# UNCOVERING CONCEPTUAL GEOMETRY IN DIFFUSION MODELS WITH SPATIALLY INVARIANT SIMILARITY

**Anonymous authors**
Paper under double-blind review

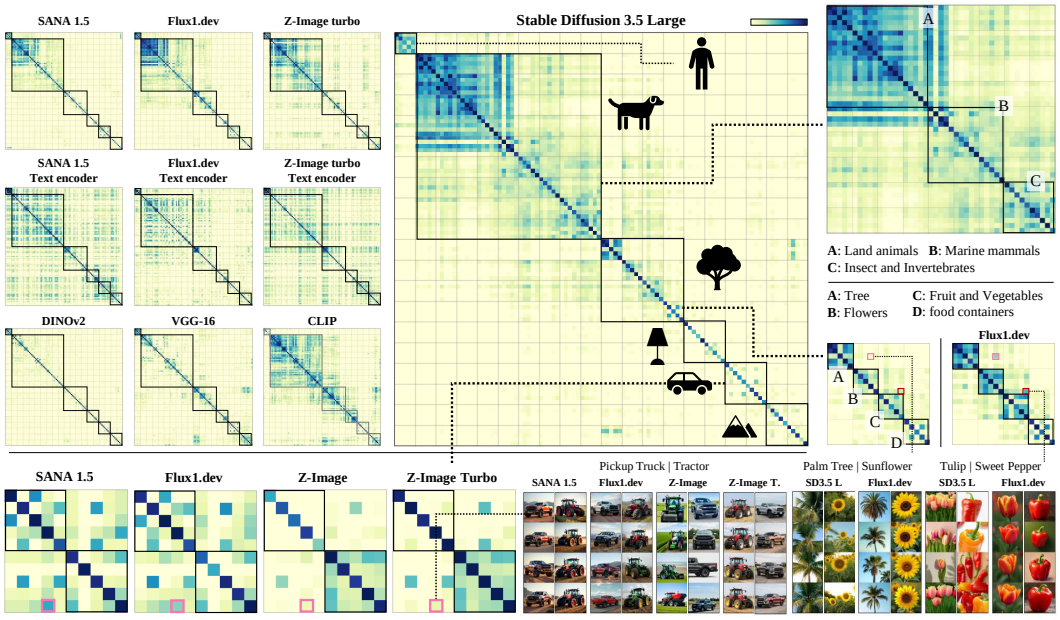



Figure 1: **Conceptual geometry revealed by Testing with Spatially Invariant Similarity (TSIS).** TSIS uncovers structured relationships among visual concepts in pretrained diffusion models, revealing both shared and model-specific organization. The resulting geometries can be compared with conditioning and pretrained vision representations and inspected at the level of individual concept relationships and generated examples.

## ABSTRACT

How do diffusion models organize the visual world they learn to generate? We introduce *Spatially Invariant Similarity (SIS)*, a training-free and model-agnostic method for comparing intermediate diffusion representations despite variations in spatial configuration. SIS establishes patch-wise feature correspondences and preserves the spatial evidence underlying their similarity. We further introduce Testing with SIS (TSIS), which aggregates these correspondences across generations to reveal concept-level relational structure. Across diverse pretrained diffusion models, TSIS uncovers coherent conceptual geometries: semantically related object categories exhibit shared coarse organization while retaining model-specific relationships. Extending beyond object categories, we show that color, material, and viewpoint exhibit distinct block–timestep trajectories, revealing where and when different visual factors become distinguishable during generation. Finally, generations that deviate from the reference geometry in intermediate representations consistently produce rarer and less similar final images. These results establish SIS and TSIS as tools for studying both the organization and evolution of conceptual structure within diffusion models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

How do diffusion models organize the visual world they have learned to generate? Recent advances in generative modeling increasingly suggest that generation and understanding are deeply intertwined. Image generators trained for synthesis can exhibit broad visual understanding capabilities (Gabeur et al., 2026), while recent models explicitly couple image generation and understanding within a shared generative process (Le et al., 2026). These developments raise a fundamental question: if a generative model can both produce and reason about diverse visual concepts, how are those concepts organized within its internal representation?

For humans, similarity is a fundamental mechanism for structuring perceptual and conceptual knowledge (Tversky, 1977). Concepts can be viewed as occupying positions in a psychological space, where their relative distances reflect perceived similarity and support categorization and generalization (Nosofsky, 1986). This motivates us to ask an analogous question of diffusion models: *what conceptual geometry emerges from representations learned through generation?*

Diffusion models provide a compelling setting for this question. Their internal features have been successfully repurposed for semantic correspondence (Tang et al., 2023; Luo et al., 2023), depth estimation (Ke et al., 2024), pose estimation, and image manipulation (Tumanyan et al., 2023). Other studies have identified semantically meaningful latent directions and manipulated or removed specific concepts and attributes (Kwon et al., 2022; Gandikota et al., 2023; 2024). These findings show that diffusion models encode rich and actionable semantic information.

However, identifying individual concepts does not reveal how they are organized *relative to one another*. Existing work largely focuses on locating, manipulating, or matching concepts and spatial regions, providing limited insight into the broader relational structure of the representation. For instance, correspondence methods may reveal matching parts across two images, but do not directly tell us whether a model represents a dog as closer to a wolf than to a cat, or how viewpoint and material affect these relationships. Thus, the *conceptual geometry* formed by relationships among diverse concepts remains underexplored.

In this paper, we introduce *Spatially Invariant Similarity (SIS)*, a method for comparing internal diffusion representations while accounting for spatial variation during generation. Because the same concept can appear at different positions, scales, and poses, directly comparing spatially aligned features can confound semantic similarity with spatial arrangement. SIS instead compares feature evidence across spatial positions and retains a *spatial similarity map*. Aggregating this map produces a global similarity score, while the map itself reveals the spatial evidence underlying the measured relationship. Using SIS, we establish spatially invariant correspondences between individual generations and retain the local evidence underlying their similarity. We further introduce Testing with SIS (TSIS), which aggregates SIS across multiple generations and both comparison directions to obtain robust concept-level similarity scores. Using TSIS, we construct pairwise affinity matrices over predefined concept sets and analyze the conceptual geometries formed within different pretrained diffusion models. We compare these geometries across models and model families, examine their organization across diffusion blocks and denoising timesteps, and extend the analysis beyond object categories to prompt-definable visual factors such as color, material, and viewpoint. Our contributions are threefold:

- We introduce *Spatially Invariant Similarity (SIS)*, a method for comparing spatial diffusion representations without requiring positional alignment. SIS establishes patch-wise correspondences across generations while retaining the spatial evidence underlying their similarity.
- We introduce *Testing with SIS (TSIS)*, which aggregates SIS across multiple generations and both comparison directions to obtain robust concept-level similarities. These pairwise similarities define a model's *conceptual geometry*, providing a model-agnostic and training-free framework for comparing internal organization across pretrained diffusion models.
- Using TSIS, we show that diffusion models exhibit structured conceptual geometries across both object categories and prompt-definable visual factors, including color, material, and viewpoint. We further characterize where and when these relationships emerge across diffusion blocks and denoising timesteps, while SIS maps provide local spatial evidence for individual concept relationships.

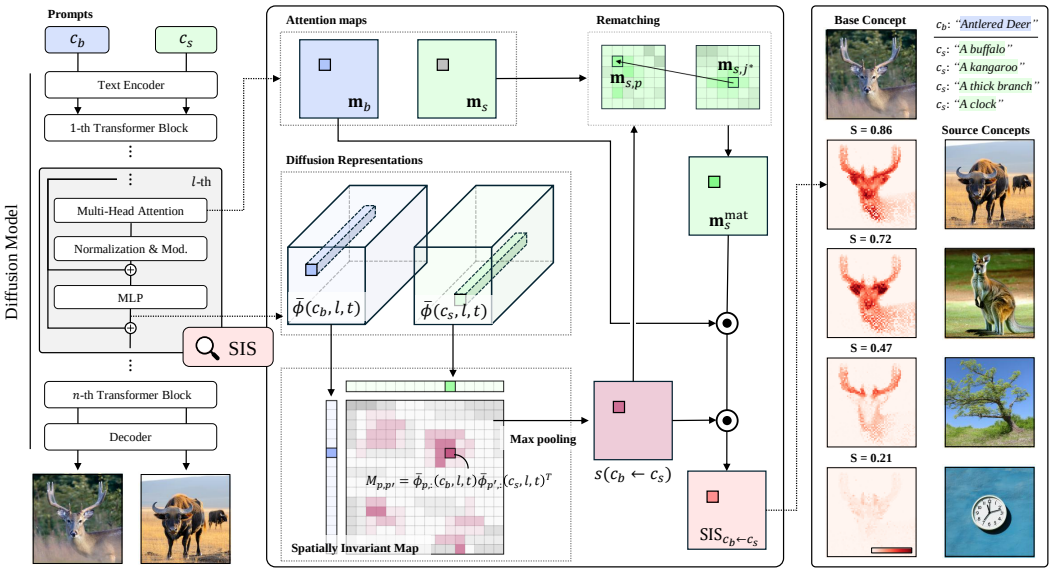



Figure 2: **Overview of Spatially Invariant Similarity (SIS).** For base and source concepts $c_b$ and $c_s$, we extract intermediate diffusion representations and concept-specific attention maps at block $l$ and timestep $t$. SIS computes all-to-all spatial feature similarities and matches each base token to its most similar source token, yielding the directional map $\mathbf{s}(c_b \leftarrow c_s)$. The resulting correspondence is also used to rematch the source attention map, and the base attention, matched source attention, and spatial similarity are combined to produce the final SIS map. Right: illustrative generated examples and their SIS maps.

