SliderSpace: Decomposing the Visual Capabilities of Diffusion Models

Rohit Gandikota[1*] Zongze Wu[2] Richard Zhang[2] David Bau[1] Eli Shechtman[2] Nick Kolkin[2]
[1]Northeastern University [2]Adobe Research

Figure 1. Given a prompt, SliderSpace identifies the principal directions of the visual capabilities of a diffusion model by decomposing the image distribution over the prompt. By manipulating these directions as sliders, users can control and combine them to explore the creative possibilities of the model. We visualize the top directions discovered by SliderSpace within Flux Schnell [4] for the concept "Toy".

Abstract

We present SliderSpace, a framework for automatically decomposing the visual capabilities of diffusion models into controllable and human-understandable directions. Unlike existing control methods that require a user to specify attributes for each edit direction individually, SliderSpace discovers multiple interpretable and diverse directions simultaneously from a single text prompt. Each direction is trained as a low-rank adaptor, enabling compositional control and the discovery of surprising possibilities in the model's latent space. Through extensive experiments on state-of-the-art diffusion models, we demonstrate SliderSpace's effectiveness across three applications: concept decomposition, artistic style exploration, and diversity enhancement. Our quantitative evaluation shows that SliderSpace-discovered directions decompose the visual structure of model's knowledge effectively, offering insights into the latent capabilities encoded within diffusion models. User studies further validate that our method produces more diverse and useful variations compared to baselines. Our code, data and trained weights are available at sliderspace.baulab.info

1. Introduction

Text-to-image diffusion models are capable of generating remarkable visual variations from a single prompt through different random initializations. However, this vast creative potential remains largely opaque to users—while we can generate diverse images, we lack understanding of the underlying structure of these variations. This presents a fundamental challenge: how can we discover and expose the latent visual capabilities encoded within these models?

The challenge touches on a key limitation in how we interact with diffusion models today. Current control methods require users to explicitly specify their desired edits in advance through prompts [16], reference images [8, 24, 29, 41, 42, 52], or attribute vectors [20, 21, 30, 32, 46, 50]. That contrasts sharply with natural human creative workflows, where artists dynamically explore creative ideas and jointly refine them toward meaningful artistic outcomes [23]. The need for pre-specified controls creates a barrier between users and the full creative potential of these models.

Interestingly, earlier generative models like GANs [6, 17, 28] naturally developed more interpretable internal structures. Their compact latent spaces often exhibited

*Correspondence to gandikota.ro@northeastern.edu

emergent disentanglement [19, 35, 45, 47], enabling continuous and compositional control over generated images. Users could explore these spaces to discover interesting variations that would be difficult to describe in words [47], then combine them to achieve their creative goals [18].

Diffusion models have largely superseded GANs in conditional image synthesis [11], achieving greater diversity through much higher-dimensional latents. And yet an understanding of the underlying structure of these larger latent spaces has remained elusive. In this work, we ask a fundamental question: *Can we automatically discover the visual structure within a diffusion model's knowledge of a concept?* Rather than requiring user-specified controls, we aim to decompose the model's internal representations into expressive directions that users can explore and combine.

To address these needs, we present **SliderSpace**, a framework that brings systematic explorability to diffusion models. Given just a text prompt, SliderSpace discovers a canonical set of meaningful, diverse, and controllable directions within the model's knowledge of that concept. Each direction is implemented as a low-rank adapter [24] that can be scaled and composed with others, allowing users to explore and smoothly combine different aspects of variation, as shown in Figure 1.

We ground SliderSpace discovery in three key requirements for meaningful decomposition of a diffusion model's visual manifold:

1. **Unsupervised Discovery:** The decomposition process should emerge from the intrinsic structure of the model's learned representation, rather than being guided by predefined attributes. This ensures we capture the true topology of the model's knowledge space rather than projecting our assumptions onto it.
2. **Semantic Orthogonality:** Each discovered control must represent a distinct semantic direction. This is enforced in a semantic feature space, like CLIP, where every slider has an orthogonal effect in embeddings. This prevents discovering multiple controls that create similar semantic effects [15, 33, 47], making the system more efficient.
3. **Distribution Consistency:** Directions must induce consistent transformations across both random seeds and prompt variations.

