Appendix

A. Background

A.1. Diffusion Models for Text-to-Image

Diffusion Models (DMs) have emerged as a powerful class of generative models, especially for tasks involving complex structured outputs such as images. In the context of text-to-image generation, DMs are conditioned on textual prompts to guide the generation process. This is typically achieved by conditioning the denoising network on a text embedding, allowing for fine-grained control over the generated content.

A diffusion model defines a **forward process** (also called the noising process) that gradually adds Gaussian noise to the input data, and a **reverse process** (the denoising process) that learns to reconstruct the original data from the noisy version.

Forward Process. Let x_0 be a data sample from the real data distribution $q(x_0)$. The forward process is defined as a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad \text{for } t = 1, \ldots, T, \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ is a predefined noise schedule. The cumulative form is:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Reverse Process. The reverse process is learned as a parameterized distribution $p_\theta(x_{t-1}|x_t)$ intended to approximate the true reverse posterior $q(x_{t-1}|x_t)$. It is modeled as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3}$$

Training Objective. Training is done by minimizing a variational lower bound (VLB) on the negative log-likelihood:

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_q \left[\log \frac{q(x_1, \ldots, x_T|x_0)}{p_\theta(x_0, \ldots, x_{T-1}|x_T)} \right]. \tag{4}$$

Ho et al. [3] showed that this loss can be simplified to a noise prediction loss:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right], \tag{5}$$

where:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{6}$$

A.2. Types of Diffusion Models

1. Denoising Diffusion Probabilistic Models (DDPM). DDPM [3] learns to approximate the reverse of the forward noising process via a deep neural network. The predicted mean of the reverse step is:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right), \tag{7}$$

and the full sampling equation becomes:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z, \quad z \sim \mathcal{N}(0, \mathbf{I}), \tag{8}$$

where σ_t can be fixed or learned.

Despite its excellent generation quality, DDPM requires hundreds to thousands of sampling steps due to its stochastic nature and slow convergence.

2. Denoising Diffusion Implicit Models (DDIM). DDIM [13] removes the stochasticity of DDPM's sampling by introducing a deterministic, non-Markovian sampling procedure. The denoising process estimates the original image x_0 from x_t as:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t) \right), \tag{9}$$

and computes the next step using:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\theta(x_t, t). \tag{10}$$

This enables faster inference (as few as 10–50 steps), with no additional training overhead. A stochastic variant of DDIM reintroduces noise via a hyperparameter η:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\eta^2\sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \eta\sigma_t z. \tag{11}$$

3. Latent Diffusion Models (LDM). Latent Diffusion Models (LDM) [10] address the computational bottleneck of pixel-space diffusion by operating in a compressed *latent space*. A pretrained autoencoder is used to transform the image into a lower-dimensional representation where diffusion is performed.

- **Encoding.** Compress the image x into a latent code $z = \mathcal{E}(x)$ using a convolutional encoder or a vector quantized autoencoder (VQ-VAE [15]).
- **Latent Diffusion.** Apply the denoising diffusion process in latent space:

  $$p_\theta(z_{t-1}|z_t)$$

  where z_t is the noisy latent at step t and θ are the parameters of the model.

- **Decoding.** The denoised latent z_0 is then decoded back to image space via:
$$\hat{x} = \mathcal{D}(z_0)$$
using the decoder $\mathcal{D}$.
- **Conditioning.** The reverse process is conditioned on additional information c (e.g., text or image embeddings), leading to a noise prediction formulation of:
$$\epsilon_\theta(z_t, t, c),$$
where the model learns to predict the noise given the latent input, time step, and condition c (e.g., CLIP embedding of the prompt).

where c is a text or image condition. Stable Diffusion [10] uses CLIP text embeddings as c and a U-Net [11] backbone with cross-attention layers to modulate generation using prompt semantics.

By working in a lower-dimensional space, LDMs are computationally efficient and enable generation of high-resolution images with fewer resources.

A.3. Fine-Tuning of Diffusion Models

While pretrained diffusion models such as DDPM, DDIM, or LDM provide strong generative priors, many real-world applications require personalization or domain adaptation. Fine-tuning allows models to specialize on specific styles, identities, or downstream datasets, often with only a few target samples.

Formally, let $\mathcal{L}_{\text{diff}}(\theta)$ denote the loss function of a pretrained diffusion model with parameters θ trained on dataset $\mathcal{D}_{\text{pretrain}}$. Fine-tuning introduces a new dataset $\mathcal{D}_{\text{target}}$ and adapts $\theta \to \theta^*$ such that:

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{diff}}(\theta; \mathcal{D}_{\text{target}}) \tag{12}$$

Depending on the desired controllability, personalization, or efficiency, different fine-tuning strategies are employed. Below, we describe some widely adopted techniques.

- **Full Fine-Tuning.** The most straightforward approach is to fine-tune all parameters of the model (e.g., the U-Net in the denoising process). However, this is computationally expensive and prone to overfitting when only a few target samples are available.
- **DreamBooth [12].** DreamBooth introduces identity-preserving fine-tuning of text-to-image diffusion models. It conditions the model on a unique identifier (e.g., $\langle S \rangle$), which is associated with a particular subject or style. The objective is to generate images $\hat{x}$ that both match the target prompt p and preserve visual fidelity to the subject image x:
$$\mathcal{L}_{\text{DreamBooth}} = \mathbb{E}_{x,p}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}z + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c_p)\|^2\right] \tag{13}$$
where c_p is the CLIP embedding of the prompt including the identifier (e.g., "a photo of $\langle S \rangle$ person in a cowboy hat"), and the loss is computed using the noise prediction objective.
- **LoRA (Low-Rank Adaptation) [4].** LoRA is a parameter-efficient technique where low-rank matrices (A, B) are injected into existing weight matrices of the model (e.g., attention layers) such that:
$$W' = W + \Delta W, \quad \Delta W = AB \tag{14}$$
with $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, where $r \ll d$. Only A and B are updated during fine-tuning, while W is frozen. This drastically reduces the number of trainable parameters and allows for faster adaptation with minimal compute.
- **Textual Inversion [1].** Instead of modifying the model weights, textual inversion learns a new token embedding e_* such that it captures the concept present in the training images. This embedding can be directly plugged into prompts:
"A painting of e_* in Van Gogh style"
The training optimizes:
$$\min_{e_*} \mathbb{E}_{x,p}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}z + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c_{p+e_*})\|^2\right] \tag{15}$$
allowing users to encode concepts without modifying model parameters.
- **Custom Diffusion [5].** Custom Diffusion combines the benefits of LoRA and personalized prompts by enabling user-specific adaptation. It trains lightweight modules on top of a frozen diffusion backbone while optionally learning prompt embeddings. This method is optimized for fast fine-tuning and minimal memory usage.
- **Pivotal Tuning [9].** This technique first generates an image from a given text prompt and then fine-tunes the model to reconstruct the same image from the same prompt, ensuring deterministic and identity-preserving outputs.