## 2 PRELIMINARIES

Modern text-to-image diffusion models employ different architectures to integrate textual conditions with visual representations. In U-Net-based models, text and image tokens typically interact through cross-attention, whereas transformer-based architectures such as MMDiT jointly process multimodal tokens through attention. Despite these architectural differences, both can be viewed as repeatedly updating visual tokens through attention-based interaction blocks and residual connections.

Let $\mathbf{X}^{l,t} \in \mathbb{R}^{P \times Ch}$ denote the image tokens at block $l$ and denoising timestep $t$, and let $\mathbf{C} \in \mathbb{R}^{T \times Ch}$ denote the text tokens corresponding to prompt $c$. We abstract an interaction block as

$$\Delta \mathbf{X}^{l,t} = \mathcal{F}^{l,t}(\mathbf{X}^{l,t}, \mathbf{C}), \tag{1}$$

where $\mathcal{F}$ represents the architecture-specific computation involving attention and subsequent transformations. The corresponding attention operation $\mathcal{A}$ produces token-wise interaction weights,

$$\mathbf{A}^{l,t} = \mathcal{A}^{l,t}(\mathbf{X}^{l,t}, \mathbf{C}), \tag{2}$$

which may include text–image as well as image–image interactions depending on the architecture. From these weights, we derive a concept-specific spatial attention map $\mathbf{m}_c(l,t) \in \mathbb{R}^P$ that indicates where the prompt-related information is expressed across image tokens.

The diffusion representation is then updated through a residual connection,

$$\mathbf{X}^{l+1,t} = \mathbf{X}^{l,t} + \Delta \mathbf{X}^{l,t}. \tag{3}$$

Rather than using the updated token stream $\mathbf{X}^{l+1,t}$, we focus on the *residual update* $\Delta \mathbf{X}^{l,t}$ immediately before it is added back to the main stream. This term captures the information introduced by the interaction block itself, without mixing it with the incoming representation (Surkov et al., 2025). We therefore define

$$\phi(c, l, t) = \Delta \mathbf{X}^{l,t}. \tag{4}$$

We regard $\phi(c,l,t)$ as the block-level representation of the concept specified by prompt $c$, and $\mathbf{m}_c(l,t)$ as its corresponding spatial importance map. These two quantities form the basis of the similarity computation introduced in the following section.

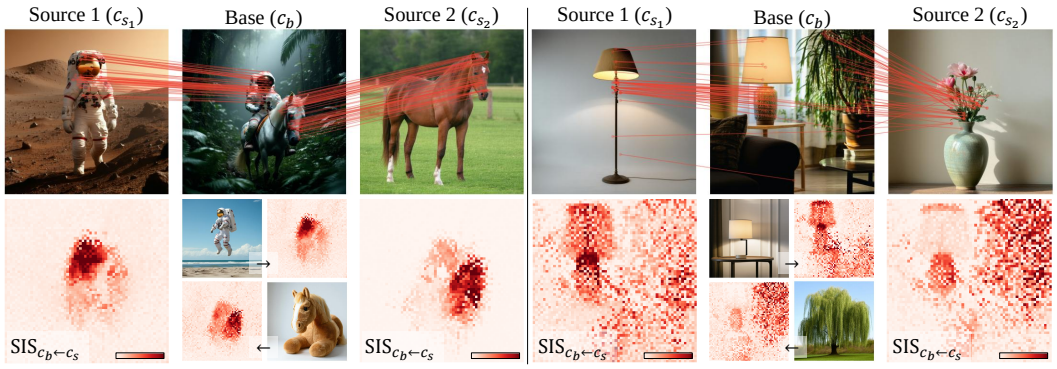



Figure 3: Illustrative SIS maps in Stable Diffusion 3.5 Large ($l = 24, t = 16$). Each example compares a base generation $c_b$ with various source generations. We visualize the top 50 patch correspondences ranked by their SIS contribution, with lines connecting matched spatial locations. Generation prompts and additional SIS maps are available in Appendix D and A, respectively.

## 3 Spatially Invariant Similarity (SIS)

Let $c_b$ and $c_s$ denote the base and source prompts, respectively. Given a diffusion block $l$ and denoising timestep $t$, we denote their corresponding diffusion representations as $\phi(c_b, l, t), \phi(c_s, l, t) \in \mathbb{R}^{P \times Ch}$, where $P$ is the number of spatial tokens and $Ch$ is the feature dimension. We first normalize each spatial feature along the channel dimension:

$$\bar{\phi}_{p,:}(c, l, t) = \phi_{p,:}(c, l, t)/||\phi_{p,:}(c, l, t)||_2. \tag{5}$$

We then construct a pairwise similarity matrix

$$M(c_b, c_s, l, t) = \bar{\phi}(c_b, l, t)\bar{\phi}(c_s, l, t)^\top \in \mathbb{R}^{P \times P}, \tag{6}$$

where each element

$$M_{p,p'}(c_b, c_s, l, t) = \bar{\phi}_{p,:}(c_b, l, t)\bar{\phi}_{p',:}(c_s, l, t)^\top \tag{7}$$

measures the cosine similarity between spatial location $p$ in the base representation and location $p'$ in the source representation. For notational simplicity, we omit $(l, t)$ in the following.

Since corresponding concepts need not appear at the same spatial locations across different generations, we do not enforce position-wise alignment. Instead, for each base location $p$, we identify the most similar source location:

$$j_p = \arg\max_{p' \in 1,\ldots,P} M_{p,p'}(c_b, c_s), \tag{8}$$

and define the spatially invariant map as

$$s_p(c_b \leftarrow c_s) = M_{p,j_p}(c_b, c_s) = \max_{p'} M_{p,p'}(c_b, c_s). \tag{9}$$

This produces a similarity map $\mathbf{s}(c_b \leftarrow c_s) \in \mathbb{R}^P$, where each base location is matched to its most similar source feature regardless of spatial position.

**Attention-Weighted Similarity.** While $\mathbf{s}(c_b \leftarrow c_s)$ captures spatially invariant feature correspondence, not all spatial locations contribute equally to the generation of the target concepts. We therefore incorporate the corresponding text-to-image attention maps as spatial importance weights. Let $\mathbf{m}_b, \mathbf{m}_s \in \mathbb{R}^P$ denote the attention maps[1] associated with $c_b$ and $c_s$, respectively.

Since each base position $p$ is matched to source position $j_p$, we first rematch the original source attention map using the same feature correspondence:

$$\mathbf{m}^{\text{mat}}_{s,p} = \mathbf{m}_{s,j_p}. \tag{10}$$

[1] Attention maps are averaged across all attention heads.

The rematching is performed before normalization so that the source attention is first expressed in the spatial layout induced by the base-to-source feature correspondence. We then independently normalize the base attention map and the rematched source attention map:

$$\bar{\mathbf{m}}_b = \mathbf{m}_b / \|\mathbf{m}_b\|_2, \quad \bar{\mathbf{m}}_s^{\text{mat}} = \mathbf{m}_s^{\text{mat}} / \|\mathbf{m}_s^{\text{mat}}\|_2. \quad (11)$$

We define the weighted *Spatially Invariant Similarity* (SIS) map as

$$\text{SIS}_{c_b \leftarrow c_s} = \bar{\mathbf{m}}_b \odot \mathbf{s}(c_b \leftarrow c_s) \odot \bar{\mathbf{m}}_s^{\text{mat}}, \quad (12)$$

where $\odot$ denotes element-wise multiplication. The resulting $\text{SIS}_{c_b \leftarrow c_s} \in \mathbb{R}^P$ retains spatial evidence from the base representation while assigning each location its strongest feature correspondence in the source representation. This allows SIS to capture concept similarity independently of absolute spatial arrangement while emphasizing regions relevant to both concepts. Figure 3 shows the illustrative examples of SIS for the given base and source concepts.

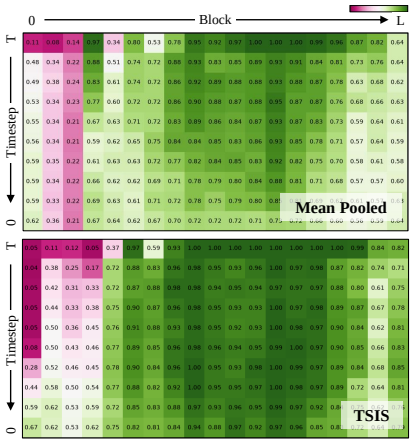



Figure 4: **Benefit of spatial correspondence.** Self-classification accuracy across blocks and timesteps using globally mean-pooled diffusion features (**top**) and TSIS (**bottom**). TSIS preserves stronger concept separability across a broader range of intermediate representations.

**Testing with SIS (TSIS).** To obtain a robust concept-level similarity, we generate multiple samples from the same base and source prompts, $c_b$ and $c_s$. Let $\phi^{(i)}(c_b, l, t)$ and $\phi^{(j)}(c_s, l, t)$ denote the representations from the $i$-th and $j$-th generations of $N_b$ and $N_s$, respectively. We compute SIS for every generation pair and obtain the directional concept-level similarity by averaging over all generation pairs and spatially aggregating the resulting maps:

$$S_{c_b \leftarrow c_s} = \frac{1}{N_b N_s} \sum_{i=1}^{N_b} \sum_{j=1}^{N_s} \sum_{p=1}^{P} \text{SIS}_{c_b \leftarrow c_s, p}^{(i,j)}. \quad (13)$$

The final bidirectional similarity is

$$\text{TSIS}(c_b, c_s) = \frac{1}{2} \left( S_{c_b \leftarrow c_s} + S_{c_s \leftarrow c_b} \right). \quad (14)$$

Figure 2 shows the entire procedure of TSIS and illustrative examples.