These requirements naturally lead to our proposed framework, which we formalize in Section 4. As we show in our experiments, SliderSpace is architecture-agnostic, working with both conventional U-Net based models like Stable Diffusion [34, 37, 40, 43, 51] and recent transformer-based architectures like Flux [4].

We demonstrate the expressiveness of SliderSpace through three applications: First, we show how SliderSpace can decompose high-level concepts into diverse and expressive components, revealing the natural axes of variation in the model's understanding. Second, we explore artistic style variation, where SliderSpace discovers directions that match or exceed the diversity of manually curated artist lists while being judged more useful by human evaluators. Finally, we show how SliderSpace can help reverse the mode collapse commonly observed in distilled diffusion models, restoring diversity while maintaining generation speed.

Beyond providing practical creative control, SliderSpace opens new avenues for understanding and utilizing the latent capabilities of diffusion models. By mapping these models' visual potential into intuitive, composable directions, we take a step toward making their creative possibilities more accessible and interpretable to users.

2. Related Works

Recent text-to-image diffusion models have demonstrated remarkable capabilities in generating diverse visual concepts [40]. While newer foundation models enhance text-image alignment through LLM-generated captions [13, 26], the fundamental challenge remains: text-conditioned generation is inherently under-determined, with multiple distinct outputs potentially satisfying the same prompt, making precise control challenging.

Prior work has explored various approaches to enhance generation control. One direction introduces additional conditioning modalities: spatial signals via adapters (ControlNet [52]), attention-based regional control [8], and image-based conditioning for identity preservation [30, 46, 50] or style transfer via attention manipulation [21]. Another line of research focuses on refining existing images through disentangled text-driven editing [5, 7, 10, 32, 48, 49], where specific attributes are modified while preserving others, often using spatial masks. However, these text-driven approaches inherit the same under-determination challenges as the base models.

Notably, disentangled control has been more naturally achieved in GANs [17], particularly through StyleGAN's low-dimensional latent space [28], which exhibits emergent disentanglement even at scale [27]. This has enabled powerful latent space and image editing capabilities [1, 2, 47]. While diffusion models offer superior generation quality and diversity, they lack two key advantages of GANs' latent space: continuous, compositional edits and emergent disentanglement. This limitation prevents users from making serendipitous discoveries about visual variations captured in the training data, instead constraining them to variations they can explicitly describe through prompts.

Recent works have explored different approaches to discovering interpretable directions in diffusion models. The weights2weights method [12] learns a manifold of personalized model weights by fine-tuning individual LoRA adapters for each identity and applying PCA to discover a weight space that enables editing. While effective, this requires training separate models per instance. NoiseCLR [9]

learns text embeddings through contrastive learning on a data distribution, but the discovered directions can be arbitrary and lack semantic interpretability. Similarly, Liu et al. [31] propose unsupervised decomposition of images into compositional concepts by learning multiple embeddings, but their approach often yields redundant or non-semantic directions. In contrast, our work directly learns a small set of semantically grounded sliders that decompose the distribution into interpretable and composable directions, enabling infinite creative variations through systematic exploration of the visual manifold.

Concept Sliders [16] addressed continuous control by leveraging LoRA adaptors [24] to learn user-defined attributes. Our work complements this by tackling emergent disentanglement through self-supervised decomposition of the model's inherent variations into composable control dimensions, enabling systematic exploration of the model's creative capabilities.

