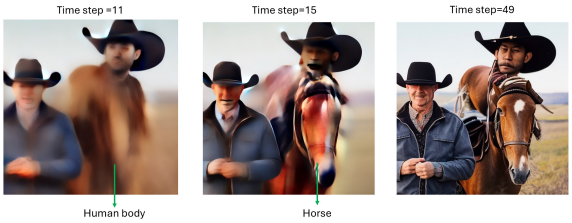


Figure 2: Artifact formation in the sampling process. The denoised images x_0 at various time steps illustrate that changes in the diffusion model's predictions between consecutive steps can cause different objects to overlap. This overlap may distort the original shapes, leading to the appearance of artifacts.

For model evaluation, we aim to investigate whether a stronger model generates fewer images with artifacts. Consequently, a stronger model is expected to exhibit fewer drops in similarity scores, resulting in a higher average *Similarity Trajectory* compared to a weaker model.

Experimental Setup

In the following, we will introduce the experimental setting on how to detect artifacts based on *Similarity Trajectories*.

Artifacts Detection Based on Trajectory

A straightforward method to determine if there is a statistical difference in the *Similarity Trajectory* between images with and without artifacts involves compiling a data set of generated images classified by the presence of artifacts. Then we investigate the statistical property of the *Similarity Trajec-*

tory for each group. This data set can then be used to train a classifier. The classifier's ability to generalize to unseen data and classify whether the images contain artifacts could serve as evidence of the existence of such a relationship.

Data set Creation

Labeling Data We randomly select 250 prompts from the 2014 MS COCO data set (Lin et al. 2014). By using SD2 with the DDIM 50 time step sampler (Rombach et al. 2022), we generate 9 images per prompt, resulting in a total of 2,250 images. To ensure the quality of our data set, we manually review these images, selecting only those with obvious artifacts or those that appears natural-looking (without or mostly without artifacts). Images that fall between these categories are discarded, resulting in 425 natural-looking and 255 artifact-exhibiting images, for a total of 680 images. We emphasize that the selection process focused solely on the quality of the final generated image.

Creating *Similarity Trajectories* Given a similarity metric $d(\cdot, \cdot)$, we calculate the trajectory for each image as $\{z_t\}_{t=1}^{T}$, where $z_t = d(x_0^{(t)}, x_0^{(t-1)})$. Here, $x_0^{(t)}$ represents the denoised latents from the U-net (Ronneberger, Fischer, and Brox 2015) of SD2, which are then decoded into pixel space via the VAE. The chosen similarity metric, DreamSim, operates in the pixel space and provides similarity scores that align closely with human perception. Note that to make DreamSim align with the sense of similarity, we document the value of one minus its output score. We select DreamSim because humans are sensitive to abrupt changes of denoised images, making a human-aligned metric beneficial for this task.

Training Classifier We select a Random Forest model as our classifier due to its strong performance, especially when the amount of training data is scarce. Before inputting the *Similarity Trajectory* into the model, we apply feature engineering, as detailed below, to better capture the information in the data.

First, we apply a Haar transformation to the entire trajectory because of its ability to detect sudden changes in time-series data (Stanković and Falkowski 2003). From the transformation, we obtain a set of coefficients corresponding to each basis function. Additionally, we divide the entire trajectory into three equal segments based on time steps, as well as considering the entire trajectory as a single segment.

For all the segments obtained—whether from the Haar transform coefficients or in the time domain—we calculate 10 statistical features for each segment. These statistical features are chosen to describe the statistical property of the segment of data. They include entropy, 5th percentile, 25th percentile, median, 75th percentile, 95th percentile, mean, standard deviation, number of mean crossings, and number of zero crossings. All of these features serve as inputs for the Random Forest model. Additionally, we trained a *K*-Nearest Neighbor (*k*-NN) model using only the last 18 dimensions of the trajectory, and the prediction probabilities from this *k*-NN model are also used as a feature in the Random Forest classifier.

Experimental Settings for Model Evaluation

In this experiment, our objective is to examine whether there is a connection between the trajectory of average similarity and the performance of the model. We use models provided by the EDM2 research group for the experiment (Karras et al. 2024). These models are trained and evaluated within a consistent framework, ensuring that differences between them are limited to factors of the number of model parameters and the number of training steps. This controlled environment allows us to assess models performance with more precision. Consequently, allowing us to establish the relationship between *Similarity Trajectories* and model performance. We conducted two types of comparison in the experiment: one controlling for model size and the other for training steps.