**Why Spatial Correspondence?** We additionally examine whether preserving spatial correspondence provides more concept-discriminative representations than global average pooling ($\phi_{\text{GAP}} = \frac{1}{P} \sum_{p=1}^{P} \phi_{p,:}$) in Stable Diffusion 3.5 Large (Esser et al., 2024) with the subset of CIFAR-100 labels. Figure 4 compares TSIS with globally mean-pooled diffusion features using the self-classification evaluation introduced in Sec. 4 and detailed in Appendix C. Across blocks and denoising timesteps, TSIS yields higher classification accuracy over a broader region, suggesting that spatial correspondence preserves concept-level structure that is otherwise weakened by global average pooling.

## 4 Experiments

We use TSIS to probe conceptual geometry at two levels: object categories and controlled visual attributes (color, material, and viewpoint). Across experiments, we compare diffusion models with diverse architectures, scales, and acceleration strategies.

**Models.** We evaluate Stable Diffusion 3.5 Large (SD3.5 L), Medium (SD3.5 M), and Large Turbo (SD3.5 LT) (Esser et al., 2024); FLUX.1-dev and FLUX.1-schnell (Labs, 2024); PixArt-$\alpha$ (Chen et al., 2023); Z-Image and Z-Image-Turbo (Z-Image Team, 2025); SANA-1.5 (Xie et al., 2025); Lumina-Image 2.0 (Lumina 2) (Qin et al., 2025); Kandinsky 3 (Kand. 3) (Arkhipkin et al., 2024); SDXL (Podell et al., 2024); and Stable Diffusion 1.5 (SD1.5) (Rombach et al., 2022). Generation settings are provided in Appendix D.

**Representation Selection.** Because diffusion representations vary across blocks and denoising timesteps, we select $(l, t)$ for each model by grid search. At each $(l, t)$, we perform 5-fold self-class prediction using TSIS and select the pair with the highest cross-validated accuracy. Details are provided in Appendix C.

Table 1: Comparison of diffusion-model affinity matrices with external vision representations ( CLIP , VGG16 , and DINOv2 ). L2 measures the Euclidean distance between upper-triangular affinity entries, whereas CS measures their cosine similarity. Lower L2 and higher CS indicate greater agreement with the corresponding vision representation.

| Arch. | Transformer | | | | | | | | | | U-Net | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SD3.5 L | SD3.5 M | SD3.5 LT | Lumina 2 | Z-Image | Z-Image Turbo | SANA 1.5 | FLUX Dev | FLUX Schnell | PixArt $\alpha$ | Kand. 3 | SDXL 1.0 | SD 1.5 |
| L2 (↓) | **57.60** | **58.23** | **58.41** | 75.43 | **62.62** | **62.04** | 62.99 | **57.93** | **56.81** | 64.91 | **60.52** | 72.38 | **61.31** |
| L2 (↓) | 70.55 | 73.11 | 70.99 | **69.88** | 75.21 | 73.80 | 68.19 | 68.57 | 68.18 | **61.76** | 64.90 | **68.29** | 66.53 |
| L2 (↓) | 64.16 | 71.86 | 64.28 | 77.48 | 68.60 | 69.08 | **60.54** | 64.57 | 64.26 | 63.54 | 66.04 | 69.94 | 62.37 |
| CS (↑) | **0.66** | **0.66** | **0.66** | 0.43 | **0.60** | **0.61** | 0.60 | **0.66** | **0.67** | 0.57 | **0.63** | 0.47 | **0.62** |
| CS (↑) | 0.50 | 0.46 | 0.49 | **0.51** | 0.43 | 0.45 | 0.53 | 0.53 | 0.53 | **0.61** | 0.57 | 0.53 | 0.55 |
| CS (↑) | 0.58 | 0.48 | 0.58 | 0.39 | 0.52 | 0.52 | **0.63** | 0.58 | 0.58 | 0.57 | 0.51 | **0.61** | **0.62** |

**Conceptual Geometry.** For a concept set $\mathcal{C} = \{c_1, \ldots, c_n\}$, we define its affinity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ as

$$\mathbf{A}_{i,j} = \mathrm{TSIS}(c_i, c_j) = \mathbf{A}_{j,i}. \tag{15}$$

For visualization and cross-model comparison, we Z-normalize the affinity values within each model. We additionally clip negative values to zero, suppressing below-average affinities to emphasize stronger concept relationships for clear visualization.

### 4.1 Class-Level Conceptual Geometry

We first examine class-level conceptual geometry using the 100 class names of CIFAR-100 (Krizhevsky & Hinton, 2009). For each class, we generate 20 images using the prompt template *"a photo of {class}"* and construct the pairwise TSIS affinity matrix. We order the classes by the CIFAR-100 superclass structure to facilitate interpretation; the complete class ordering is provided in Appendix E.

**Conceptual geometry across diffusion models.** Figure 1 shows representative affinity matrices across diffusion models. Entire results are available in Appendix G. Despite substantial differences in architecture, scale, and training strategy, the models exhibit broadly similar class-level organization. The consistency is particularly pronounced among biological categories, where semantically related classes form coherent blocks of high affinity. At the same time, the geometries are not identical: individual relationships and group boundaries vary across models, as illustrated by the plant-related categories in Figure 1. Thus, diffusion models appear to share a coarse conceptual organization while retaining model-specific relational structure.

**Comparison with text and vision representations.** We next ask whether this geometry can be explained by the representations used to condition or visually evaluate the models. For each conditioning text encoder, we embed the same class prompts and construct its pairwise affinity matrix. We similarly construct visual affinity matrices using DINOv2 (Oquab et al., 2024), VGG16 (Simonyan & Zisserman, 2015), and CLIP (Radford et al., 2021); image features are extracted from the generated samples and averaged across diffusion models to obtain a common geometry for each vision encoder.

The conditioning text geometries show no clear correspondence to the geometry of their associated diffusion models (Figure 1), suggesting that the relational structure emerging inside the diffusion process is not simply inherited from the text embedding space. The vision baselines reveal a different contrast. DINOv2 and VGG16 exhibit strong within-class structure, but substantially weaker affinities across semantically related classes. In comparison, CLIP and the diffusion models show broader group-level organization, including the coherent biological regions visible in Figure 1. We quantify these differences in Table 1 by comparing the upper-triangular entries of each diffusion affinity matrix with those of the three vision baselines using Euclidean distance (L2) and cosine similarity (CS). For almost of evaluated models, diffusion geometries are closer to CLIP than to DINOv2 or VGG16, indicating greater agreement with a vision-language representation than with purely visual representations. As a result, these observations suggest that diffusion representations

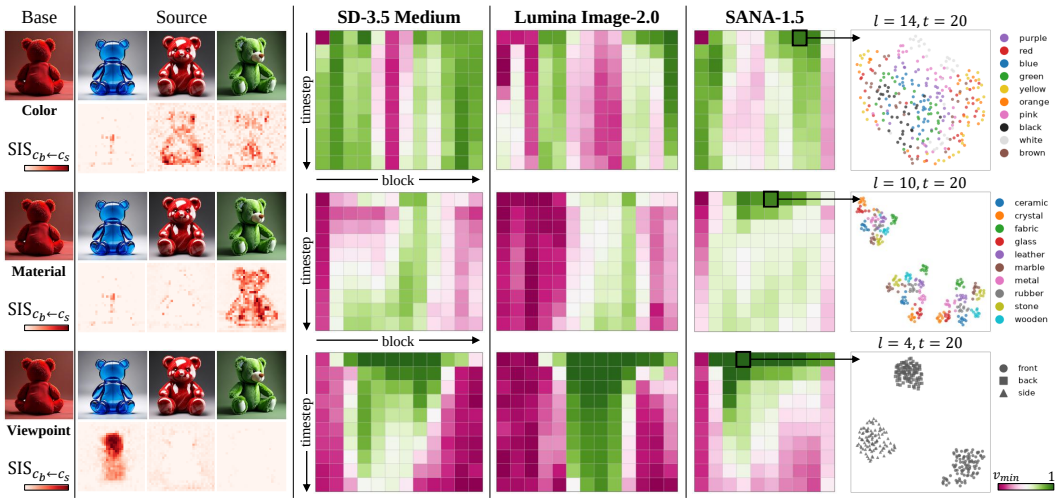



Figure 5: Visual-level conceptual geometry across color, material, and viewpoint. **Left:** Representative generations under controlled attribute variations. **Center:** TSIS self-classification accuracy across diffusion blocks and denoising timesteps for SD3.5 M, Lumina 2, and SANA-1.5. Higher accuracy (green) indicates representations in which the corresponding attribute is more distinguishable; $v_{\max} = 1$ and $v_{\min}$ is set to random-chance accuracy ($1/10$ for color and material and $1/3$ for viewpoint). **Right:** UMAP projections of SANA-1.5 TSIS affinity matrices at representative high-accuracy $(l, t)$ locations, illustrating the relational organization of each visual attribute.

encode relational structure beyond both their conditioning text space and purely visual class discrimination. Their greater agreement with CLIP may reflect the role of cross-modal supervision, although establishing its origin requires further investigation.

### 4.2 Visual-Level Conceptual Geometry

While the previous experiment examines conceptual geometry at the level of coarse object categories, we next consider finer-grained and more abstract visual factors: color, material, and viewpoint. Beyond asking whether these attributes form meaningful relational structures, we investigate *where and when* they become distinguishable during the diffusion process. This analysis is motivated by prior findings that different blocks and denoising timesteps play distinct roles in image generation (Surkov et al., 2025; Tinaz et al., 2025). To control object identity, we use *teddy bear* as the common object and generate all combinations of 10 colors, 10 materials, and 3 viewpoints:

$$\text{“a photo of } \{color\}\ \{material\}\ \text{teddy bear, } \{view\}\ \text{view.”} \tag{16}$$

This yields 300 generations, allowing each attribute to be analyzed while treating the other two as nuisance factors. At every block and denoising timestep $(l, t)$, we apply the TSIS self-classification procedure described in Appendix C. We use the resulting classification accuracy as a measure of how distinctly an attribute is represented at that location in the diffusion process: higher accuracy indicates a stronger attribute-specific generative role.

