Navigating h-space for Multi-Attribute Editing in Diffusion Models

Jinhyeong Park*
School of Computer Science & Engineering
Kyungpook National University
Daegu, South Korea
hini2245@knu.ac.kr

Muhammad Shaheryar*
School of Computer Science & Engineering
Kyungpook National University
Daegu, South Korea
shaheryar@knu.ac.kr

Seangmin Lee
School of Computer Science & Engineering
Kyungpook National University
Daegu, South Korea
smlee0610@knu.ac.kr

Soon Ki Jung†
School of Computer Science & Engineering
Kyungpook National University
Daegu, South Korea
skjung@knu.ac.kr

***Abstract*—Multi-attribute editing in generative models has been a challenging problem, especially in achieving realistic and disentangled transformations across multiple attributes simultaneously. In this work, we propose an approach for multi-attribute editing in the h-space of diffusion models, where multiple attributes such as aging, gender and eyeglasses can be edited simultaneously. Unlike existing methods that require separate models for each attribute or operate in a highly coupled latent space, our method harnesses the power of a unified framework. We learn interpretable attribute directions in the latent space through supervised training, enabling fine-grained control over specific attributes without affecting others. This disentangled editing allows for complex transformations, such as modifying both age and hairstyle while preserving identity. By performing edits in the h-space, we ensure high-quality, coherent transformations, demonstrating the potential for rich and flexible editing capabilities. The ability to perform multi-attribute modifications in a single, unified model opens up new possibilities for applications in computer vision, digital media, and personalized content creation, making our method a significant advancement in generative modeling.**

***Index Terms*—Multi attribute Editing, Image Transformation, Generative Models, Latent space**

I. Introduction

Semantic image editing involves modifying specific attributes of an image, such as changing hair color or simulating aging, while preserving the overall identity and realism. This capability has diverse applications in design, visualization, and targeted data augmentation. Effective editing requires disentangled attribute manipulation and precise control, particularly when handling multiple attributes simultaneously.

Generative adversarial networks (GANs) [2] have demonstrated significant potential in image synthesis and semantic editing, leveraging latent space representations to achieve fine-grained attribute control [3]–[5]. Methods like InterFaceGAN [1] have explored the rich latent properties of GANs, enabling precise manipulation of facial attributes such as age, gender, and expression. Similarly, StarGAN [4] allows multi-domain image-to-image translation, making it possible to modify multiple attributes simultaneously within a single model. Despite their successes, GAN-based approaches often face challenges in disentanglement, limiting their ability to edit multiple attributes independently and precisely.

Denoising Diffusion Models (DDMs) [6] have emerged as a powerful alternative to GANs, offering superior image quality and diversity in synthesis [12]. DDMs operate by iteratively denoising latent representations, which can be computationally expensive when applied directly in pixel space. Latent diffusion models [7] address this issue by operating in a compact latent space, significantly reducing computational costs. However, these models often require fine-tuning or additional training to incorporate new attributes, making them computationally expensive and less practical.

Recently, Kwon et al. [11] introduced h-space, a rich latent representation derived from the deepest feature maps of pre-trained DDMs. This h-space encodes semantically meaningful information, making it particularly suited for interpretable and disentangled editing. Building on this, Boundary Diffusion [8] demonstrated a method for single-attribute editing by defining semantic boundaries in h-space, enabling editing without requiring fine-tuning. While effective, Boundary Diffusion is limited to single-attribute editing, leaving the challenge of multi-attribute manipulation unaddressed.

In this paper, we extend the capabilities of DDMs to enable multi-attribute editing in h-space. We propose a simple yet efficient method that identifies semantic boundaries in h-space using minimal supervision. Unlike previous approaches that rely on extensive attribute annotations or pre-trained classifiers, our method requires only 100 curated image pairs

* Equal contribution.
† Corresponding author.
This work is supported by IITP grant funded by MSIT (IITP-2024-RS-2022-00156389) and the ICT R&D program of MSIT/IITP. (RS-2024-00336663)

for each attribute, containing examples with and without the desired attribute. By projecting these image pairs into h-space and training a linear classifier to define attribute-specific boundaries, we identify dominant directions corresponding to semantic attributes. These learned directions enable realistic and coherent multi-attribute editing for attributes such as age, smile, glasses, and hairstyle. Our approach demonstrates the flexibility and power of leveraging h-space for disentangled and controlled editing, even in a few-shot setting, while advancing the realism in DDM-based semantic editing.