To obtain the averaged *similarity trajectory*, we first generate 5,000 images for each model using the Heun sampler (Karras et al. 2024) with 32 inference steps and set the S_{churn} parameter at 40. The guidance scale is configured to 1.5, which helps align the generated images more closely with the real data distribution when utilizing classifier-free guidance. In particular, the unconditional model for classifier-free guidance is selected to be identical to the conditional model but does not receive any prompt. This approach ensures that our evaluation is focused solely on the performance of the model being tested.

Next, we directly averaged the 5,000 *Similarity Trajectories* without decoding them into pixel space. The similarity metric used is the Root Mean Square Error (RMSE) between adjacent time steps. We intentionally avoid decoding the latent representations into pixel space to eliminate the influence of decoder sensitivity. This is because we aim to establish a more robust metric for evaluating the diffusion model, independent of decoder-induced variations.

Experimental Results

To assess whether there is a statistical difference between the trajectories of artifact-exhibiting and natural-looking images, we perform experiments using the labeled data set. It is important to note that, to ensure a fair assessment, we do not examine the trajectories during the labeling process.

Statistical Properties for *Similarity Trajectories*

We evaluate the statistical properties of *Similarity Trajectories* from two perspectives. First, we analyze the maximum decline within individual trajectories, which follows directly from our assumption. Second, we assess the standard deviation of similarity scores across different trajectories.

Calculating Maximum Decline Within *Similarity Trajectory* We begin by investigating whether there are statistical differences between the trajectories of artifact-exhibiting and natural-looking images. Our comparison focuses in time steps 13 to 34 of the sampling process because, as indicated by a significant decrease in impurities in the random forest classifier (see Figure 4), this region offers more critical information about the presence of artifacts. We will provide more details on this region in the subsequent section.

A straightforward way to examine our assumption is to compare the maximum decline in the *Similarity Trajectory* for the two groups of images. The maximum decline is defined as the largest cumulative decrease in similarity scores over any continuous decline within the trajectory. We then average these maximum decline values separately for artifact-exhibiting and natural-looking images. We expect artifact-exhibiting images to have a more severe decline because artifacts often accompany a more severe drop in the similarity score.

Our findings support this expectation. For natural-looking images, the average drop is 0.017 ± 0.0011. In comparison, the images artifacts-exhibiting images show an average drop of 0.027 ± 0.0015, reflecting a 1.6-fold difference compared to the natural-looking images. To estimate the precision of the average drop, we report the standard error of the mean. These results indicate that the drop for artifact-exhibiting images is indeed larger than that for images without artifacts, thereby providing statistical support for our observations.

Calculating Standard Deviation Across *Similarity Trajectories* To better understand the statistical properties of *Similarity Trajectories* with respect to time steps, we calculate the standard deviation of similarity scores in different trajectories at the same time step. The underlying intuition is that a drop in similarity scores increases the scatter of the scores, thereby enlarging the overall standard deviation at each time step. As a result, we expect artifacts-exhibiting images to have a larger standard deviation of similarity scores compared with natural-looking images at the same time steps.

In Figure 3, we present the standard deviation of similarity scores at each time step, calculated across all trajectories for two groups of images. It is evident that the standard deviation for artifact-exhibiting images is consistently higher than that for images without artifacts between time steps 13 and 34. In particular, the area under the curve for the standard deviation of *Similarity Trajectories* is 0.24 for images natural-looking images and 0.34 for artifact-exhibiting images, indicating a 1.4-fold difference. This observation reaffirms our hypothesis that images with artifacts are more likely to experience sudden drops in similarity, leading to a larger standard deviation.

Image Quality Classification Based on *Similarity Trajectory*

To explore whether the statistical differences between trajectories can serve as predictors of image quality, we develop a Random Forest classifier comprising 1,000 trees and apply feature engineering to the trajectory discussed in the previous section. We then train the classifier on the balanced data set of 255 natural-looking images and 255 artifact-exhibiting images. The 255 natural-looking images are selected as the first 255 images from that of the original training data set. Given this balanced set-up, random guessing would result in an accuracy of only 50%.

By performing a 10-fold validation test, we obtained an accuracy of 72.35% with a standard error of the mean of 2.1%. This is significantly higher than random guessing, indicating a notable correlation between the *Similarity Trajectory* and image quality.