Fine-tuning is essential for personalizing diffusion models or adapting them to domain-specific data. While full model tuning is often expensive, parameter-efficient approaches like DreamBooth, LoRA, and Textual Inversion enable rapid and memory-efficient specialization. The choice of technique depends on the trade-off between fidelity, data availability, and compute constraints.

A.4. Latent Representation of Images

To conduct fine-grained analysis of attribution behavior, we extract visual latent representations from three distinct feature spaces, each emphasizing a different visual dimension—semantics, structure, and style.

- **CLIP Embeddings.** We use OpenAI's ViT-B/32 model from the CLIP [8] family, which is pretrained using a contrastive objective to align images and natural language

descriptions. CLIP captures high-level *semantic* information and has been shown to be effective for tasks like zero-shot classification, concept retrieval, and semantic alignment across domains. We extract the 512-dimensional embedding from the image encoder as the semantic feature vector.

- **DINOv2 Embeddings.** For structural representations, we use the ViT-S/14 backbone from DINOv2 [7], a self-supervised vision transformer trained on a large-scale dataset without labels. DINOv2 embeddings preserve information related to spatial arrangements, edge contours, and object topology. We obtain the global [CLS] token from the final transformer layer as the 384-dimensional structural descriptor.
- **Gram Matrix Representations.** To capture style, we follow the procedure proposed in Gatys et al. [2] by computing Gram matrices from feature maps of a pretrained VGG-19 network (layer `relu3_1`) available in the PyTorch 'torchvision' module. The Gram matrix encodes correlations between activation maps, which reflect *style* elements such as texture, color repetition, and artistic patterns. The resulting matrix is flattened into a 512-dimensional style embedding.

Each representation is extracted from a fixed, frozen backbone to ensure that our comparisons are driven by attribution behavior rather than network training dynamics. Together, these latent features allow us to perform robust, multi-perspective evaluations of influence across different axes of visual similarity.

A.5. KernelSHAP

KernelSHAP [6] is an efficient model-agnostic method to approximate Shapley values for explaining individual predictions of complex models. It frames Shapley value computation as a weighted linear regression problem over a simplified binary input space, making it particularly effective for attribution when the exact Shapley value is computationally infeasible to compute directly.

Background: Shapley Values

Given a utility function $v : 2^N \to \mathbb{R}$ over a finite set of N players (in our case, candidate training images in T_{reduced}), the Shapley value ϕ_i for player i is defined as the average marginal contribution of i across all possible subsets $S \subseteq N \setminus \{i\}$:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right]. \tag{16}$$

This formulation ensures desirable properties: symmetry, efficiency, linearity, and null player. However, computing this directly requires exponential time ($2^{|N|}$ subsets), making it impractical for large N.

KernelSHAP Approximation

KernelSHAP sidesteps this intractability by modeling the utility function $v(S)$ as a weighted linear function over binary feature masks:

$$v(S) \approx \sum_{i=1}^{N} \phi_i z_i = \langle z, \phi \rangle, \tag{17}$$

where $z \in \{0, 1\}^N$ is the binary mask indicating which players (training images) are in the coalition S. Each sampled mask $z^{(m)}$ corresponds to a coalition S_m whose utility $v(S_m)$ is evaluated using the negative energy distance.

Weighted Linear Regression Formulation

The Shapley values $\phi = [\phi_1, \ldots, \phi_N]$ are obtained by solving the following weighted least-squares problem:

$$\phi = \arg\min_{\phi \in \mathbb{R}^N} \sum_{m=1}^{M} w^{(m)} \left(v(S_m) - \langle z^{(m)}, \phi \rangle \right)^2, \tag{18}$$

where: - M is the number of sampled subsets, - $z^{(m)}$ is the binary mask vector for subset S_m, - $v(S_m)$ is the utility score of the subset (negative ED), - $w^{(m)}$ is the **SHAP kernel weight** for subset S_m, given by:

$$w^{(m)} = \frac{(N-1)}{\binom{N}{|S_m|} \cdot |S_m| \cdot (N - |S_m|)}. \tag{19}$$

This kernel assigns higher weights to subsets near the center of the coalition space (i.e., with size close to $N/2$), and lower weights to very small or very large coalitions, ensuring better stability and fairness.

A.6. KernelSHAP in Our Framework

In our setup, we apply KernelSHAP to attribute training examples (players) from the reduced set T_{reduced} to the generated sample $\hat{z}$. Each subset S_m is a randomly sampled coalition of training images, and its utility $v(S_m)$ is computed as the negative energy distance between the generated set G and subset S_m.

Solving the regression in Equation 18 yields a Shapley value vector ϕ where each entry ϕ_i indicates the influence of image $x_i \in T_{\text{reduced}}$ on the generation of $\hat{z}$.

This approach retains the axiomatic guarantees of Shapley values while being scalable and flexible to plug into our attribution framework.

A.7. Energy Distance as Utility in Shapley

In our proposed attribution framework, we employ **Energy Distance (ED)** as the utility function within the KernelSHAP algorithm. We now present a theoretical justification for this choice, grounded in the properties of ED established by Székely and Rizzo [14].

A.8. Definition of Energy Distance

Let F and G denote two probability distributions over $\mathbb{R}^d$, with random variables $X \sim F$ and $Y \sim G$. The **population Energy Distance** between F and G is defined as:

$$\mathcal{E}(F, G) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\| \quad (20)$$

where X, X' are i.i.d. samples from F, and Y, Y' are i.i.d. samples from G. Here, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$.

Key Property

As shown in [14], Energy Distance satisfies:

$$\mathcal{E}(F, G) = 0 \iff F = G \quad (21)$$

This makes ED a valid metric for comparing distributions. It is sensitive to all moments of the distributions and captures differences in both location and scale, unlike simpler divergences such as KL or JS which may be undefined or unreliable in high dimensions.

Empirical Estimate

Given empirical samples:

- $A = \{a_1, \ldots, a_m\}$ drawn from F
- $B = \{b_1, \ldots, b_n\}$ drawn from G

The unbiased empirical estimate of Energy Distance is:

$$\begin{aligned} \text{ED}(A, B) = \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \|a_i - b_j\| \\ - \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \|a_i - a_j\| \\ - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|b_i - b_j\| \end{aligned} \quad (22)$$

This closed-form expression is nonparametric and scalable—making it particularly suitable for high-dimensional latent representations (e.g., CLIP, DINOv2, Gram matrices) used in generative modeling.

Energy Distance as a Utility Measure for Subset Influence

We define the SHAP utility function for a subset $S \subseteq T$ (training examples) as:

$$v(S) = -\text{ED}(L_S, L_G) \quad (23)$$

Where:

- L_S: Latent representations of subset S
- L_G: Latent representations of generated set G

We negate the Energy Distance because KernelSHAP assumes that higher utility values imply greater contribution, and lower ED indicates stronger alignment in distributional structure.