3. Background

3.1. Latent Diffusion Models

State-of-the-art text-to-image diffusion models [4, 34, 38, 39] often belong to the class of latent diffusion. Unlike traditional diffusion models that operate in pixel space, latent diffusion models work in a compressed latent space, offering significant computational advantages [40]. The diffusion modelling can be formalized as follows:

Let $\mathbf{x}_0$ be an initial image and $\mathbf{x}_T$ be pure Gaussian noise. The forward diffusion process gradually adds noise to the image. The generative process aims to reverse this diffusion, starting from x_T and progressively denoising to reconstruct $\mathbf{x}_0$. At a timestep t the model takes $\mathbf{x}_t$ as input and predicts a noise ϵ_t such that the next step $\mathbf{x}_{t-1}$:

$$\mathbf{x}_{t-1} \leftarrow \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_t}{\sqrt{\alpha_t}} \tag{1}$$

We can estimate the final image $\tilde{\mathbf{x}}_{0,t}$ by taking the same direction ϵ_t for remaining diffusion steps. This can be achieved by recursively applying the denoising from Equation 1 with the same direction. We label this "Final Image Extrapolation":

$$\tilde{\mathbf{x}}_{0,t} \leftarrow \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}. \tag{2}$$

This enables us to visualize the final image $\tilde{\mathbf{x}}_{0,t}$ that the diffusion model is planning of at each timestep t without actually running denoising forward passes through all timesteps.

3.2. LoRA: Low Rank Adaptors

Low-rank adaptator (LoRA) [24] are a class of light-weight adaptors that are attachable to the weights of the model. Given a pre-trained model layer with weights $W_0 \in \mathbb{R}^{d\times k}$, LoRA decomposes the weight update ΔW as:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}, \tag{3}$$

where $r \ll \min(d, k)$ is a small rank that constrains the update to a low-dimensional subspace. This decomposition allows for efficient parameter updates and has shown success in various downstream tasks. However, the application of LoRA to unsupervised discovery of semantic directions in diffusion model weight space remains unexplored.

4. Method

We present SliderSpace, a framework for decomposing a diffusion model's visual capabilities into semantically orthogonal control dimensions (Fig. 1). Given a pre-trained text-to-image diffusion model θ and a prompt c, our goal is to discover n independent directions that capture the principal modes of variation in the model's learned distribution.

4.1. Problem Formulation

Let $\mathcal{M}_\theta(c)$ denote the manifold of possible images that model θ can generate for prompt c. We aim to identify a set of controllable directions $\{\mathcal{T}_i\}_{i=1}^n$ that: (1) span the major modes of variation in $\mathcal{M}_\theta(c)$, (2) maintain semantic consistency across different initializations, and (3) are mutually orthogonal in semantic space.

Building on recent advances in model adaptation [16], we formulate each control dimension as a LoRA adaptor [24], updating $\mathcal{T}_i$, where $i \in \{1, ..., n\}$. These lightweight adapters introduce targeted modifications to the model's cross-attention layers, enabling precise control over specific generative attributes.

4.2. SliderSpace: Unsupervised Visual Discovery

SliderSpace discovery process consists of three key steps as depicted in Figure 2:

Distribution Sampling First, we generate a diverse set of samples $\{x_j\}_{j=1}^m$ from $\mathcal{M}_\theta(c)$ by varying the random seed (for stability, $m \approx 5000$). For each sample, we extract the estimated final image $\tilde{\mathbf{x}}_{0,t}$ at each timestep t using Eq. 2.

Semantic Decomposition We map each sample to a semantic embedding space, like CLIP, $\phi(\tilde{\mathbf{x}}_{0,t})$ and compute the principal components $V = \{v_i\}_{i=1}^n$ of the resulting distribution. These components represent orthogonal directions of maximal variation in semantic space:

$$V = \text{PCA}(\{\phi(\tilde{x}_{0,t})\}_{j,t}) \tag{4}$$

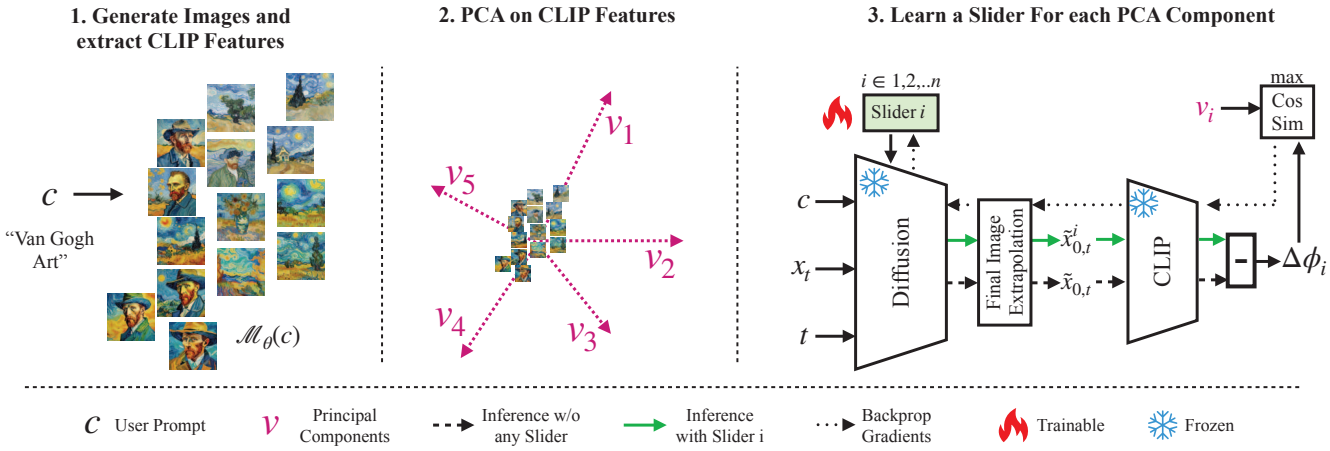

Figure 2. Given a prompt, SliderSpace first generates images and extracts the CLIP features. We then compute the spectral decomposition of the CLIP features and align the each slider with extracted principle components. Each slider is therefore trained to represent a unique semantic direction that are relevant in the diffusion model's knowledge of the prompt.

Slider Training For each principal direction v_i, we train a corresponding adapter $\mathcal{T}_i$ to induce transformations that align with v_i in semantic CLIP space. The training objective for each slider is:

$$\mathcal{L}_{\text{sliderspace}} = \sum_{i=1}^{n} 1 - \cos(\Delta\phi_i, v_i), \tag{5}$$

where $\Delta\phi_i = \phi(\tilde{\mathbf{x}}^i_{0,t}) - \phi(\tilde{\mathbf{x}}_{0,t})$ represents the transformation induced by slider i in embedding space and $\tilde{x}^i_{0,t}$ is the estimated final image with the slider $\mathcal{T}_i$ attached. This ensures that each slider's effect is semantically aligned with its corresponding principal direction while maintaining orthogonality with other sliders, building on established work showing that embedding differences encode semantic relationships [15, 33].

Our formulation satisfies the three key requirements outlined in Section 1. First, unsupervised discovery is achieved by deriving directions directly from the model's intrinsic variation through PCA in a semantic embedding space, without imposing predefined attributes or external supervision. Second, distribution consistency is enforced by our training objective $\mathcal{L}_{\text{sliderspace}}$, which ensures each slider's transformation maintains consistent direction in embedding space across different seeds and timesteps. Finally, semantic orthogonality is guaranteed through the PCA-based initialization of directions as each principal components are mutually orthogonal, ensuring each slider captures a distinct mode of variation. Together, these components enable the discovery of interpretable and reliable control dimensions that effectively decompose the model's learned distribution. In majority of this paper, we use CLIP [36] as our primary semantic encoder.

4.3. Interpretability & Control

SliderSpace provides a dual contribution: it serves as both a framework for discovering expressive dimensions of control and as a mechanism for decomposing the model's learned concept space into semantically meaningful components. Each adapter functions as a "slider" controlling a specific attribute of the generated image, enabling fine-grained manipulation of the output while maintaining semantics.

The resulting set of adapters provides valuable insights into the model's conceptual understanding, revealing nuanced semantic relationships that may not be immediately apparent from the text prompt alone.

Moreover, the low-rank structure of our adapters ensures computational efficiency and minimal memory overhead, making them particularly suitable for real-time interactive applications. This enables users to explore the concept space dynamically by modulating the influence of each adapter, facilitating intuitive creative control over the image generation process while maintaining the underlying semantic integrity of the original prompt.