Figure 5 visualizes these block–timestep landscapes for three diffusion models. Although the exact locations of high-accuracy regions vary across models, similar attribute-specific patterns emerge. Color remains highly distinguishable across a broad range of denoising timesteps, with stronger roles concentrated toward the earlier and later blocks. Viewpoint exhibits a different trajectory: at early denoising timesteps, viewpoint information is distributed broadly across blocks, whereas at later timesteps its discriminative region becomes increasingly concentrated in intermediate blocks. Material follows another pattern, with its strongest regions overlapping substantially with those associated with viewpoint. These results suggest that different visual attributes follow distinct block–timestep trajectories during generation, while the overall patterns are partially shared across model architectures. We further examine the relational structure associated with these generative roles using SANA-1.5. For each attribute, we construct the TSIS affinity matrix at a representative high-accuracy $(l, t)$ and treat each row as a relational representation of the corresponding generation. We

Table 2: Image-level evaluation of the Top and Bottom TSIS groups across visual attributes and object categories. For each category, 500 generated images are fixed as the reference set. RS (↑ is better) measures rarity relative to this reference distribution, with the number of detected outliers among the 30 samples reported in parentheses. CLIP similarity (CS) (↓ is better) is the average inner product between the pooled CLIP embedding of each candidate image and those of the same 500 reference images.

| | | Vase | | Handbag | | Hat | | Teapot | |
|---|---|---|---|---|---|---|---|---|---|
| Attr. | Group | RS (↑) | CS (↓) | RS (↑) | CS (↓) | RS (↑) | CS (↓) | RS (↑) | CS (↓) |
| Color | Top | 5.10 (0) | 0.86 | 5.76 (0) | 0.79 | 5.56 (0) | 0.86 | 3.41 (0) | 0.87 |
| | Bttm | **7.89** (13) | **0.80** | **7.89** (19) | **0.74** | **8.80** (14) | **0.74** | **3.66** (4) | **0.85** |
| Material | Top | 3.23 (0) | 0.86 | 3.49 (0) | 0.80 | 2.98 (0) | 0.85 | 5.21 (0) | 0.88 |
| | Bttm | **3.69** (8) | **0.80** | **4.33** (9) | **0.73** | **4.52** (9) | **0.69** | **5.50** (6) | **0.86** |
| View | Top | 2.33 (0) | 0.87 | 3.07 (0) | 0.78 | 3.20 (0) | 0.86 | 5.33 (0) | 0.87 |
| | Bttm | **4.27** (12) | **0.76** | **5.10** (9) | **0.72** | **5.58** (10) | **0.69** | **5.63** (5) | **0.85** |

then project these representations using UMAP (McInnes et al., 2018). As shown on the right of Figure 5, color and viewpoint exhibit clear attribute-dependent organization, with samples sharing the same attribute values forming distinct groups. Material remains less clearly separated and is substantially entangled with viewpoint (See Figure 11 in Appendix H. This is consistent with their overlapping block–timestep regions, suggesting that visual factors with similar generative roles may also exhibit more entangled relational geometry.

### 4.3 From Representational Atypicality to Rare Generations

What can we learn from the conceptual geometry encoded in internal diffusion representations? If this geometry captures the organization of typical generations, samples that lie farther from this organization may correspond to unusual or rare outcomes in the final image space. We test this connection by asking whether representational atypicality measured during generation is preserved in the completed image. For each object category, we generate 1,500 samples and fix 500 as a reference set. For each of the remaining candidate generations $x$, we measure its similarity to the reference set using TSIS:

$$T(x) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{TSIS}(x, x_i), \tag{17}$$

where $N = 500$ and $x_i$ denotes a reference sample. We perform this comparison separately for color, material, and viewpoint, using the attribute-specific representations identified in Sec. 4.2. A lower $T(x)$ therefore indicates greater deviation from the reference geometry with respect to the corresponding visual attribute.

We rank the 1,000 candidate generations by $T(x)$ and compare the 30 highest-scoring samples (*Top*) with the 30 lowest-scoring samples (*Bottom*). The ranking is determined entirely from intermediate diffusion representations. After generation is completed, we independently evaluate the final images against the same reference set using two image-space measures. Rarity Score (RS) (Han et al., 2022) measures how uncommon a candidate is relative to the reference distribution,[2] while CLIP similarity (CS) measures the average inner product between the pooled CLIP embedding of each candidate and those of the 500 reference images. Thus, higher RS and lower CS both indicate greater deviation in the final image space.

Table 2 shows that the Bottom TSIS group consistently has higher RS and lower CS than the Top group across all categories and attributes, with RS outliers appearing only in the Bottom groups. Figure 6 qualitatively supports this result, showing attribute-specific deviations in the Bottom generations. These results indicate that representational atypicality is reflected in the final image space.

[2]Following the original RS implementation, samples outside the estimated reference distribution are identified as outliers. We additionally report the number of detected outliers within each group.

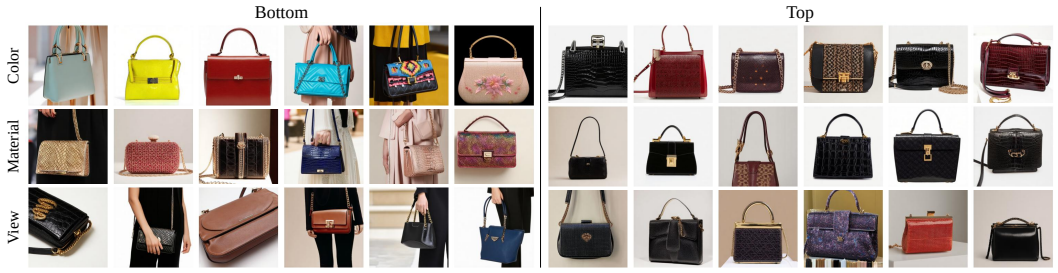



Figure 6: Qualitative comparison of the Bottom and Top TSIS groups for handbag generations. For each visual attribute (color, material, and viewpoint), we show representative samples from the Bottom (lowest TSIS) and Top (highest TSIS) groups. Additional examples are available in Appendix I.

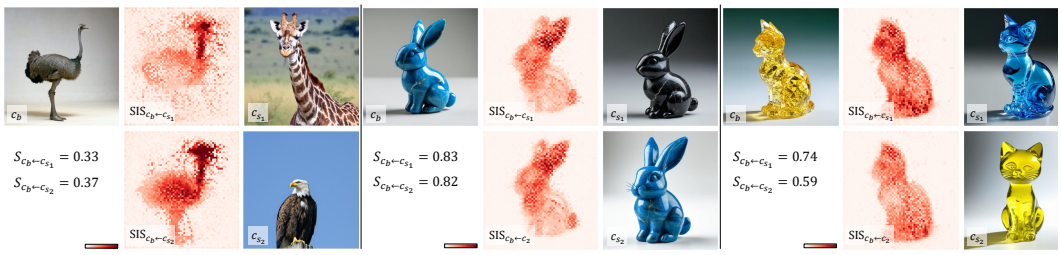



Figure 7: Spatial evidence beyond scalar similarity. Similar SIS scores can arise from different visual regions (ostrich; left), while similar maps may still leave the underlying factor ambiguous (rabbit and cat; right).

## 5 Related Work

Deep feature similarity has been used to characterize perceptual relationships and learned representations (Zhang et al., 2018; Eberle et al., 2020; Raghu et al., 2017; Kornblith et al., 2019), while diffusion representations have been studied for segmentation, correspondence, and image manipulation (Baranchuk et al., 2022; Li et al., 2023; Tang et al., 2023; Luo et al., 2023; Tumanyan et al., 2023). Recent work also suggests that generative representations support broader visual understanding (Gabeur et al., 2026; Le et al., 2026). Our work instead focuses on the relational organization among concepts, using SIS and TSIS to characterize concept-level geometry across generations. Extended discussion is provided in Appendix B.

## 6 Discussion and Conclusion

A key advantage of SIS is that it preserves the spatial evidence underlying a similarity score. Figure 7 illustrates this with the ostrich example. Although giraffe and eagle yield similar SIS scores (0.33 and 0.37), their maps reveal different sources of similarity: the giraffe correspondence is concentrated around the head and neck, whereas the eagle correspondence extends over the body. Thus, relationships that appear similar when reduced to a scalar can arise from different visual cues, which remain visible in SIS. Spatial evidence, however, does not fully explain what drives a similarity. In the rabbit example, the two comparisons produce nearly identical scores (0.83 and 0.82) and similar spatial patterns, making it difficult to determine whether the correspondence arises from object identity, color, viewpoint, or their combination. One promising direction is therefore to extend SIS beyond a single representation. Our block–timestep analysis shows that different visual attributes become distinguishable at different locations in the diffusion process (Figure 5). Combining SIS maps across such locations could provide a multi-axis view of similarity, separating complementary factors underlying the same spatial correspondence. Overall, SIS and TSIS reveal that diffusion representations contain structured conceptual geometries while retaining local evidence for individual relationships. Across object categories and visual attributes, these geometries vary across models and evolve throughout generation, while deviations from them are reflected in atypical final images. We hope this framework provides a basis for studying not only how similar concepts are in diffusion models, but also how those relationships are formed and expressed during generation.