Benefits

The use of ED as a utility function offers several benefits:

- **Faithfulness:** $\mathcal{E}(F, G) = 0$ iff $F = G$, ensuring alignment under perfect matching.
- **Nonparametric:** No assumptions about the form of F or G.
- **Scalability:** Efficient to compute over latent representations in high dimensions.
- **Smoothness:** ED behaves continuously over latent space, enabling gradient-based or regression-based approximations in KernelSHAP.

This mathematically grounded formulation validates Energy Distance as a robust utility metric for attributing influence based on distributional similarity.

Axioms of Shapley Values

We justify the use of Energy Distance (ED) as a utility function in our KernelSHAP-based attribution framework by verifying its compliance with the Shapley axioms. Recall that our utility function is defined as:

$$v(S) = -\text{ED}(L_S, L_G), \quad (24)$$

where L_S and L_G denote the latent representations (e.g., CLIP, DINOv2, or Gram features) of subset $S \subseteq T$ and the generated image set G, respectively. The negative sign ensures that lower distances correspond to higher utility.

Axiom 1: Efficiency

The Shapley value satisfies efficiency if the total utility is distributed among all players:

$$\sum_{i \in T} \phi_i(v) = v(T). \quad (25)$$

KernelSHAP approximates Shapley values via linear regression over sampled coalitions. While exact equality may not hold due to approximation, the property is preserved in expectation under the KernelSHAP setting. Thus, efficiency is approximately satisfied.

Axiom 2: Symmetry

If two training examples t_i and t_j contribute identically to every coalition, i.e.,

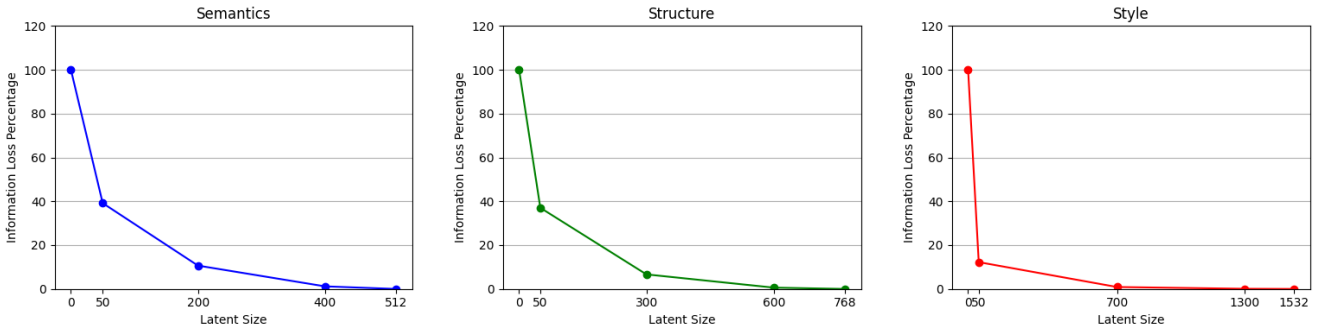
$$v(S \cup \{t_i\}) = v(S \cup \{t_j\}) \quad \forall S \subseteq T \setminus \{t_i, t_j\}, \quad (26)$$

then $\phi_i = \phi_j$. Since Energy Distance is a symmetric function of the two distributions, if t_i and t_j yield the same distance when added to any subset, the symmetry condition holds.

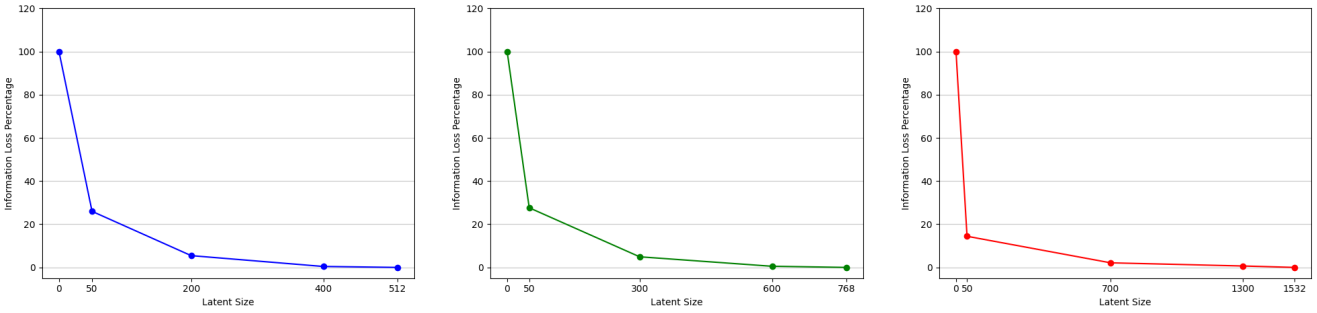
Axiom 3: Dummy

If a player t_i does not affect the value of any subset, i.e.,

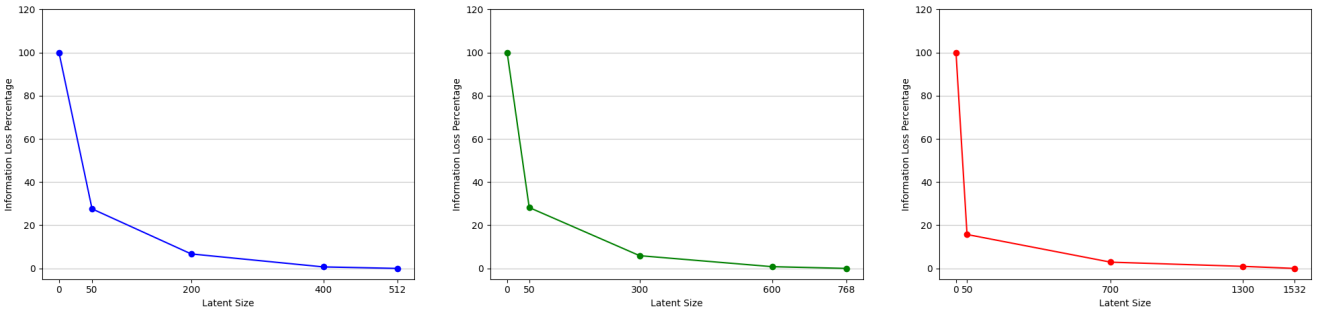
$$v(S \cup \{t_i\}) = v(S) \quad \forall S \subseteq T, \quad (27)$$



(a) PopArt



(b) Ukiyo-e



(c) Munch

Figure 1. Percentage of latent information loss after PCA-based dimensionality reduction, computed across latent representations for the different datasets. This analysis helps determine optimal latent dimensionality while preserving the maximum relevant signal for attribution.

then $\phi_i = 0$. This condition is satisfied for ED because adding a latent vector that does not reduce the distance to G yields zero marginal contribution.