II. Related Work

A. Latent Space and Diffusion Models for Semantic Editing

Latent space plays a critical role in modern image generation and editing tasks, offering a compact and structured representation of high-dimensional data. By leveraging latent space representations, it is possible to isolate and manipulate specific semantic attributes while preserving the underlying details of the image. [5], [9]. This capability makes latent space essential for applications such as multi-attribute editing and targeted transformations.

Generative models, such as GANs, have demonstrated the power of latent spaces for controlling semantic attributes. AttGAN [3] enables facial attribute editing by focusing on specific features, such as age. Also, StarGAN [4] enables multi-attribute editing by performing image-to-image translation across multiple domains. However, they often lack precise disentanglement and require extensive re-training for fine-grained attribute editing. DDMs have emerged as a strong alternative, introducing a well-structured latent space with semantically meaningful representations. This enables precise control and manipulation of attributes without the need for additional training.

B. Diffusion Models and Their Generative Process

DDMs generate high-quality images by reversing a noise-adding process, where the forward process introduces noise step-by-step into a clean image x_0. This process results in a fully noised latent vector x_T, sampled from a Gaussian distribution $\mathcal{N}(0, I)$. Mathematically, the forward process is defined as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \tag{1}$$

where α_t is the noise scheduling factor controlling the level of noise at each step. The reverse process, parameterized by ϵ_θ, predicts the noise ϵ at each step, enabling the recovery of the clean image x_0 from x_T. The objective of the reverse process is to minimize the reconstruction loss over all timesteps:

$$L = \mathbb{E}_{x,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{2}$$

This iterative denoising process not only ensures high-quality synthesis but also facilitates flexibility in semantic manipulation by working with intermediate latent representations.

C. The Emergence of h-Space in Diffusion Models

While DDMs traditionally operate in pixel space, recent advancements by Kwon et al. [11] introduced h-space, a latent space derived from the bottleneck feature maps of U-Net architectures within DDMs. h-space offers a semantically enriched and disentangled representation with the following unique properties:

- **Consistency Across Samples:** A direction Δh_t produces similar semantic changes across different images, enabling generalized attribute control.
- **Intensity Control:** The magnitude of Δh_t directly influences the degree of semantic change, allowing for fine-grained attribute manipulation.
- **Additivity:** Linear combinations of different directions $(\Delta h_{t,1}, \Delta h_{t,2}, \ldots)$ enable simultaneous and independent editing of multiple attributes, making h-space particularly effective for multi-attribute control.

D. Semantic Editing in h-Space

Semantic editing in h-space is achieved by defining attribute-specific boundaries and applying offsets Δh_t during the denoising process. Given the arithmetic nature of h-space, these offsets align closely with semantic directions, allowing precise manipulation. For instance, the edited latent representation is given as:

$$h_{\text{edit}} = h + \sum_j \alpha_j \Delta h_{t,j}, \tag{3}$$

where α_j controls the intensity of the attribute modification, and $\Delta h_{t,j}$ represents the learned direction for a specific attribute j. This formulation ensures that edits are disentangled and semantically coherent, preserving the overall identity of the original image.

E. Applications and Advancements in h-Space

The structured properties of h-space address the limitations of traditional approaches, such as Boundary Diffusion [8], which focuses on single-attribute editing. By leveraging linear combinations of directions, h-space enables efficient multi-attribute editing, demonstrating its potential for advanced applications such as face de-identification, targeted editing, and multi-domain transformations.

Overall, DDMs and their latent h-space offer a robust framework for high-quality, disentangled, and interpretable image editing. This approach not only reduces computational overhead but also introduces a scalable solution for complex attribute manipulation, setting a new benchmark for semantic editing tasks.

III. Proposed Approach

This work focuses on uncovering semantic directions within the h-space of DDMs, that enables precise and interpretable editing. This section first introduces our supervised method to find interpretable directions in DDMs' h-space. In the second part, we show how to utilize discovered directions in the inference process for responsible Multi-attribute editing.