## AI USE STATEMENT

The core research questions, methodology, experimental design, and scientific contributions of this work were developed by the authors. We used generative AI tools to assist with English-language polishing and with drafting and debugging code for large-scale experiments. All AI-assisted text was reviewed and revised by the authors, and all generated code was manually inspected, tested, and validated against the intended experimental procedures. Generative AI was not used to originate the central research idea or to determine the scientific conclusions of this work. The authors take full responsibility for the final content of the paper, including all text, code, analyses, claims, and artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

This work analyzes representations of pretrained text-to-image diffusion models and does not involve human subjects, personally identifiable information, or private data. The proposed framework may reveal biases or undesirable concept associations inherited from model training data, and we view such analyses as potentially useful for improving the transparency and understanding of generative models. We are not aware of additional ethical concerns specific to the experiments presented in this work.

## REPRODUCIBILITY STATEMENT

We provide the methodological details required to reproduce our experiments throughout the main paper and Appendix. The main text specifies the SIS and TSIS formulations, including spatial matching, attention weighting, aggregation, and concept-level similarity computation. The Appendix provides pseudocode for SIS, model-specific image generation settings, and the grid-search procedure used to select diffusion blocks and denoising timesteps. We additionally report the prompt templates, evaluated model configurations, sampling procedures, and concept sets used in the experiments. These details are intended to enable reproduction of both the proposed similarity framework and the conceptual-geometry analyses across diffusion models.

## REFERENCES


Vladimir Arkhipkin, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Anton Bukashkin, Konstantin Kulikov, et al. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 475–485, 2024.

Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SlxSY2UZQT.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1149–1161, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Distillation of diffusion features for semantic correspondence. *WACV*, 2025.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 114432–114462. Curran Associates, Inc., 2025. doi: 10.52202/085713-3816. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a654298de44ec95dec28aa37e8d0af51-Paper-Conference.pdf.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502. IEEE, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Minh-Quan Le, Armand Comas, Alexandros Lattas, Stylianos Moschoglou, Pedro Vélez, Amit Raj, Aaron Germuth, Thabo Beeler, Dimitris Samaras, and Di Qiu. Concurrent image understanding and generation: Self-correcting coupled markov jump processes. *arXiv preprint arXiv:2607.13188*, 2026.

Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7633–7642. IEEE, 2023.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8100–8110, June 2024.

Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/oquab2024tmlr-dinov2/.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24129–24142. Curran Associates, Inc., 2023. doi: 10.52202/075280-1048. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4bfcebedf7a2967c410b64670f27f904-Paper-Conference.pdf.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20031–20042, October 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. URL https://mlanthology.org/iclr/2015/simonyan2015iclr-very/.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 99956–100027. Curran Associates, Inc., 2025. doi: 10.52202/085713-3342. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/90291c8d685db65a8a1e32fd2b66f8bb-Paper-Conference.pdf.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. *Advances in neural information processing systems*, 38:166943, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930. IEEE, 2023.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.

Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18892–18903. IEEE, 2023.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 45533–45547. Curran Associates, Inc., 2023. doi: 10.52202/075280-1973. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8e9bdc23f169a05ea9b72ccef4574551-Paper-Conference.pdf.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

# Appendix:

## Uncovering Conceptual Geometry in Diffusion Models with Spatially Invariant Similarity

## CONTENTS


**A Additional Examples of SIS Map** . . . . . **14**

**B Extended Related Work** . . . . . **14**

**C Grid Search for Block and Timestep** . . . . . **15**

**D Implementation Details** . . . . . **16**

**E Superclass Order in CIFAR-100** . . . . . **17**

**F Implementation Code** . . . . . **18**

**G Additional Results for Class-Level Conceptual Geometry** . . . . . **18**

**H Additional Results on Visual-Level Conceptual Geometry** . . . . . **19**

**I Additional Results for Rare Generations** . . . . . **21**


## A ADDITIONAL EXAMPLES OF SIS MAP

Figure 8 provides additional examples of SIS maps across diverse concept pairs. For each base generation, different source concepts produce distinct spatial correspondence patterns, revealing the visual regions that contribute to their similarity. These examples further illustrate that SIS complements scalar similarity scores with spatial evidence for inspecting individual concept relationships.




Figure 8: Additional examples of SIS maps. Each group shows a base generation and three source generations together with their directional SIS maps.

## B EXTENDED RELATED WORK

We provide a more detailed discussion of prior work on deep feature similarity and diffusion representations, complementing the summary in Sec. 5.

**Deep Feature Similarity.** Similarity in learned feature spaces has long been used to characterize both perceptual relationships between inputs and structural relationships between neural representations. Deep features, for example, provide perceptual similarity measures that align more closely with human judgments than conventional pixel-level metrics (Zhang et al., 2018). BiLRP (Eberle et al., 2020) further showed that deep similarity scores can be decomposed into contributions from individual input features, revealing the visual evidence underlying similarity judgments. Beyond input-level similarity, methods such as SVCCA (Raghu et al., 2017) and centered kernel alignment (CKA) (Kornblith et al., 2019) have been widely used to compare internal representations across layers, models, and training conditions. These approaches establish similarity as a useful tool for probing learned representations, but they are not directly designed for spatially varying representations produced during image generation. Perceptual feature metrics typically assume spatial correspondence or aggregate local distances, whereas representation-level methods compare aligned observations or globally summarized structures. This limitation becomes particularly relevant for diffusion models, where semantically related content may emerge at different spatial locations across generations.

**Diffusion Representations.** Modern diffusion models (Esser et al., 2024; Labs, 2024; Chen et al., 2023; Z-Image Team, 2025; Xie et al., 2025) learn rich visual representations through large-scale generative training. A growing body of work has shown that their intermediate features encode semantic information useful beyond image generation, including semantic segmentation (Baranchuk et al., 2022; Li et al., 2023; Yang & Wang, 2023), interpretable latent structure (Kwon et al., 2022; Park et al., 2023), and controllable image manipulation (Surkov et al., 2025). In particular, diffusion features have proven highly effective for semantic correspondence across images (Tang et al., 2023; Luo et al., 2023; Zhang et al., 2023; Gan et al., 2025; Fundel et al., 2025), suggesting that individual spatial features preserve semantically meaningful local information. Such correspondences are sufficiently reliable to guide the generation process itself (Nam et al., 2024).

Most prior work, however, has used diffusion representations as visual priors for downstream tasks or examined the presence and localization of individual semantic concepts. More recent work suggests a broader connection between generative representations and visual understanding, showing that image generators can serve as general-purpose vision learners (Gabeur et al., 2026) and that understanding and generation can be explicitly coupled within a shared generative process (Le et al., 2026). Building on these observations, we investigate the relational organization of concepts within diffusion representation spaces.

## C Grid Search for Block and Timestep

The conceptual geometry revealed by TSIS depends on the diffusion block $l$ and denoising timestep $t$ from which the representations are extracted. We therefore select a model-specific operating point $(l_m^*, t_m^*)$ through a grid search based on self-class prediction.

Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ denote a set of $N$ concepts. For each concept, we divide all generated samples into five folds and perform 5-fold cross-validation. At each candidate block and timestep $(l, t)$, a test sample $x$ is compared with the training samples of every concept using TSIS. Specifically, we compute the mean TSIS between $x$ and the training samples belonging to concept $c_n$:

$$q_n(x; l, t) = \frac{1}{|\mathcal{D}_n^{\text{train}}|} \sum_{x' \in \mathcal{D}_n^{\text{train}}} \text{TSIS}_{l,t}(x, x'), \tag{18}$$

where $\mathcal{D}_n^{\text{train}}$ denotes the training samples of concept $c_n$. The predicted concept is the one with the highest average similarity:

$$\hat{y}(x; l, t) = \underset{n \in \{1,\ldots,N\}}{\arg\max}\ q_n(x; l, t). \tag{19}$$

For each fold, we compute the classification accuracy over all held-out samples. The self-class prediction accuracy $\text{Acc}(l, t)$ is obtained by averaging the accuracy across the five folds. We then average this accuracy over timesteps for each block and select the block with the highest mean accuracy:

$$l_m^* = \underset{l \in \mathcal{L}_m}{\arg\max}\ \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \text{Acc}(l, t), \tag{20}$$

Table 3: Generation settings and selected diffusion representations for the evaluated models. $l$ and $t$ denote the selected block and denoising timestep for the class-level experiment, respectively. SDXL uses a resolution of $1024 \times 1024$ to ensure sufficient image quality, while SANA uses the same resolution to match the required latent dimensions. *For SDXL and SD1.5, the reported block indices correspond to blocks within the mid-block.

| | | Image Size | Inference Steps | Guidance Scale | $t$ | $l$ |
|---|---|---|---|---|---|---|
| Transformer | Stable Diffusion 3.5 Large | 512 | 20 | 4.5 | 12 | 24 |
| | Stable Diffusion 3.5 Medium | 512 | 20 | 4.5 | 12 | 10 |
| | Stable Diffusion 3.5 L Turbo | 512 | 4 | 0.0 | 3 | 24 |
| | Lumina-Image-2.0 | 512 | 20 | 4.0 | 12 | 14 |
| | Z-Image | 512 | 20 | 4.0 | 12 | 20 |
| | Z-Image Turbo | 512 | 6 | 0.0 | 4 | 20 |
| | SANA-1.5 | 1024 | 20 | 4.5 | 12 | 6 |
| | FLUX.1-Dev | 512 | 20 | 3.5 | 14 | 12 |
| | FLUX.1-Schnell | 512 | 4 | 0.0 | 3 | 12 |
| | PixArt-$\alpha$ | 512 | 20 | 4.5 | 16 | 12 |
| U-Net | Kandinsky-3 | 512 | 20 | 4.5 | 16 | 11 |
| | SDXL 1.0-Base | 1024 | 20 | 5.0 | 16 | 2* |
| | SD v1.5 | 512 | 20 | 7.5 | 16 | 1* |

where $\mathcal{L}_m$ and $\mathcal{T}_m$ denote the candidate blocks and timesteps for model $m$, respectively. For the selected block, we fix the timestep to $t_m^* \approx 0.8T$, where $T$ denotes the full denoising trajectory.