5. Experiments

We conduct our main experiments to evaluate SliderSpace using SDXL-DMD [51], a 4-step distilled diffusion model. Our implementation requires less than 24GB VRAM and can discover 64 semantic directions in under 2 hrs on a single A100 GPU. When concepts exhibit severe mode collapse, the discovery process can be enhanced by generating data using undistilled base models or LLM-expanded prompts for increased sample diversity. We demonstrate SliderSpace's generalization to SDXL [34], SDXL-Turbo [43], and transformer-based FLUX Schnell [4] in appendix. Our analysis focuses on three key applications:

Figure 3. SliderSpace decomposes the visual variation of diffusion model's knowledge corresponding to a concept. These directions can be perceived as interpretable directions of the model's hierarchical knowledge. We show the decomposed slider direction for a concept using SliderSpace and the corresponding labels generated by Claude 3.5 Sonnet.

concept decomposition, art styles exploration, and diversity enhancement in distilled models.

5.1. Concept Decomposition

We first demonstrate how SliderSpace can serve as an exploratory tool by decomposing high-level concepts into semantic directions that align with the diffusion model's internal representations. Summarizing and exposing the dominant variations a model is capable of for a particular prompt.

Given a concept prompt (e.g., "picture of a monster"), we discover SliderSpace directions.Figure 3 shows discovered sliders for "Monster", and "car" concepts. We label the sliders using Claude 3.5 Sonnet [3] by showing multiple image pairs showing the effect of each slider and prompting to identify the semantic transformation being applied. Through these discovered directions, we demonstrate the manipulation of individual attributes. As these directions are intended to capture the model's visual possibilities for a prompt, we naturally ask the question, *"How much variation is enabled by these directions?"*. To quantitatively evaluate the diversity enabled by SliderSpace against base model, we compute DreamSim [14] distance to measure inter-image variation across 2500 generated samples per concept (Figure 4). For SliderSpace-augmented generation, we generate the images by randomly activating a sparse subset of 3 sliders (out of 32 discovered directions) for each generation. Our analysis reveals that SliderSpace-generated samples exhibit significantly higher inter-image diversity compared to the baseline model outputs. We measure the CLIP-Score [22] between the input prompts and the generated images to measure text-alignment with the prompt. We find that they have similar CLIP Scores to the images generated by the base model. This suggests that SliderSpace effectively expands the achievable variation in the model's knowledge, while maintaining semantic consistency. To validate our findings through human perception, we conducted pairwise comparisons of image grids generated by SliderSpace versus baseline methods. As shown in Table 1, users consistently preferred SliderSpace outputs across diversity, utility, and creative potential.

	User Study (Win Rate %)		
Method vs.	**"Diverse"**	**"Useful"**	**"Creative"**
SDXL-DMD	72.4	66.0	68.1
LLM + SDXL-DMD	62.5	62.5	62.5
SDXL	65.3	61.2	59.2

Table 1. Users perceive SliderSpace generated images to be more diverse, useful, and interesting. We show win-rate percentages of SliderSpace samples against baselines.

5.2. Art Styles Exploration

We also evaluate to what degree SliderSpace can expose "all" of the art styles that the diffusion model has learned from its training data. As a proxy for "all", we use a diverse set of art styles, manually discovered and documented by ParrotZone [25]. We explore the visual artistic space by decomposing the prompt "artwork in the style of a famous artist" into 64 directions. This process enables us to discover the SliderSpace of art and create a comprehensive dictionary of discoverable art styles in a diffusion model.

To compare how this measures up against supervised methods like Concept Slider [16], we train 64 manually curated concept sliders using LLM generated training