Axiom 4: Additivity

For two utility functions v_1 and v_2, if $v = v_1 + v_2$, then:

$$\phi_i(v) = \phi_i(v_1) + \phi_i(v_2). \tag{28}$$

Since Energy Distance is a metric, additive combinations of utility functions preserve linearity, satisfying this axiom.

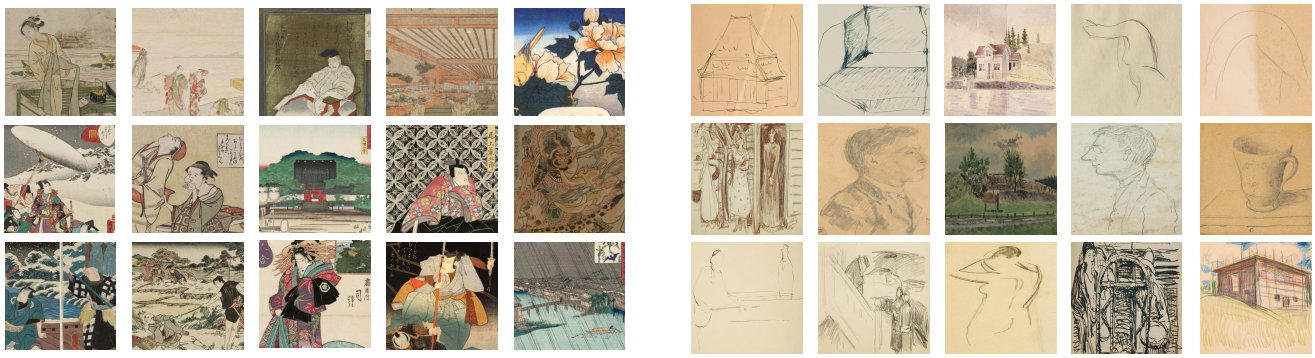
Conclusion

All four Shapley axioms: Efficiency, Symmetry, Dummy, and Additivity, are satisfied or well-approximated when Energy Distance is used as a utility function under the KernelSHAP framework. This provides strong theoretical justification for our distribution-level attribution approach.

B. Experimental Details and Results

B.1. Datasets

In our experiments, we utilize the WikiArt PopArt dataset, which contains 1482 images that represent the Pop Art movement—a cultural phenomenon that emerged in the 1950s and gained widespread traction in the 1960s across America and Britain. Unlike many contemporary image datasets (such as ImageNet or CelebA) that exhibit structured scenes or homogeneous styles, the PopArt dataset features high intra-class variance and low stylistic monotonicity. This makes PopArt particularly challenging and appropriate for evaluating attribution, as it demands disentangling



(a) Ukiyo-e (b) Munch

Figure 2. Example artworks from the (a) Ukiyo-e and (b) Munch datasets.

not only semantics but also subtle structural and stylistic cues.

In addition to PopArt, we incorporate the Ukiyo-e dataset, which consists of traditional Japanese woodblock prints. Ukiyo-e provides a contrasting testbed due to its distinctive stylistic attributes such as flattened perspective, bold contouring, and repetitive structural motifs, which differ markedly from Western artistic traditions. The dataset introduces stronger correlations in style and structure, thereby allowing us to evaluate whether attribution methods can still disentangle semantic relevance from stylistic alignment under a more uniform yet stylistically dominant regime.

We further construct an artist-specific dataset of Edvard Munch, whose artworks (including The Scream and The Dance of Life) are known for their emotionally charged, expressionistic style and unique brushwork. This dataset is particularly valuable for attribution as it emphasizes artist-specific semantics and style interplay—many works share common semantic themes (e.g., figures, landscapes) but are visually dominated by Munch's distinct stylistic fingerprint. Evaluating attribution on this dataset ensures that methods can correctly identify influence beyond shared content, recognizing contributions grounded in stylistic signature.

Taken together, these three datasets—PopArt, Ukiyo-e, and Munch—provide complementary challenges that conventional generic datasets fail to capture. Whereas ImageNet or CelebA predominantly emphasize semantic content (objects, faces), our chosen datasets stress the full triad of semantics, structure, and style, which are central to attribute-aware attribution. This choice ensures a more rigorous and representative evaluation of the proposed DoTA framework.

B.2. Dimensionality of latent representations

To our observation, we noticed that Energy Distance fails to compare two distributions of uneven size and of high dimensions. This is due to several underlying statistical and geometric factors that degrade its discriminative power in such settings.

Curse of Dimensionality and Distance Concentration: In high-dimensional spaces, pairwise distances between points from the same or different distributions tend to concentrate, making it difficult for distance-based metrics like Energy Distance to distinguish between distributions. Formally, let $x, x' \sim \mathcal{N}(0, I_d)$ be i.i.d. random vectors in $\mathbb{R}^d$, then:

$$\mathbb{E}[\|x - x'\|^2] = 2d, \quad \text{Var}[\|x - x'\|^2] = O(d) \tag{29}$$

However, as $d \to \infty$, the normalized pairwise distances $\|x - x'\|/\sqrt{d}$ concentrate around a constant value, reducing the contrast between intra- and inter-distribution distances.

Unequal Sample Size Bias: Let the empirical Energy Distance between two distributions F and G, represented by samples $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$, be given by:

$$\begin{aligned} \text{ED}^2(F, G) = \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \|x_i - y_j\| \\ - \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \|x_i - x_j\| \\ - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|y_i - y_j\| \end{aligned} \tag{30}$$

When $m \neq n$, the cross-term $\frac{2}{mn} \sum \|x_i - y_j\|$ dominates the expression, while the within-distribution terms (which require m^2 or n^2 pairings) may be poorly estimated if one of the distributions has very few samples. This leads to biased and unreliable distance estimation.

Outlier Sensitivity in High Dimensions: Since Energy Distance is based on raw Euclidean distances, it is highly sensitive to outliers. In high-dimensional spaces, the influence of even a single point with large magnitude grows with