This procedure selects the block at which samples of the same concept are most consistently identified through their TSIS relationships across the denoising process, without training an additional classifier. When multiple timesteps achieve the same maximum accuracy for the selected block, we use a timestep near $t = 0.8T$ as a tie-breaking operating point, rather than defaulting to the first timestep, where $T$ denotes the full denoising trajectory.

**Cross-validation splits.** For the class-level experiment on CIFAR-100, we generate 20 samples for each of the 100 classes, resulting in 2,000 images per model. Under 5-fold cross-validation, each fold contains 16 training and 4 test images per class, corresponding to 1,600 training and 400 test images in total.

For the visual-level experiments, the dataset consists of 300 images generated from all combinations of 10 colors, 10 materials, and 3 viewpoints. When evaluating color or material, the corresponding attribute defines 10 classes, with 30 images per class. Each fold therefore contains 24 training and 6 test images per class. For viewpoint, the three viewpoints define three classes, each containing 100 images, with 80 training and 20 test images per class in each fold. The same 5-fold self-class prediction procedure described above is applied to each attribute independently.

## D Implementation Details

We use publicly available pretrained checkpoints for all diffusion models, obtained from the Hugging Face repositories linked to their corresponding official releases, without additional training or fine-tuning. Table 3 summarizes the generation settings and selected block and timestep $(l, t)$ for each model. For the vision-encoder baselines, we use the official implementation and pretrained weights for DINOv2[3], the pretrained VGG16 provided by `torchvision`, and the official OpenAI implementation and pretrained weights for CLIP [4]. All vision encoders are used in inference mode without additional training or fine-tuning.

[3] dinov2-vitb14

[4] openai/clip-vit-base-patch32

Table 4: Superclass and class ordering used for the CIFAR-100 affinity matrices. Indices denote the row/column positions in the matrices. The blue highlighted labels are used to generate Figure 4.

| Group | Index | Superclass | Classes |
|---|---|---|---|
| People & mammals | 1–5 | people | baby, boy, girl, man, woman |
| | 6–10 | small mammals | hamster, mouse, rabbit, shrew, squirrel |
| | 11–15 | medium mammals | fox, porcupine, possum, raccoon, skunk |
| | 16–20 | large carnivores | bear, leopard, lion, tiger, wolf |
| | 21–25 | large omnivores and herbivores | camel, cattle, chimpanzee, elephant, kangaroo |
| | 26–30 | aquatic mammals | beaver, dolphin, otter, seal, whale |
| Other animals | 31–35 | fish | aquarium fish, flatfish, stingray, shark, trout |
| | 36–40 | reptiles | crocodile, dinosaur, lizard, snake, turtle |
| | 41–45 | insects | bee, beetle, butterfly, caterpillar, cockroach |
| | 46–50 | non-insect invertebrates | crab, lobster, snail, spider, worm |
| Plants | 51–55 | trees | maple tree, oak tree, palm tree, pine tree, willow tree |
| | 56–60 | flowers | orchid, poppy, rose, sunflower, tulip |
| | 61–65 | fruit and vegetables | apple, mushroom, orange, pear, sweet pepper |
| Household objects | 66–70 | food containers | bottle, bowl, can, cup, plate |
| | 71–75 | household electrical devices | clock, keyboard, lamp, telephone, television |
| | 76–80 | household furniture | bed, chair, couch, table, wardrobe |
| Vehicles | 81–85 | vehicles 1 | bicycle, bus, motorcycle, pickup truck, train |
| | 86–90 | vehicles 2 | lawn mower, rocket, streetcar, tank, tractor |
| Outdoor scenes | 91–95 | large man-made outdoor things | bridge, castle, house, road, skyscraper |
| | 96–100 | large natural outdoor scenes | cloud, forest, mountain, plain, sea |

**Denoising timestep indexing.** We use $t$ to denote the inference-step index rather than the raw diffusion timestep (e.g., $[0, 999]$). For a model generated with 20 inference steps, $t \in \{1, \ldots, 20\}$ refers to the corresponding step in the denoising trajectory (e.g., $t = 12$ is the 12th inference step), regardless of the scheduler timestep associated with that step. Models with different numbers of inference steps are indexed analogously.

**Generation Prompts.** For Figure 3, all images were generated using the prompt template *"a photo of {concept}"*, where {concept} was replaced by one of the following: *"an astronaut riding a horse in a jungle"*; *"an astronaut on mars"*; *"the american quarter horse"*; *"an astronaut jumping on a beach, side view"*; *"a large stuffed horse"*; *"a living room with a lamp and a potted plant"*; *"a floor lamp"*; *"a flower vase"*; *"willow tree"*; and *"a table lamp"*.

## E SUPERCLASS ORDER IN CIFAR-100.

For the class-level analysis in Sec. 4.1, we reorder the 100 CIFAR-100 classes to make the affinity matrices easier to interpret. The original CIFAR-100 superclass indices follow alphabetical order, which places semantically unrelated superclasses next to each other (e.g., *large carnivores* next to *large man-made outdoor things*). We instead arrange the 20 superclasses so that semantically related groups are contiguous: people and mammals, ordered roughly by body size and habitat, are followed by the remaining animal superclasses (fish, reptiles, insects, and non-insect invertebrates), plants, household objects, vehicles, and large outdoor scenes. Within each superclass, we retain the original CIFAR-100 class order. Table 4 lists the resulting ordering, which defines the row and column order of all CIFAR-100 affinity matrices. We rename the class *ray* to *stingray* to avoid ambiguity with non-animal meanings of the word.

## F Implementation Code

We provide PyTorch-style pseudocode for SIS and TSIS. SIS computes directional spatial similarity between a pair of generations. TSIS aggregates SIS across multiple generations and both comparison directions to obtain a concept-level similarity score.

```
import torch
import torch.nn.functional as F

def SIS(feat_b, feat_s, attn_b, attn_s,
        eps=1e-8):
    """
    feat_b, feat_s: (P, C)
    attn_b, attn_s: (P,)
    """

    # Feature normalization
    feat_b = F.normalize(
        feat_b, p=2, dim=-1, eps=eps)
    feat_s = F.normalize(
        feat_s, p=2, dim=-1, eps=eps)

    # All-to-all cosine similarity
    M = feat_b @ feat_s.T  # (P, P)

    # Base-to-source matching
    spatial_sim, match_idx = M.max(dim=-1)

    # Rematch before normalization
    attn_s_match = attn_s[match_idx]

    # Attention normalization
    attn_b = F.normalize(
        attn_b, p=2, dim=0, eps=eps)
    attn_s_match = F.normalize(
        attn_s_match, p=2, dim=0, eps=eps)

    # Directional SIS map
    sis_map = (
        attn_b * spatial_sim * attn_s_match
    )

    return sis_map
```

```
def directional_TSIS(
    feat_b, feat_s, attn_b, attn_s
):
    """
    feat_b: (N_b, P, C)
    feat_s: (N_s, P, C)
    attn_b: (N_b, P)
    attn_s: (N_s, P)
    """

    sis_maps = []

    # All base-source pairs
    for i in range(feat_b.shape[0]):
        for j in range(feat_s.shape[0]):
            sis_maps.append(
                SIS(
                    feat_b[i], feat_s[j],
                    attn_b[i], attn_s[j]
                )
            )

    # Average across pairs
    sis_map = torch.stack(
        sis_maps
    ).mean(dim=0)

    # Spatial aggregation
    return sis_map.sum()


def TSIS(feat_b, feat_s, attn_b, attn_s):
    """Bidirectional concept similarity."""

    score_b2s = directional_TSIS(
        feat_b, feat_s, attn_b, attn_s
    )
    score_s2b = directional_TSIS(
        feat_s, feat_b, attn_s, attn_b
    )

    return 0.5 * (score_b2s + score_s2b)
```

## G Additional Results for Class-Level Conceptual Geometry

This section provides additional results for the class-level conceptual geometry analysis in Sec. 4.1. Figure 9 presents the full CIFAR-100 affinity matrices together with the block–timestep grid-search results for the evaluated diffusion models.