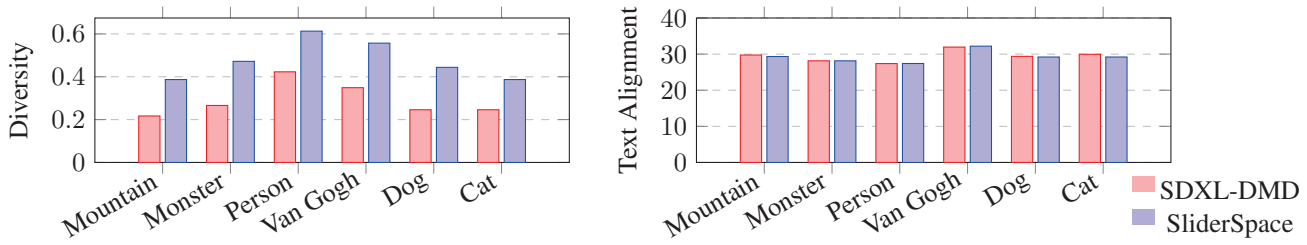

Figure 4. (left) SliderSpace generates diverse variations of a concept, as measured by DreamSim [14] distance across generated samples (higher is better). (right) The method maintains similar text-to-image alignment, as measured by CLIP Scores [22] (lower is better).

prompts. We compute FID scores against the generated samples conditioned on actual artist names from Parrot-Zone [25] dataset. The dataset contains 4388 artists mimicked by SDXL, are discovered through exhaustive manual search. We also establish two other baselines. First, we use GPT/Claude to generate 4388 different artistic style names (e.g., "cubism", "Warhol"). Second, we use a generic prompt like "[subject] in the style of a famous artist". Using our discovered SliderSpace, we randomly sample a sparse set of 3 sliders to generate an equivalent number of images. Specifically, for each baseline, we match the provided samples of [25] by creating two images per style for "building landscape" and "character portraits" each (total of four images per style) to ensure same semantic structure across the datasets and that FID measures the artistic spread. Figure 5 shows random, non-cherry picked images from all the methods and their corresponding FID scores. We find that SliderSpace generates a distribution significantly more closely matches the manually curated artist names than the baselines, including supervised methods like Concept Sliders. Figure 6 shows qualitative comparison of art styles discovered by SliderSpace and styles discovered manually [25].

To validate the practical utility of our "artistic" SliderSpace, we conduct a user study examining both usefulness and diversity of generated samples (Table 2). We conduct comparitive study with 2 grids of 9 images (total of 1000 pairs), one with slider generated images and other with baselines. Users find the SliderSpace-generated images are far more diverse and useful than the baseline methods. We also find that users would prefer using SliderSpace images over images prompted with the real artist names [25], while finding these both to be equally diverse. We provide user-study details and qualitative examples in Appendix.

5.3. Diversity Enhancement

Finally, we explore SliderSpace's ability to addressing mode collapse in distilled models. Instead of discovering SliderSpace for narrow distributions, we train it on a larger spread of 8000 randomly selected COCO-30k prompts and enhance them with LLM. We use DMD model for generating training dataset since we wish to undo mode collapse of DMD models. We discover a generic SliderSpace of 64 sliders that captures the generally applicable visual variations within SDXL-DMD. Figure 7 showcases qualitative examples of two distributions for the prompts "Car driving through a forest" and "Picture of a person". We demonstrate how generating images by randomly sampling from these SliderSpace directions effectively increases output diversity. This demonstrates that the mode collapse in distilled model can be reversed by discovering and exploring the visual structure. We show more examples in appendix.

Method	User study (win rate %) "Diverse"	"Useful"
vs. Real Artist Prompts [25]	54	63
vs. LLM Prompts	73	67
vs. Generic Prompts	87	88

Table 2. Pairwise comparison user study (win rates in %) reveals that SliderSpace-extracted artistic styles achieve comparable diversity to manually curated artist [25], while being significantly preferred for creative utilization.

Method	FID-30k (↓)	CLIP (↑)
Real	-	30.14
SDXL	11.72	29.41
SDXL-DMD	15.52	28.92
DMD-SliderSpace	12.12	29.13

Table 3. SliderSpace trained on larger range of knowledge can improve the diversity of the distilled models (FID) while having a good text-image alignment (CLIP).

In Table 3, we evaluate this improvement by measuring FID and CLIP scores on all COCO-30k prompts before and after applying SliderSpace to SDXL-DMD. We find that DMD-SliderSpace improves FID from the distilled version, almost matching the FID of undistilled SDXL.