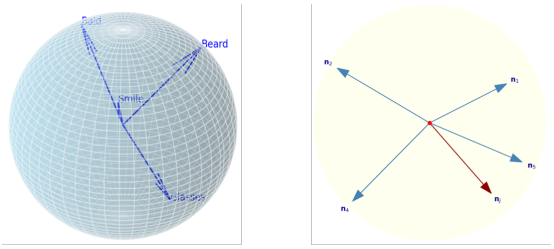

Fig. 1. (a) The attribute directions $\mathbf{n}_j$ are represented on a latent space, visualized as a spherical manifold, where each vector corresponds to a semantic attribute such as "Beard," "smile," or "Glasses" (b) A hyperplane representation is shown, where the attribute directions $\mathbf{n}_j$ are derived by training linear classifiers to distinguish the presence or absence of specific attributes. The red dot represents the latent point of an image, which can be moved along any learned direction $\mathbf{n}_j$ to manipulate the corresponding attribute.

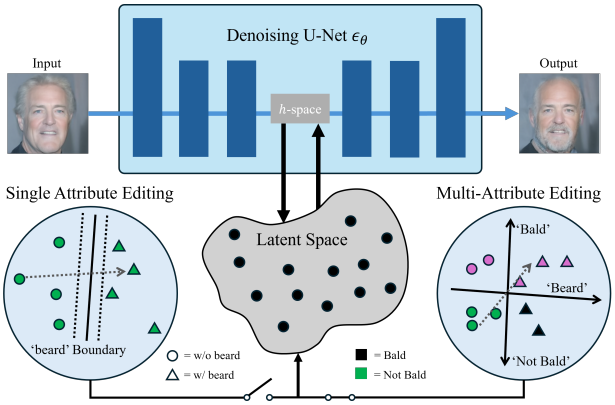

Fig. 2. Framework Overview: The proposed methodology operates within the latent h-space of the DDM. The top section illustrates the denoising process of the U-Net, where h-space captures hierarchical semantic information. Multi-attribute editing is performed by perturbing the latent representations in$h$-space, guided by boundaries learned for individual attributes (e.g., "beard" and "bald"). The bottom section shows how attribute boundaries and semantic directions are identified and applied to achieve coherent and disentangled transformations.

The proposed methodology focuses on leveraging the latent h-space of the DDM for multi-attribute editing. During the denoising process, h-space serves as a critical intermediate representation, encoding rich semantic details of the image. Attribute-specific boundaries (e.g., "beard" or "bald") are identified in the latent space using labeled examples. Multi-attribute editing is achieved by perturbing the h-space representation along these boundaries, enabling simultaneous and independent control over multiple attributes. The learned attribute directions ensure disentanglement, allowing the model to edit features like aging, facial hair, or hairstyle without compromising the coherence or identity of the subject. This flexible approach demonstrates the ability to handle complex and realistic transformations.

IV. Problem Statement

The objective of this work is to enable controllable semantic image editing in the latent h-space of DDMs. Given N attributes, the goal is to discover semantically meaningful directions $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_N\}$ in the h-space. Each direction $\mathbf{n}_j \in \mathbb{R}^m$ corresponds to a specific attribute a_j and enables the manipulation of that attribute during the denoising process. The modification of a latent representation is expressed as:

$$h_{\text{edit}} = h + \alpha_j \mathbf{n}_j, \tag{4}$$

where h is the original latent representation, α_j is a scalar controlling the intensity of the change, and $\mathbf{n}_j$ represents the direction associated with attribute a_j. The objective is to apply these modifications while preserving the overall structure and coherence of the generated image.

V. Finding Directions

To uncover interpretable directions in h-space, we utilize labeled data with binary labels indicating the presence (1) or absence (-1) of specific attributes. Linear classifiers (SVM) are trained on the latent representations to separate regions in the h-space corresponding to these attributes. Each classifier imposes a hyperplane that separates regions associated with the presence and absence of a given attribute. The direction orthogonal to this hyperplane defines the semantic direction $\mathbf{n}_j$ for that attribute.

For attribute manipulation, the latent representation is updated as:

$$h_{\text{edit}} = h + \alpha_j \mathbf{n}_j, \tag{5}$$

where traversing along $\mathbf{n}_j$ increases or decreases the presence of the attribute a_j. For example, this approach can adjust features like smiling, facial hair, or aging in a disentangled manner.

A. Multi-Attribute Editing

For multi-attribute editing, multiple semantic directions $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\}$ are combined. The modified representation is given by:

$$h_{\text{edit}} = h + \sum_{j=1}^{m} \alpha_j \mathbf{n}_j, \tag{6}$$

where α_j controls the intensity of the contribution of each attribute a_j. This enables simultaneous and independent manipulation of multiple attributes, such as making a person appear younger while adding glasses.