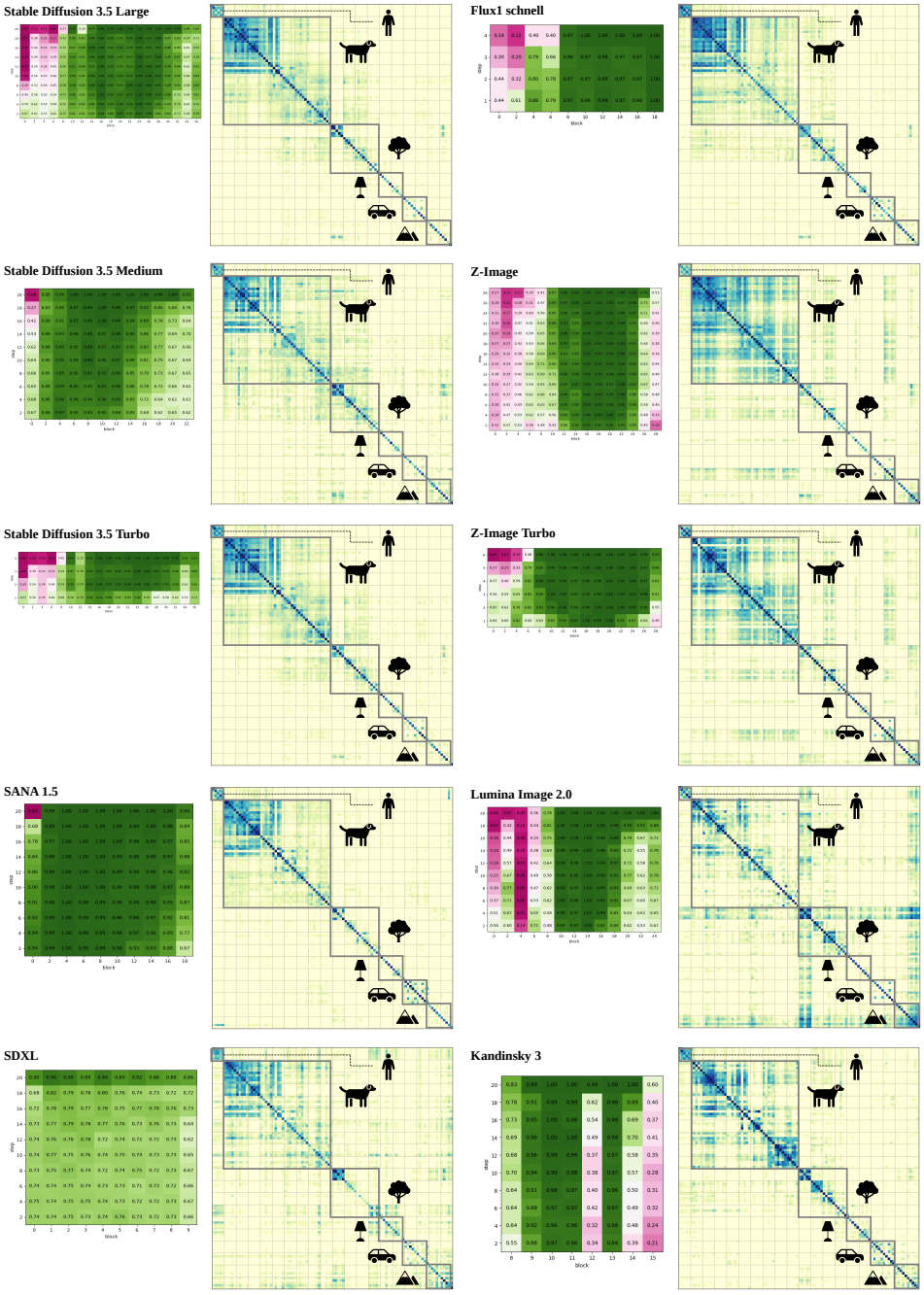
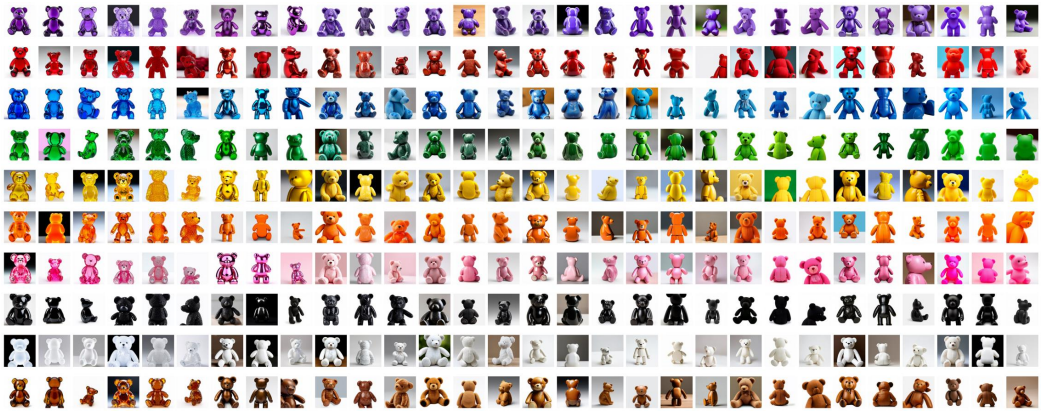



Figure 9: Additional class-level conceptual geometry results across diffusion models. For each model, the left panel shows the TSIS self-classification accuracy across blocks and denoising timesteps, and the right panel shows the CIFAR-100 TSIS affinity matrix at the selected $(l, t)$. The results reveal model-dependent differences in both where class-level information is most distinguishable and the conceptual geometry formed at the selected representation.

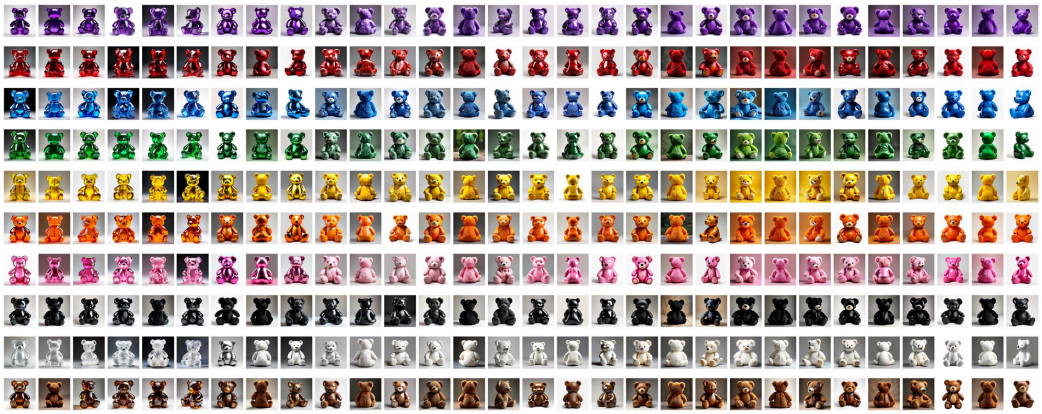
## H Additional Results on Visual-Level Conceptual Geometry

This section provides additional results for the visual-level conceptual geometry analysis in Sec. 4.2. We first provide generated samples (Figure 10) covering the controlled color variations used in the analysis. We then present the full UMAP visualizations (Figure 11) obtained from the TSIS

Stable Diffusion 3.5 - Medium



SANA 1.5



Lumina Image 2.0

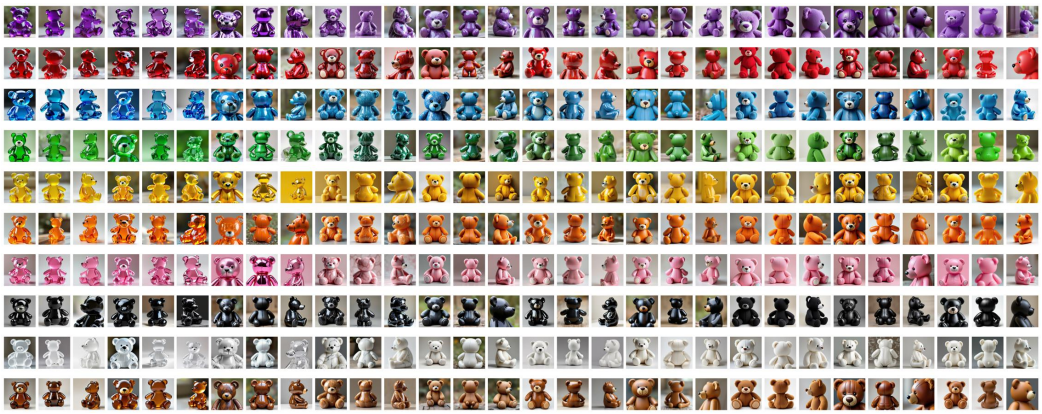


Figure 10: Generated samples for the visual-level conceptual geometry analysis. Rows correspond to the 10 color classes, and columns show generations with varying material and viewpoint conditions. We show the complete set of color-controlled samples for SD3.5 Medium, SANA-1.5, and Lumina-Image 2.0.

affinity matrices for SD3.5 Medium, SANA-1.5, and Lumina Image 2.0 across color, material, and viewpoint.