5.4. Slider Transferability

Using FaceNet [44] as the semantic encoder, we train SliderSpace on the concept "person" and discover interpretable

Figure 5. SliderSpace demonstrates broader artistic style coverage, as evidenced by the lower FID scores compared to both supervised Concept Sliders. Comparison of artistic style diversity using FID scores against reference distribution (a) derived from the complete artist dataset [25]. We compare against outputs from (b) generic art prompts (b), (c) LLM-generated art prompts, and (d) Concept Sliders.

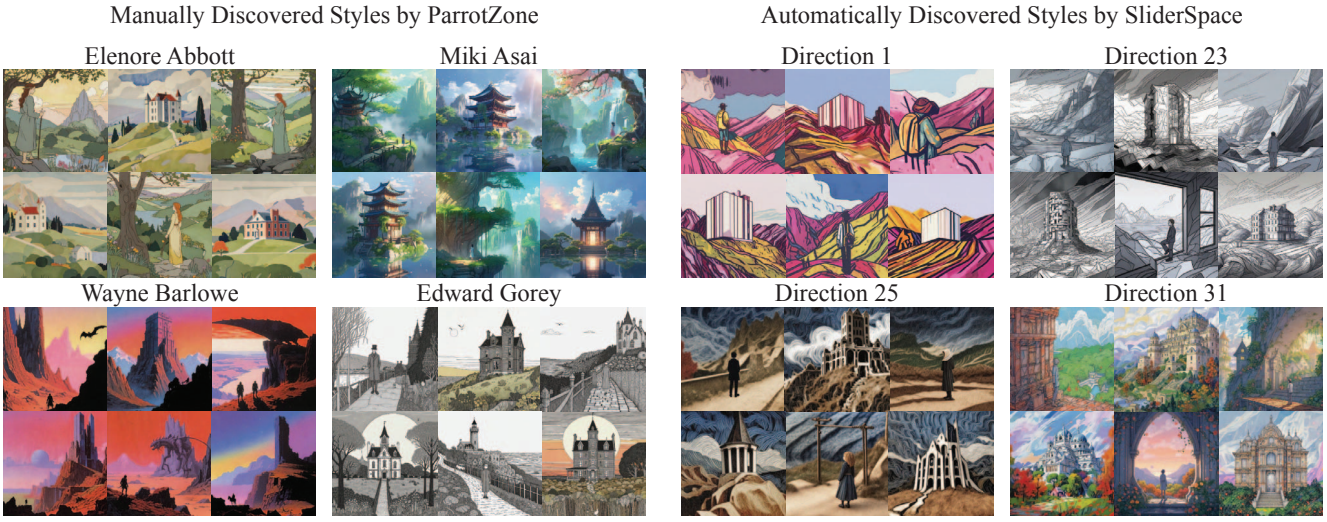

Figure 6. Comparison of artistic styles discovered in SDXL-DMD: (Left) Representative samples from artist-specific prompts manually curated by the Parrotzone community [25] through extensive exploration. (Right) Automatically discovered artistic directions using SliderSpace, which captures diverse and semantically meaningful variations without requiring explicit artist references.

directions controlling attributes like appearance and age (Figure 8). These directions not only transfer effectively to related concepts like "police" and "athlete" but also generalize surprisingly well to out-of-domain concepts like "dog", suggesting that SliderSpace captures fundamental visual transformations in the model's knowledge space.

Additional experiments analyzing hyperparameter choices (App. B.2), alternative semantic embeddings (App. B.1), and ablation studies (App. E) are in Appendix.

6. Limitations

Our method can work with several semantic embeddings - each encoder might introduce different inherent biases present in its training data. While these embeddings enable semantic consistency, they may not capture certain culturally-specific or nuanced concepts. This highlights the need for application specific semantic embedding choice. The current discovery process requires ≈ 65 mins on A100. We note that our method trains 8 times faster than Concept Sliders for same number of sliders. To further optimize, it opens avenues for future research into training time optimizations. For art style discovery, the discovered directions are not one-to-one matched with the real artists, this can be limiting based on users' application. Further work can address discovery that nudges the directions to be aligned with real artists.