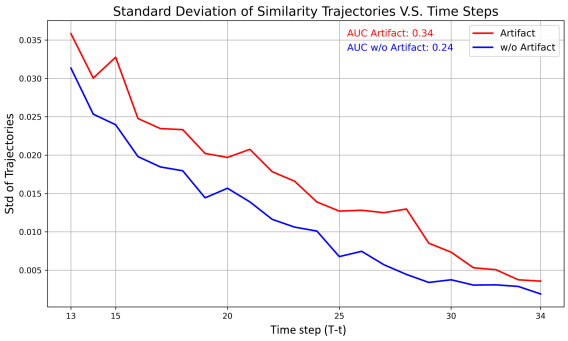


Figure 3: Standard deviation of the *Similarity Trajectories* across 255 artifacts and normal images at each time step. Artifact images consistently exhibit a higher standard deviation compared to images without artifacts, indicating more varied trajectories.

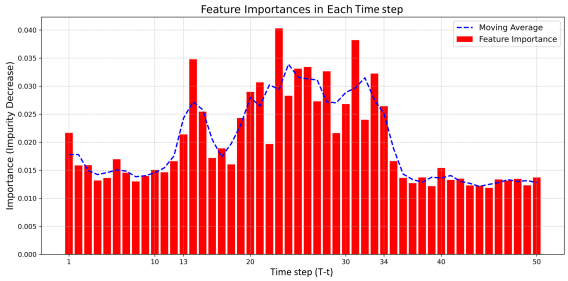


Figure 4: Average Gini impurity reduction at each time step from the random forest classifier. We input raw *Similarity Trajectory*, without any transformations, into the random forest classifier and analyze the average Gini impurity reduction.

To determine which segments of the trajectory are significant for evaluating image quality, we feed the unprocessed trajectory directly into the random forest classifier and examine the average decrease in Gini impurity. A decrease in Gini impurity indicates that specific segments of the trajectory are important for classifying the data. The results are shown in Figure 4. We observe an obvious drop in impurity between time steps 13 and 34, indicating that this range is most indicative of image quality. This critical middle section aligns with the findings in Zhang et al. (2023b), which demonstrated that the middle phase of the sampling process has the greatest influence on the final image.

Intuitively, artifacts are generated during this middle section of the sampling process. At the beginning, the noise level is extremely high, making the image too chaotic for the denoiser to produce meaningful results. Conversely, when the noise level is too low toward the end of the process, the image is too certain to introduce any artifacts or additional objects. Therefore, it stands to reason that the middle section, where the noise level is moderate, is most critical for artifact detection.

Classifing Image Quality in a Realistic Scenario

To determine whether the trained random forest can effectively classify image with and without artifacts in real-world scenarios, we adapted the testing criteria from prior work (Kirstain et al. 2023), with some modifications.

First, we randomly select 100 prompts from the Pick-a-Pic data set. For each prompt, we generate 100 images using SD2, each with a unique random seed. We then evaluate the quality of these images using the trained random forest classifier, which assesses the image artifacts based on their trajectories. For each prompt, we identify the images with the highest and lowest predicted quality and paired them together. After randomly shuffling images within these pairs, we create 100 test pairs, each containing one high-quality and one low-quality image corresponding to the same prompt.

To assess the alignment between our classifier's predictions and human perception of image artifacts, we recruite 10 human participants and divide the 100 test pairs into two equal-sized groups. Five participants review each group. Each participant is asked to carefully examine the image pairs and select one that exhibited more severe artifacts, defined as significant structural or shape deviations from realistic images. If participants are unable to decide which image is worse, they are allowed to declare a tie, reducing random guessing and improving the accuracy of the evaluation. We then evaluate whether the participants consistently identified the images our random forest classifier had predicted as low quality as having more severe artifacts compared to those predicted as high quality. This experiment demonstrates whether the classifier's selections can generalize to unseen data and align with human perception.

In the human evaluation test, we find that, on average, the selections of human participants matched those of the classifier 72.8% of the time and differed 27.2% of the time, as shown in Table 1. In 20.2% of all image pairs, participants are unable to select between the quality of the images, resulting in a tie; these instances are excluded from the calculation. The experimental results demonstrate that the random forest classifier can distinguish image quality in a real-world setting and therefore confirm that the *Similarity Trajectory* is indeed linked with artifacts in images.

Outcome	Percentage (%)
Selections Matched	72.8
Selections Differed	27.2

Table 1: Results of Human Evaluation Test (excluding discarded images due to ties). Human participants' selections matched that of the classifier in 72.8% of cases and differed in 27.2%.