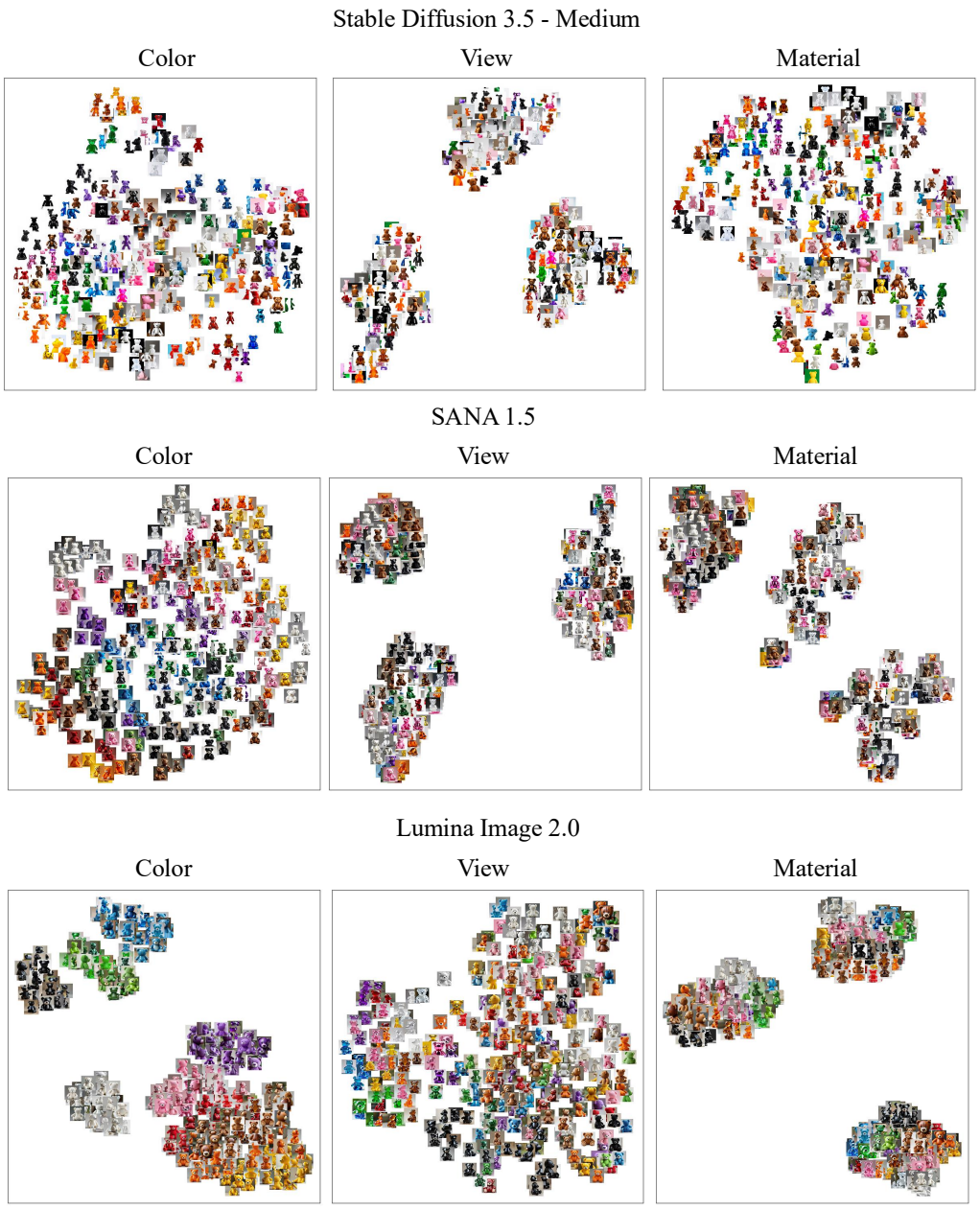



Figure 11: UMAP visualization of generated samples for the visual-level conceptual geometry analysis. Rows correspond to SD3.5 Medium, SANA-1.5, and Lumina-Image 2.0, and columns correspond to the color, view, and material attributes. Each sample is shown as its image thumbnail at its UMAP-projected position computed from the corresponding attribute-specific TSIS affinity matrix, illustrating how each model organizes the generated samples along each attribute.

# I Additional Results for Rare Generations

This section provides additional results for rare generations in Sec. 4.3.

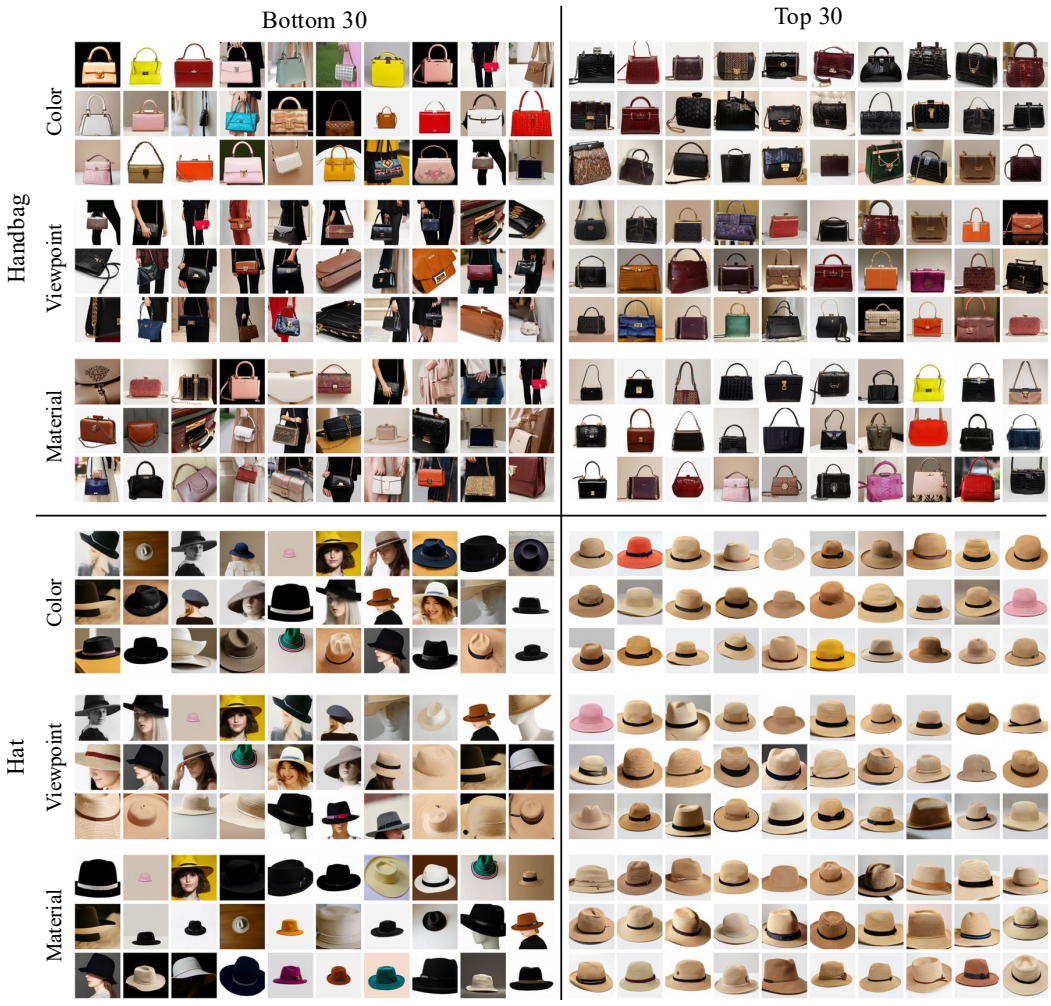



Figure 12: Qualitative comparison of the Bottom and Top TSIS groups for generations.

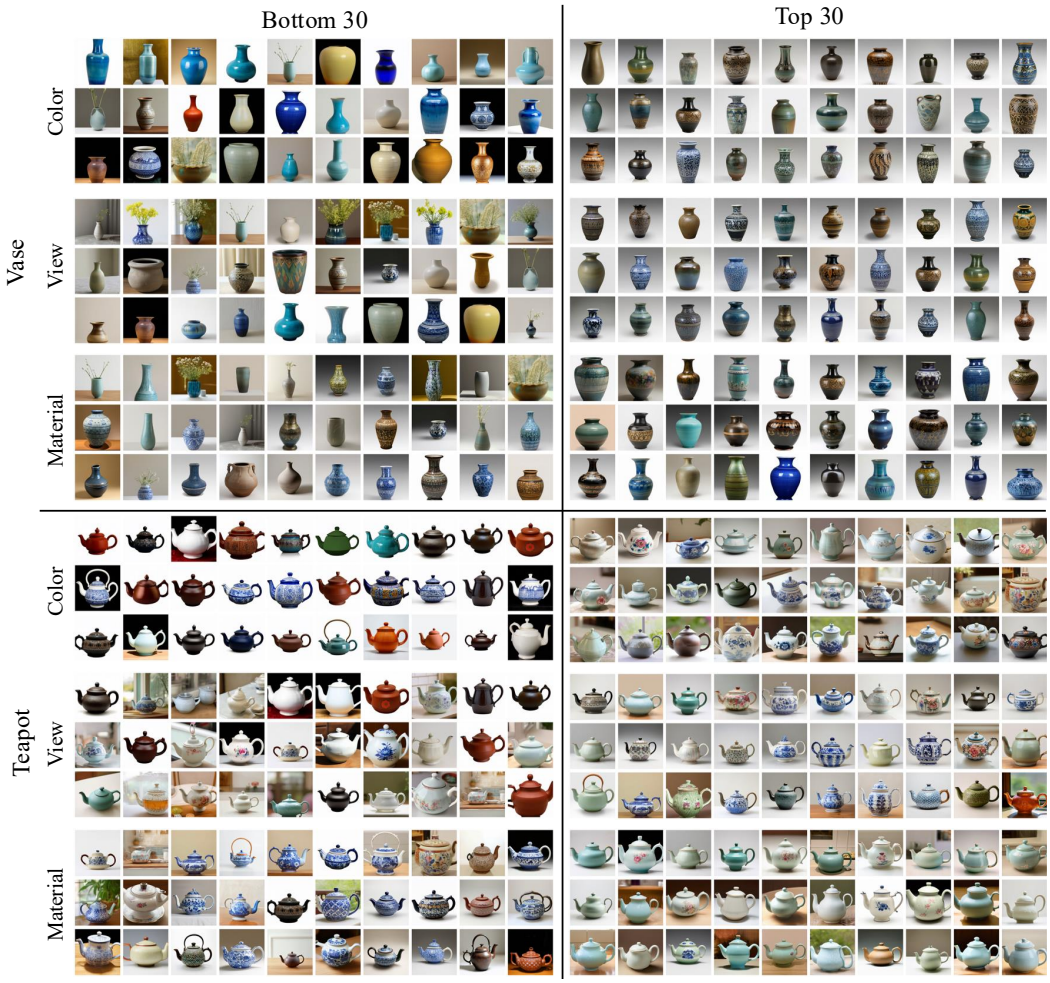



Figure 13: Qualitative comparison of the Bottom and Top TSIS groups for generations.