7. Conclusion

SliderSpace is a simple framework that automatically decomposes diffusion models' capabilities into semantically meaningful and controllable directions. By leveraging spectral decomposition in semantic space combined with low-rank adaptation, our method enables systematic exploration of a model's latent creative space without requiring manual

Figure 7. SliderSpace can also decompose general broad visual variation of diffusion model's and can be used to overcome model collapse in distilled models. We randomly sample a sparse set of sliders and generate the sample showing higher variation than base distilled model.

Figure 8. SliderSpace directions for the "person" concept successfully generalize to related "police" and "athlete" concepts. They also transfer to out-of-domain concepts like "dog"

attribute specification. Through extensive experiments, we demonstrated SliderSpace's effectiveness across three key applications. First, our concept decomposition revealed interpretable variations within the model's knowledge representation, enabling fine-grained control while maintaining semantic consistency. Second, our exploration of artistic capabilities showed that SliderSpace can discover directions matching or exceeding the diversity of manually curated artist lists, while being rated more useful by human evaluators. Finally, we demonstrate how SliderSpace can help address mode collapse in distilled diffusion models, restoring diversity while preserving computational efficiency.

The ability of SliderSpace to uncover interpretable directions suggests that diffusion models may develop structured internal representations of visual concepts during training, without explicit supervision. By mapping these models' vast creative potential into intuitive, composable directions, our work takes a step toward making their capabilities more transparent and accessible.

Acknowledgment

This work was done by RG at Adobe during their internship. RG and DB are supported by Open Philanthropy and NSF grant #2403304.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 2

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 2

[3] Anthropic. Introducing claude 3.5 sonnet, 2024. 5, 2

[4] BlackForestLabs. Announcing state-of-the-art flux.1 dev and schnell models, 2024. 1, 2, 3, 4, 17

[5] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv e-prints*, pages arXiv–1809, 2018. 1

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[8] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024. 1, 2

[9] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24209–24218, 2024. 2, 3

[10] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models, 2024. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. *arXiv preprint arXiv:2406.09413*, 2024. 2

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2

[14] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 6

[15] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2, 4

[16] 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*, pages 172–188. Springer, 2024. 1, 3, 5

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2

[18] Imke Grabe, Miguel Gonz'alez-Duque, Sebastian Risi, and Jichen Zhu. Towards a framework for human-ai interaction patterns in co-creative gan applications. In *Joint Proceedings of the ACM IUI Workshops*, 2022. 2

[19] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. 2

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[21] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 1, 2

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5, 6

[23] Oliver Hoffmann. On modeling human-computer co-creativity. In *Knowledge, Information and Creativity Support Systems: Selected Papers from KICSS'2014-9th International Conference, held in Limassol, Cyprus, on November 6-8, 2014*, pages 37–48. Springer, 2016. 1

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 2, 3

[25] Surea I, Proxima Centauri B, Erratica, and Stephen Young. Image synthesis style studies, 2022. 5, 6, 7

[26] et al James Betker. Improving image generation with better captions. *OpenAI Reports*, 2023. 2

[27] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10124–10134, 2023. 2

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2

[29] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 3

[30] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[31] Nan Liu, Yilun Du, Shuang Li, Joshua B Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2095, 2023. 3

[32] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1, 2

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 2, 4

[34] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 4, 8

[35] Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[37] Robin Rombach. Stable diffusion 2.0 release, 2022. 2

[38] Robin Rombach and Patrick Esser. Stable diffusion v1-4 model card, 2022. 3

[39] Robin Rombach and Patrick Esser. Stable diffusion v2 model card, 2022. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[41] 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. 1, 3

[42] Simo Ryu. Cloneofsimo/lora: Using low-rank adaptation to quickly fine-tune diffusion models.s. *GitHub*, 2023. 1

[43] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025. 2, 4, 16

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[45] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[46] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 1, 2

[47] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 2

[48] Zongze Wu, Nicholas Kolkin, Jonathan Brandt, Richard Zhang, and Eli Shechtman. Turboedit: Instant text-based image editing. *ECCV*, 2024. 2

[49] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. In *Conference on Computer Vision and Pattern Recognition 2024*, 2024. 2

[50] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2

[51] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 2, 4, 5, 6, 7, 9, 10, 11, 13, 14, 15

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2