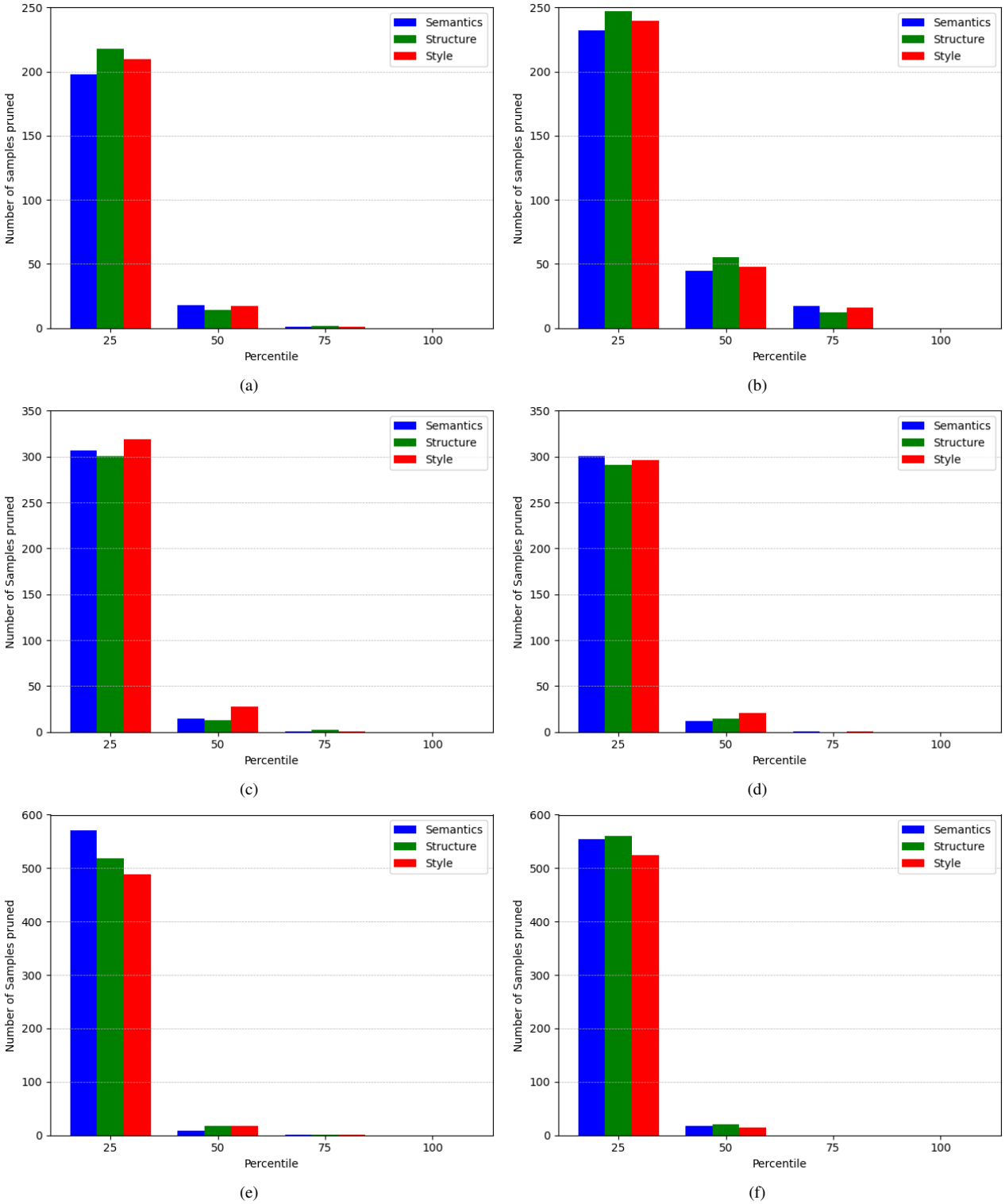


Figure 3. Number of samples pruned during SSP with different representations: (a) without adversarial attack and (b) with adversarial attack for PopArt, (c) without adversarial attack and (d) with adversarial attack for Ukiyo-e, (e) without adversarial attack and (f) with adversarial attack for Munch.

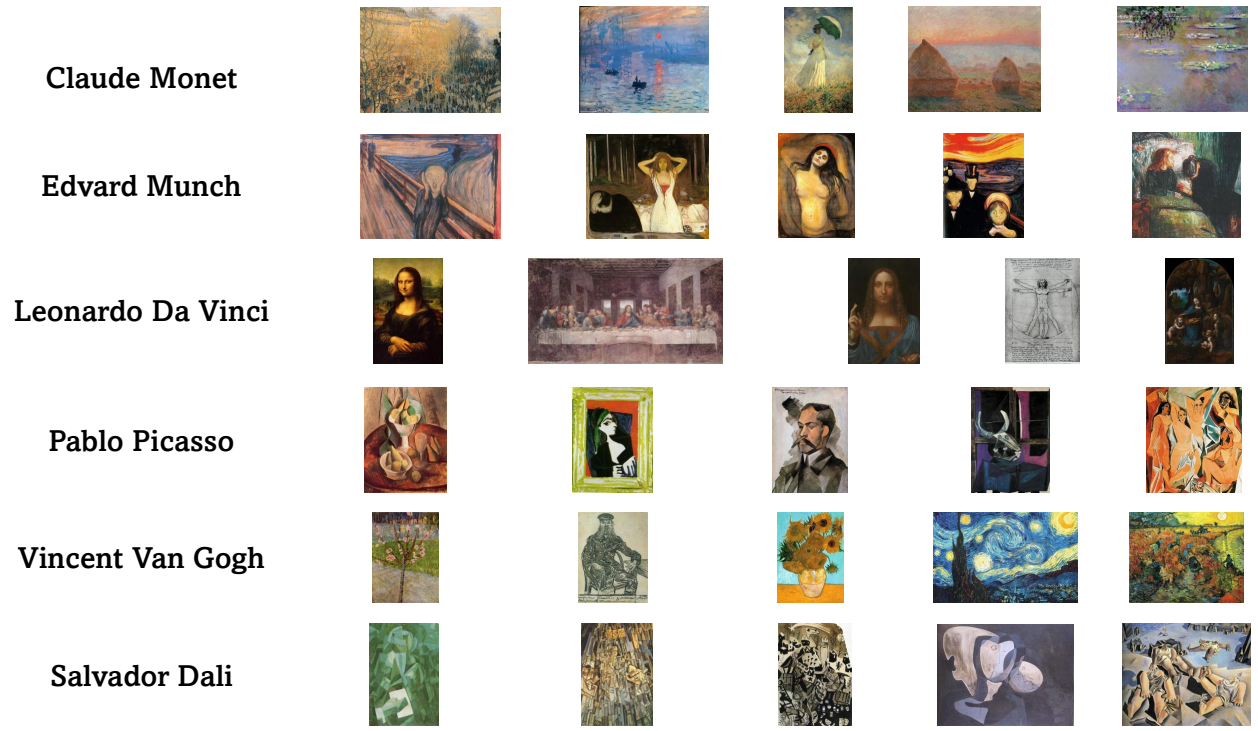


Figure 4. Adversarial set of popular artworks across artists: Claude Monet, Edvard Munch, Leonardo Da Vinci, Pablo Picasso, Vincent Van Gogh, and Salvador Dali. The set considers some of the widely popular artworks such as Starry Night, The Scream, Mona Lisa, etc.

dimension:

$$\|x_i - y_j\|^2 = \sum_{k=1}^{d} (x_i^{(k)} - y_j^{(k)})^2 \propto d \tag{31}$$

Thus, outliers or mismatched supports between the two distributions can disproportionately affect the final ED score.