As a point of reference, the prior work (Kirstain et al. 2023) reported that human participants preferred their model's image selections over random choices 71.4% of the time, underscoring the challenge inherent in the task. Although our evaluation setting is different — our image pairs are formed by pairing the best and worst images from the model's selection instead of random choices, and we allowed participants to declare a draw. It is important to note that our method uses fewer than one thousand labeled images compared with the prior work. This further underscores the valuable insights provided by the *Similarity Trajectory* in artifacts detection.

Averaging Trajectories for Model Assessment

In this section, we present the results of experiments designed to explore the relationship between model performance and *Similarity Trajectory*. The first experiment examined how the averaged *Similarity Trajectory* evolves as a model is trained. The second experiment investigated how the averaged *Similarity Trajectory* changes as the model size increases.

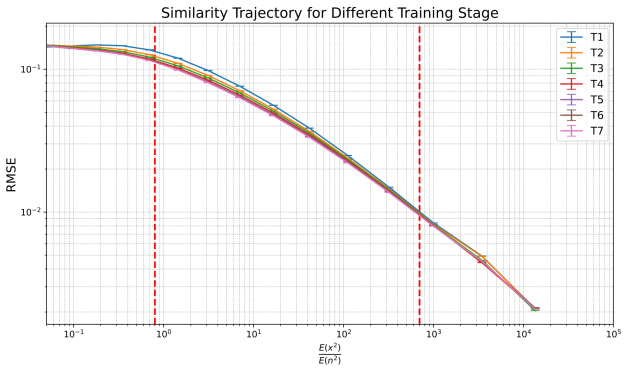
Our assumption is that a more capable model should achieve higher similarity scores, as it is less likely to generate images with artifacts that cause drops in similarity. As training progresses, the model becomes a more effective generator, producing images with fewer artifacts. Additionally, as the model size increases, the model is expected to become stronger. Consequently, we anticipate that the similarity between adjacent denoised latents will be higher in larger models.

For the first experiment, we utilized the snapshots of the large model with a fixed exponential moving average length parameter $\sigma_{\text{rel}} = 0.05$, as provided by the EDM2 repository. The snapshots begin with the model trained on 67 million images, which is represented as T1. As training progress, we used snapshots taken after the model was trained on an extra 128 million images and represented as T2, T3, T4, and so on, resulting in a total of 7 models at different stages of training. For the second experiment, we used the fully trained modeled of size XS, S, M, L, XL, XXL.

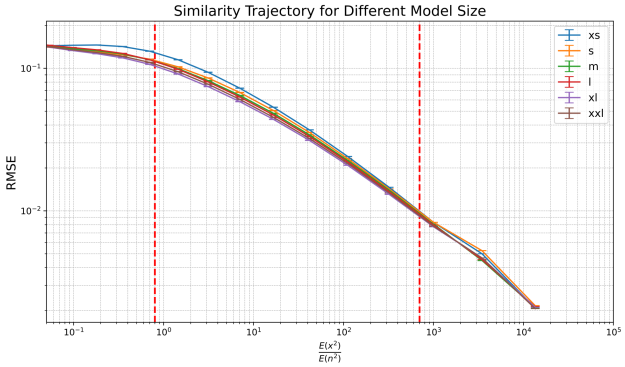
In Figures5a and 5b, we present the average Root Mean Square Error (RMSE) between denoised latents from consecutive time steps during the sampling process, along with the standard error of the mean. The y-axis represents this average RMSE, where larger values indicate more dissimilar latents. The x-axis represents the Signal-to-Noise Ratio (SNR) of the noisy latents, calculated based on the noise level at each time step during training. The latents are normalized by the EDM framework to have a mean of 0 and a standard deviation of 0.5.

We chose to use SNR for the x-axis instead of time steps because the time step arrangement differs between the SD2 and EDM frameworks. Reporting SNR directly avoids confusion and provides a clearer assessment of the noise level. We display only the portion of the averaged trajectory where the noisy latents' SNR ranges from 5×10^{-2} to 10^4. We excluded SNR values below 5×10^{-2} because, at such low levels, the sampling process cannot extract meaningful information from the noisy latents, which falls outside the scope of our discussion.

Within the region marked by the two red lines (SNR between 8×10^{-1} and 7×10^2) in Figure 5a, a clear trend emerges: as training progresses, the RMSE score between adjacent denoised latents gradually decreases compared to earlier training models. This matches our initial assumption



(a) Averaged trajectory at different training stages.



(b) Averaged trajectory for fully trained models of varying sizes.