VI. Real Image Editing for Multi-Attribute Modifications

We extend the editing approach to real images by first mapping them into the latent h-space using DDIM inversion. This allows edits to be applied in h-space in a manner consistent with generated images. The editing process for real images is formulated as:

$$\bar{\epsilon}_\theta(x_t, d_e) = \epsilon_\theta(x_t, \phi) + \lambda_e \left(\epsilon_\theta(x_t, d_e) - \epsilon_\theta(x_t, \phi)\right), \tag{7}$$

where $\epsilon_\theta(x_t, \phi)$ represents the original denoising process, $\epsilon_\theta(x_t, d_e)$ incorporates the semantic modifications via direction d_e, and λ_e controls the strength of the edits. By

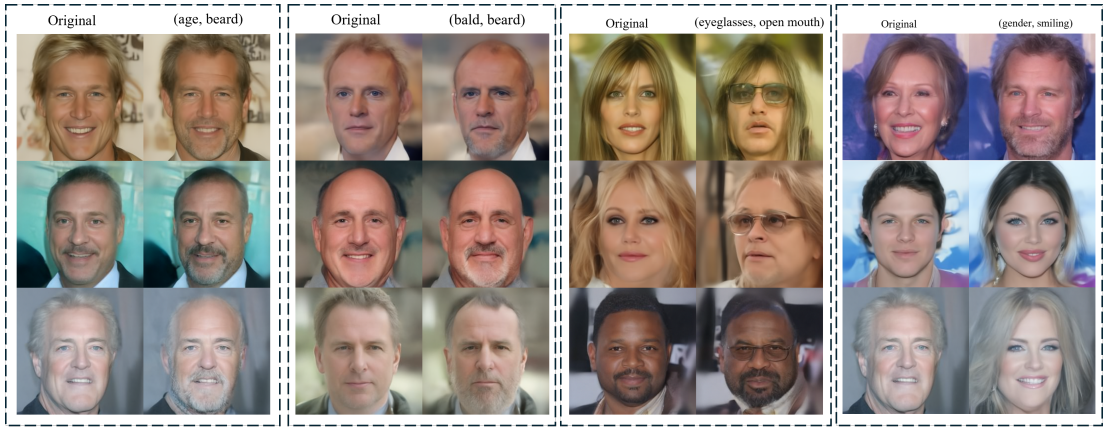

Fig. 3. Multi-Attribute Editing Results: (a) Aging with beard addition illustrates natural and identity-preserving transformations. (b) Baldness paired with beard modification demonstrates smooth and coherent attribute integration. (c) Eyeglasses and mouth expression changes showcase fine-grained control over distinct facial features. (d) Gender and smile adjustments highlight the versatility and accuracy of the model in managing diverse and intricate edits.

Fig. 4. Simultaneous modifications of multiple attributes: the first column displays the original images, the second column showcases editing results with two attributes, and the third column highlights edits involving three attributes, including aging, beard addition, and male features, demonstrating consistent and realistic transformations.

combining multiple directions d_e, we enable coherent and simultaneous edits of attributes such as hairstyle, expression, and accessories.

A. Advantages of the Latent Space Approach

This approach leverages the inherent structure of the h-space for precise and interpretable edits without requiring pixel-level supervision or synthetic datasets. By identifying disentangled directions for each attribute, the method ensures that attributes remain independent, even during multi-attribute modifications. The framework excels at maintaining semantic coherence and generating realistic transformations, making it versatile for real-image and generated-image editing tasks.

VII. Experiments

We evaluated the effectiveness of our multi-attribute editing approach by identifying semantic directions in the h-space of a pretrained DDPM model trained on CelebA [14]. The h-space, defined by bottleneck activations across T timesteps, captures hierarchical semantic features. Input images in pixel space $(3, 256, 256)$ are mapped to the deepest feature map in h-space $(T, 512, 8, 8)$, where edits are applied by introducing perturbations $(\Delta h_{T:1})$ to the latent representations during generation.

We conducted experiments on various attribute combinations, such as "smile," "wearing glasses," and "age," to validate the generalization of our method. By visualizing results, we demonstrated precise multi-attribute editing with disentangled and independent control of each attribute, confirming the scalability and robustness of the approach across different attribute combinations.