Sparse Sample Coverage: High-dimensional distributions require exponentially more samples for sufficient coverage. With limited samples, the empirical estimate of Energy Distance suffers due to poor coverage of the support, further reducing its reliability.

Given these limitations, we adopt a PCA-based projection into lower-dimensional latent spaces before computing Energy Distance, ensuring stable estimation and tractable comparison even under sample size imbalance.

In Table 1, we report the percentage of information loss incurred due to dimensionality reduction using PCA. Based on our observations, reducing the latent representation to a dimensionality of 50 strikes an effective balance—preserving most of the meaningful information while creating favorable conditions for Energy Distance to provide reliable insights into distributional alignment. This choice mitigates the adverse effects of high dimensionality and ensures stable estimation of distances in the latent space.

Leveraging Latent Embeddings for Robust Attribution: Beyond mitigating the statistical limitations of Energy Distance, operating in latent space offers inherent advantages for attribution. Latent embeddings derived from pretrained encoders serve as compact, information-preserving representations of complex visual inputs, where semantic features such as style, structure, and composition are disentangled from pixel-level noise. This mapping suppresses irrelevant high-frequency variations while retaining the characteristics most relevant for attribution, thereby ensuring that similarity or distance computations reflect meaningful artistic attributes rather than raw pixel correlations. Moreover, latent representations provide a stable and model-agnostic foundation for attribution, circumventing the fragility of gradient-based methods that depend on local parameter dynamics and the impractical requirement of training logs, as in runtime attribution frameworks. Crucially, working in a reduced latent space enables efficient pruning of candidate samples, which both lowers the computational cost of Shapley value estimation and improves robustness by focusing on semantically aligned examples. In this way, latent space functions as a principled and scalable domain where distributional alignment can be reliably assessed.

B.3. Threshold for Subset Selection

To determine an appropriate threshold for search space pruning in our framework, we empirically analyzed how

varying the percentile cutoff of the Energy Distance (δ_{ED}) impacts the number of training samples removed. Figure 3 illustrates the pruning effect across different percentiles for the three latent representations—Semantics, Structure, and Style—across Popart, Ukiyo-e, and Munch datasets of varying sizes (1482, 2000, and 5800 samples, respectively).

For Popart, we observed that at $q = 25\%$, around 200 samples were pruned, corresponding to nearly 10% of the dataset. These removed examples exhibited high Energy Distances, i.e., weak statistical alignment with the generated outputs, making them unlikely to have meaningfully influenced the generative process. This moderate pruning effectively improved attribution quality while retaining the majority of candidates (over 1200) for stable Shapley estimation. Increasing the threshold beyond 25% led to marginal additional pruning, whereas more aggressive thresholds (below 15%) destabilized the attribution process due to overly small candidate pools.

For Ukiyo-e ($N = 2000$), pruning at $q = 25\%$ removed between 300–320 samples depending on the latent type, leaving roughly 1680–1700 candidates. This corresponds to pruning around 15% of the dataset—more substantial than in Popart but still leaving a sufficiently large retained pool. Under adversarial perturbations, the pruning counts shifted slightly (e.g., 291–301 removals in structure/semantic latents and 296 in style), indicating sensitivity to distributional changes. Importantly, these perturbations remained within the range of 14–16% pruning, showing that the mechanism naturally filtered out off-distribution samples injected during adversarial attacks.

For Munch ($N = 5800$), pruning at $q = 25\%$ consistently removed 480–570 images across different latents, leaving approximately 5200–5320 candidates. This corresponds to pruning around 9–10% of the dataset, comparable in proportion to Popart despite the larger dataset scale. In the adversarial setting, pruning rates remained similar, with slightly higher removals in certain latents (e.g., 554 semantic, 561 structure), again confirming robustness to distributional perturbations. Because of the dataset's larger size, the retained pool remained very large, ensuring stability of Shapley estimation.

Overall, $q = 25\%$ emerges as a principled and empirically validated trade-off across datasets: it prunes statistically misaligned training samples while retaining a sufficiently large and stable candidate pool for attribution. While stricter pruning (e.g., $q < 25\%$) could, in principle, remove additional weakly aligned examples, it does not guarantee improved attribution quality and may destabilize Shapley estimation by discarding genuinely influential samples. Thus, we adopt $q = 25\%$ as the default threshold, balancing pruning aggressiveness with attribution stability across datasets of varying scales.

B.4. Adversarial Set

To assess the robustness and reliability of our attribution methods under distributional shifts, we construct an **adversarial set** tailored to each dataset. The adversarial set consists of carefully selected artworks that are visually or semantically similar to the target distribution but originate from artists outside the training data. Such samples mimic key stylistic or structural elements (e.g., brushwork, color palette, or subject matter) while being irrelevant to the actual source dataset, thereby serving as potential confounders in the attribution process.

For the **PopArt** and **Ukiyo-e** datasets, the adversarial set includes five iconic artists: Claude Monet, Edvard Munch, Leonardo da Vinci, Pablo Picasso, and Vincent van Gogh. These artists were chosen because their artworks exhibit stylistic overlaps with PopArt and Ukiyo-e characteristics (such as bold colors, fine linework, or thematic abstraction) yet are not part of the respective training collections.

In contrast, for the **Munch** dataset, we substitute Edvard Munch with Salvador Dalí, yielding the adversarial set: Claude Monet, Salvador Dalí, Leonardo da Vinci, Pablo Picasso, and Vincent van Gogh. This adjustment prevents leakage of in-domain Munch artworks while ensuring that the adversarial set remains equally challenging. Dalí's surrealist distortions and bold compositions introduce stylistic ambiguities that closely resemble Munch's expressionist elements, making attribution particularly difficult.

The inclusion of such adversarial sets is critical for comprehensive evaluation. In real-world scenarios, generative models often face out-of-distribution samples or contaminated training data that may resemble the target domain. A reliable attribution method must demonstrate both *sensitivity*, by correctly identifying genuine contributors, and *specificity*, by rejecting stylistically similar but irrelevant sources.

Figure 4 presents representative adversarial samples. This setup functions as a stress test, allowing us to evaluate whether the attribution method maintains precision and robustness in identifying true sources of influence when challenged with visually confounding examples.