Figure 5: Comparison of averaged trajectories under different conditions. (a) Effect of training progress on the same model, showing increased latent consistency over time. (b) Influence of model size on fully trained models, with larger models exhibiting greater adjacent latent similarity.

that as the model becomes stronger, fewer artifact-exhibiting images are generated, and hence a higher average similarity is expected between adjacent denoised latents. Although the trajectory order is correct, we want to note that the differences in RMSE scores across different training stages are small. However, the results are statistically significant, as the standard error of the mean is even smaller, as shown in the graph.

In Figure 5b, we observe a similar trend: as model capacity increases, the similarity between denoised latents at adjacent time steps also increases. The only exception is a change in the order between the XL and XXL models, which we attribute to the possibility that the XXL and XL models may have similar performance.

Overall, we demonstrate a correlation between the similarity of adjacent denoised latents and both training time and model capacity. This suggests that stronger models—whether due to extended training or increased capacity—are associated with higher similarity scores.

Conclusion and Limitations

In this paper, we introduce the concept of the *Similarity Trajectory* and demonstrate its relationship with the artifacts presented in images generated by diffusion models. Our findings suggest that even with limited training data—680 labeled images—a classifier can still utilize *Similarity Trajectory* to aid in the selection of artifact-exhibiting images. Although the results are promising, there are several limitations to our work.

While inconsistency in the *Similarity Trajectory* can indicate potential image artifacts, it is not a definitive measure. This limitation has been validated through experimental results. On the other hand, the evaluating of artifacts based on the final image although offers a more direct approach, the wide variety of possible image malformations makes this approach challenging and data-intensive. Future work could explore integrating both methods to leverage their complementary strengths and enhance detection accuracy.

Furthermore, in terms of model evaluation, our results suggest that a more capable model is associated with a higher similarity between adjacent latents in the EDM framework. However, it would be intriguing to investigate whether diffusion models trained under different settings exhibit similar behavior. We hope our study sparks further exploration into the connection between the sampling process and image quality and generator performance.

Acknowledgments

AAAI is especially grateful to Peter Patel Schneider for his work in implementing the original aaai.sty file, liberally using the ideas of other style hackers, including Barbara Beeton. We also acknowledge with thanks the work of George Ferguson for his guide to using the style and BibTeX files — which has been incorporated into this document — and Hans Guesgen, who provided several timely modifications, as well as the many others who have, from time to time, sent in suggestions on improvements to the AAAI style. We are especially grateful to Francisco Cruz, Marc Pujol-Gonzalez, and Mico Loretan for the improvements to the BibTEX and LATEX files made in 2020.

The preparation of the LATEX and BibTEX files that implement these instructions was supported by Schlumberger Palo Alto Research, AT&T Bell Laboratories, Morgan Kaufmann Publishers, The Live Oak Press, LLC, and AAAI Press. Bibliography style changes were added by Sunil Issar. \pubnote was added by J. Scott Penberthy. George Ferguson added support for printing the AAAI copyright slug. Additional changes to aaai25.sty and aaai25.bst have been made by Francisco Cruz, Marc Pujol-Gonzalez, and Mico Loretan.

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!

References

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2023. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 8780–8794. Curran Associates, Inc.

Fu, S.; Tamir, N. Y.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Karras, T.; Aittala, M.; Lehtinen, J.; Hellsten, J.; Aila, T.; and Laine, S. 2024. Analyzing and Improving the Training Dynamics of Diffusion Models. In *Proc. CVPR*.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. In *Advances in Neural Information Processing Systems*.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 36652–36663. Curran Associates, Inc.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; Ke, J.; Dvijotham, K. D.; Collins, K. M.; Luo, Y.; Li, Y.; Kohlhoff, K. J.; Ramachandran, D.; and Navalpakkam, V. 2024. Rich Human Feedback for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19401–19411.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'a r, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11461–11471.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Stanković, R. S.; and Falkowski, B. J. 2003. The Haar wavelet transform: its status and achievements. *Computers Electrical Engineering*, 29(1): 25–44.

Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; and Li, H. 2023. Better Aligning Text-to-Image Models with Human Preference. *ArXiv*, abs/2303.14420.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In

Thirty-seventh Conference on Neural Information Processing Systems.

Zhang, L.; Xu, Z.; Barnes, C.; Zhou, Y.; Liu, Q.; Zhang, H.; Amirghodsi, S.; Lin, Z.; Shechtman, E.; and Shi, J. 2023a. Perceptual Artifacts Localization for Image Synthesis Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7579–7590.

Zhang, Z.; Zhao, Z.; Yu, J.; and Tian, Q. 2023b. Shiftddpms: Exploring conditional diffusion models by shifting diffusion trajectories. *arXiv preprint arXiv:2302.02373*.