VIII. Qualitative Visualization

The qualitative results, presented in fig. 3, highlight the versatility and effectiveness of our multi-attribute editing framework. In fig. 3, (a) and (b) demonstrate complex transformations such as aging combined with beard addition and baldness respectively, showcasing the model's ability to capture realistic, disentangled edits. fig. 3, (c) illustrates the seamless integration of eyeglasses and mouth expression modifications, with subtle changes applied naturally to maintain coherence. In fig. 3, (d), simultaneous gender and smile transformations reveal the system's capability to independently control facial features while ensuring smooth transitions. Lastly, fig. 4, highlights multi-attribute edits across various attribute combinations, such as old, beard, and male, emphasizing the robustness of the approach to handle diverse and intricate modifications. Collectively, these visualizations demonstrate the model's ability to apply precise, realistic, and identity-preserving edits across a range of attributes and scenarios.

Moreover, fig. 5 illustrates the capability of controlling attribute intensity in multi-attribute editing. The progression from left to right demonstrates incremental modifications in attributes such as aging and beard, showcasing fine-grained control while maintaining the subject's identity and coherence.

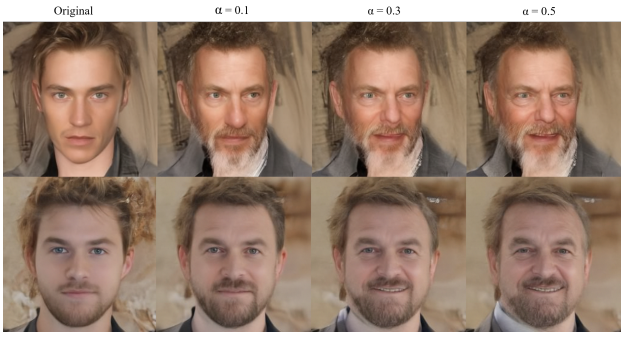

Fig. 5. Demonstration of intensity control for multi-attribute editing, where attributes such as aging and beard are progressively modified with increasing intensity from left to right, while preserving identity and realism.

IX. Conclusion

In this paper, we propose a novel multi-attribute editing mechanism based on h-space manipulation within a symmetrical U-Net-like architecture of DDMs. By leveraging interpreted semantics and a conditional manipulation technique, our method enables precise and independent control over facial attributes using any pre-trained DDM. This effectively transforms unconditional DDMs into versatile tools for controllable editing. Our approach not only enhances attribute editing accuracy but also improves detail preservation, achieving a superior balance between reconstruction fidelity and attribute modification. Extensive experiments demonstrate its strong multi-attribute editing capabilities and potential for real-image editing, showcasing its practicality and flexibility.

Acknowledgment

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2024-RS-2022-00156389, 50%) and the ICT R&D program of MSIT/IITP. [RS-2024-00336663, Development of AI technology for tracking and investigating drug criminals through multimedia sources] ※ MSIT: Ministry of Science and ICT

References

[1] Shen, Yujun, et al. "Interfacegan: Interpreting the disentangled face representation learned by gans." IEEE transactions on pattern analysis and machine intelligence 44.4 (2020): 2004-2018.

[2] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.

[3] He, Zhenliang, et al. "Attgan: Facial attribute editing by only changing what you want." IEEE transactions on image processing 28.11 (2019): 5464-5478.

[4] Choi, Yunjey, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[5] Li, Bingchuan, et al. "DyStyle: Dynamic neural network for multi-attribute-conditioned style editings." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[6] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

[7] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[8] Zhu, Ye, et al. "Boundary guided learning-free semantic control with diffusion models." Advances in Neural Information Processing Systems 36 (2024).

[9] Kwon, Mingi, Jaeseok Jeong, and Youngjung Uh. "Diffusion models already have a semantic latent space." arXiv preprint arXiv:2210.10960 (2022).

[10] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in neural information processing systems 34 (2021): 8780-8794.

[11] Kwon, Mingi, Jaeseok Jeong, and Youngjung Uh. "Diffusion models already have a semantic latent space." arXiv preprint arXiv:2210.10960 (2022).

[12] Nichol, Alexander Quinn, and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models." International conference on machine learning. PMLR, 2021.

[13] Haas, René, et al. "Discovering interpretable directions in the semantic latent space of diffusion models." 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, 2024.

[14] "Hugging Face." https://huggingface.co/google/ddpm-ema-celebahq-256.