B.5. Results

We analyze the behavior of different attribution methods, such as Cosine Similarity, DoTA without Search Space Pruning (SSP), and DoTA with SSP. Apart from the results presented earlier, we conduct an ablation study to study the impact of key components. In Table 1, we present our results for the ablation study. We consider different combinations of representation size (50 and 400) and K (5 and 10) on the Ukiyo-e dataset. We observe that with an increase in the representation size from 50 to 400, the number of instances where style confidence for Leave-K is higher than Take-K increases. This behaviour is consistent for both the values

Method	Semantics		Structure		Style	
	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)
Relative Style Confidence						
Cosine Similarity	0.0313	0.0424	0.0227	**0.0139**	0.0303	**0.0313**
DoTA wo SSP	0.0323	**0.0204**	0.0348	0.0268	0.0372	0.0332
DoTA with SSP	**0.0466**	0.0303	**0.0445**	0.0245	**0.0479**	0.0394
Absolute Style Confidence						
Cosine Similarity	0.3568	0.3690	0.4121	0.3707	0.3273	**0.3241**
DoTA wo SSP	**0.4752**	**0.3234**	0.4439	**0.3543**	0.5278	0.3373
DoTA with SSP	0.4510	0.3318	**0.4927**	0.3674	**0.5641**	0.3534

(a) 50, 5

Method	Semantics		Structure		Style	
	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)
Relative Style Confidence						
Cosine Similarity	0.0230	0.0183	0.0201	0.0259	0.0235	0.0201
DoTA wo SSP	0.0295	**0.0168**	0.0215	0.0263	0.0256	**0.0176**
DoTA with SSP	**0.0326**	0.0259	**0.0287**	**0.0256**	**0.0348**	0.0219
Absolute Style Confidence						
Cosine Similarity	0.3198	0.3375	0.3480	0.3508	0.3612	0.3394
DoTA wo SSP	0.3721	**0.3198**	0.3466	0.3510	0.3692	**0.3272**
DoTA with SSP	**0.3788**	0.3304	**0.3972**	**0.3276**	**0.3814**	0.3368

(b) 50, 10

Method	Semantics		Structure		Style	
	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)
Relative Style Confidence						
Cosine Similarity	0.0244	**0.0311**	0.0212	0.0295	**0.0351**	0.0282
DoTA wo SSP	0.0191	0.0346	**0.0335**	0.0261	0.0220	0.0304
DoTA with SSP	**0.0309**	0.0331	0.0210	**0.0193**	0.0216	**0.0270**
Absolute Style Confidence						
Cosine Similarity	0.3316	**0.3455**	0.3454	**0.3253**	**0.3763**	0.3833
DoTA wo SSP	0.3284	0.3534	**0.3756**	0.3546	0.3449	0.3836
DoTA with SSP	**0.3453**	0.3745	0.3249	0.3370	0.3479	**0.3377**

(c) 400, 5

Method	Semantics		Structure		Style	
	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)	Take-K (↑)	Leave-K (↓)
Relative Style Confidence						
Cosine Similarity	0.0189	0.0297	0.0242	**0.0165**	0.0212	**0.0262**
DoTA wo SSP	0.0194	0.0316	**0.0335**	0.0261	**0.0344**	0.0331
DoTA with SSP	**0.0211**	**0.0244**	0.0211	0.0307	0.0210	0.0405
Absolute Style Confidence						
Cosine Similarity	0.2977	0.3466	0.3591	**0.3132**	0.3360	**0.3484**
DoTA wo SSP	0.3229	0.3861	**0.3756**	0.3546	**0.3790**	0.3511
DoTA with SSP	**0.3468**	**0.3394**	0.3386	0.3360	0.3444	0.4001

(d) 400, 10

Table 1. **Ablation of counterfactual evaluation of attribution methods on the Ukiyo-e dataset**. We highlight the best result for each counterfactual setting for a latent representation with blue color. The cases where the style confidence for Leave-K is higher than Take-K are highlighted with red color. We present the results with different combinations of representation reduction (50 and 400) and K (5 and 10).

Method	PopArt					
	Semantics		Structure		Style	
	Time	Memory	Time	Memory	Time	Memory
Cosine	4763.84	1189.45	5429.43	1291.87	7542.74	1558.83
DoTA wo SSP	1592.12	324.58	1624.71	517.42	1973.91	735.41
DoTA with SSP	11.73	549.87	14.33	820.31	35.41	1071.87
Method	Ukiyo-e					
	Semantics		Structure		Style	
	Time	Memory	Time	Memory	Time	Memory
Cosine	10122.66	1427.83	12728.32	1768.72	14327.22	2136.42
DoTA wo SSP	1918.36	595.79	1925.78	594.81	1900.81	786.32
DoTA with SSP	12.09	779.11	13.94	780.02	451.04	1093.56
Method	Munch					
	Semantics		Structure		Style	
	Time	Memory	Time	Memory	Time	Memory
Cosine	35871.21	5874.73	37287.92	6174.87	45879.12	6843.41
DoTA wo SSP	19164.86	769.45	20831.71	1377.36	21536.87	1591.11
DoTA with SSP	894.07	944.32	978.87	1527.32	1137.28	1879.23

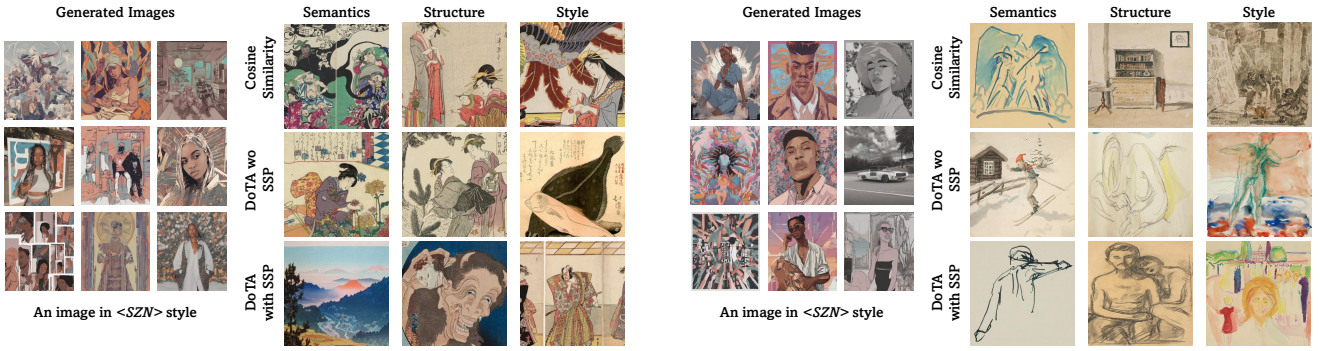
(a)

Method	PopArt					
	Semantics		Structure		Style	
	Time	Memory	Time	Memory	Time	Memory
Cosine	4959.41	1391.31	5715.33	1529.74	7903.79	1864.37
DoTA wo SSP	1734.63	461.14	1894.87	525.37	2287.98	843.25
DoTA with SSP	15.49	679.81	19.64	601.29	49.13	1017.47
Method	Ukiyo-e					
	Semantics		Structure		Style	
	Time	Memory	Time	Memory	Time	Memory
Cosine	12417.31	1766.24	15468.72	1817.94	19923.71	2641.33
DoTA wo SSP	1987.27	712.48	2131.42	582.87	2459.29	870.37
DoTA with SSP	14.72	840.21	18.34	870.71	67.17	964.34
Method	Munch					
	Semantics		Structure		Style	
	Time	Memory	Time	Memory	Time	Memory
Cosine	37513.86	5985.47	3953.51	6321.81	46713.89	6971.37
DoTA wo SSP	2003.62	920.39	2207.19	1421.07	2374.41	1714.99
DoTA with SSP	912.71	1197.89	997.84	1821.00	1194.78	1941.17

(b)

Table 2. **Attribution time and run time memory allocation (a) without and (b) with adversarial attack**. We report the results for attribution performed with with representation dimensions reduced to 50. We notice a significant drop in the time taken for the attribution due to inclusion of proposed SSP technique with a slight trade-off with memory allocation. We report the time in seconds and memory in megabytes.

of K. Such a performance degradation could be attributed to the curse of dimensionality and distance concentration discussed earlier. On a given representation size, increasing the value of K (from 5 to 10) increases the number of instances where DoTA with SSP outperforms the other two techniques. It also indicates the need to perform a hyperparameter search on K to minimize the penalty on informative signals. We present qualitative examples in Figure 5 and report the run-time and memory usage for attribution in Table 2. We observe a consistent and significant improvement in computation time when incorporating the SSP module into DoTA across all datasets and representations. On average, DoTA with SSP achieves a reduction of over **99%** in computation time compared to DoTA without SSP, corresponding to savings of two to three orders of magnitude across PopArt and Ukiyo-e, and approximately



(a) Ukiyo-e (b) Munch

Figure 5. Qualitative comparison of different approaches for data attribution on the (a) Ukiyo-e and (b) Munch datasets. We present the **best attributed image** with each method across representations. The best attributed images with DoTA effectively captures the semantics, structure, and style present in the generated images.

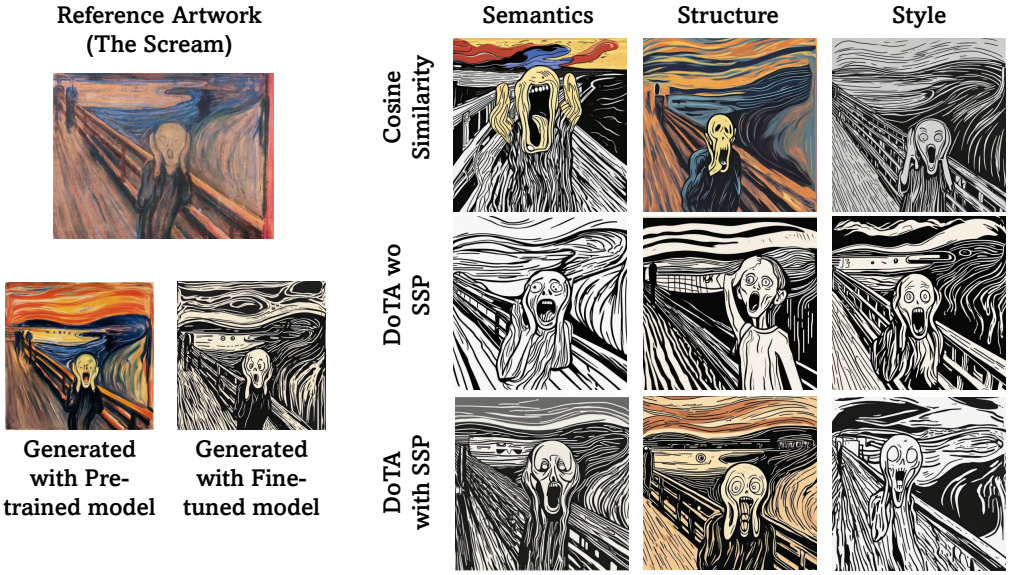


Figure 6. Qualitative comparison of attribution approaches under adversarial attack with different latent representations for the PopArt dataset. We present the **images generated for "The Scream" after retraining with Take-K under adversarial attack** in different settings. We observe maximum impact of the attack with the cosine similarity based attribution showing drastic changes in semantics, structure, and style in the respective generated images.

95–97% on the larger Munch dataset. Even when compared to cosine similarity, which is inherently faster than DoTA without SSP, the SSP-enhanced variant consistently demonstrates substantial efficiency gains, saving **98–99%** time on PopArt and Ukiyo-e and about **80–85%** on Munch. These improvements establish SSP as a critical enabler for scaling Shapley-value–based data attribution, offering near-real-time attribution while preserving attribution fidelity.

The additional time and memory analysis indicates that the increase in memory footprint for DoTA with SSP compared to DoTA without SSP arises from the internal mechanics of the search space pruning stage. Specifically, during Algorithm 2 (Data Subset Selection), a large number of random subsets S are generated, and their corresponding Energy Distance (ED) values are computed and stored. These stored ED values are critical for applying the pruning threshold δ_{ED}, which determines whether a subset should be retained or discarded. Subsequently, Algorithm 3 (Search Space Pruning) requires access to the complete set of subset–ED mappings to consistently apply the pruning criteria across multiple iterations. Furthermore, in Algorithm 4 (Ridge Regression for Shapley Value Estimation), these mappings are used as inputs for the regression stage, where subset contributions are weighted according to their

ED values. As a result, the need to maintain and reuse this expanded subset–ED mapping throughout the entire SSP pipeline results in higher memory consumption than in the non-SSP setting, where such intensive simulation and storage are not required.

We present the images generated for "The Scream" after retraining with Take-K under adversarial attack across different in Figure 6. For DoTA-based methods, the generated outputs remain highly consistent with those produced by the fine-tuned model across all three latent representations, preserving the underlying semantics, structure, and style of the original artwork. In contrast, cosine similarity-based attribution shows a pronounced vulnerability to the adversarial intrusion, with the corresponding generated outputs deviating significantly from those of the fine-tuned model. Specifically, cosine-based attribution fails to maintain semantic coherence, structural integrity, and stylistic fidelity, highlighting the instability of this approach under adversarial conditions. These observations reinforce the robustness of DoTA in preserving perceptual and semantic alignment, even in the presence of adversarial perturbations.

References

[1] Rinon Gal, Yuval Alaluf, Matan Atzmon, Or Patashnik, Hila Chefer, Cordelia Schmid, Shai Shalev, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *CVPR*, 2023. 3

[2] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 4

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[4] Edward J. Hu et al. Lora: Low-rank adaptation of large language models, 2021. 3

[5] Nupur Kumari, Aniruddha Singh, Pengchuan Wang, Agata Lapedriza, Kihwan Kim, Joshua Tenenbaum, and Arun Mallya. Custom diffusion: Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2306.00941*, 2023. 3

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[7] Maxime Oquab, Tete Darcet, Theo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. In *CVPR*, 2023. 4

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[9] Dana Roich, Rinon Gal, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 3

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2, 3

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 2

[14] Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. 2004. 4, 5

[15] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2