GAN-based Transfer of Interpretable Directions for Disentangled Image Editing in Text-to-Image Diffusion Models

Yusuf Dalva **Hidir Yesiltepe** **Pinar Yanardag**
{ydalva, hidir, pinary}@vt.edu
Virginia Tech

Abstract

The rapid advancement in image generation models has predominantly been driven by diffusion models, which have demonstrated unparalleled success in generating high-fidelity, diverse images from textual prompts. However, these models are often characterized as black boxes due to their complex, less-understood mechanisms, highlighting a significant gap in interpretability research. In contrast, Generative Adversarial Networks (GANs) are praised for their well-structured latent spaces that offer rich semantics, enabling more straightforward exploration and understanding of model behaviors. `GAN2Diff` bridges this gap by transferring the structured, interpretable latent directions from pre-trained GAN models—representative of specific, controllable attributes—into diffusion models. This approach enhances the interpretability of diffusion models, preserving their generative quality while providing new avenues for exploring and manipulating complex image attributes.

1 Introduction

Denoising Diffusion Models (DDMs) [12] and Latent Diffusion Models (LDMs) [27] gained popularity in generative modeling landscape due to the capability to generate high-quality, high-resolution images across diverse domains. Their performance, particularly highlighted by text-to-image models like Stable Diffusion [27], has led researchers to leverage them for image editing tasks. These tasks range from text prompt-driven edits to modifications based on scribbles or segmentation maps [39], underpinning a growing interest in using DDMs and LDMs for fine-grained image manipulation.

Despite their success, these models remain black boxes, limiting interpretability and control. Advancing interpretability in generative models requires understanding their semantic structures, particularly how specific attributes can be manipulated within the generative process. GANs excel in this area, featuring well-structured and informative latent spaces [37, 10, 28]. Studies have identified up to 2000 semantically meaningful latent directions in GANs [35], enabling nuanced control and deep understanding of various image aspects. In contrast, current interpretability research on diffusion models has only discovered a handful of latent directions [6, 18]. This disparity stems from the inherently more complex architecture of diffusion models, which involve independent forward noise estimation and management of numerous latent variables across multiple recursive timesteps. As a result, mapping disentangled directions within diffusion models presents a significantly greater challenge compared to GANs. This stark difference in interpretability highlights the need for further research into diffusion models' latent spaces, aiming to bridge the gap with the rich latent architecture of GANs and enhance our ability to explore and manipulate image attributes in these powerful but opaque systems.

To bridge this gap, we introduce `GAN2Diff`, a novel framework designed to transfer the comprehensive and structured latent capabilities of GANs to the generative domain of large-scale text-to-image

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

diffusion models. By incorporating the detailed latent directions from GANs into diffusion models, we aim not only to enhance the interpretability of these models but also to unlock rich editing capabilities that were previously constrained by the limited understanding of their latent spaces. Our contributions are outlined as follows: (1) Our approach represents the first study to transfer latent semantics from a pre-trained GAN model to a pre-trained text-to-image diffusion model without finetuning. (2) Our approach showcases the capability to transfer a wide range of fine-grained semantics spanning various categories, including faces, cats and dogs. Notably, we extend the application of transferred semantics beyond simple headshots—a common limitation in GAN-based methods—to real-world examples, showcasing the broader applicability and robustness of our approach. (3) Our experiments show that our method transfers semantics comparable to state-of-the-art diffusion-based and GAN-based latent space discovery methods. (4) Our method demonstrates versatility in semantic transfer, drawing from diverse GAN techniques such as StyleSpace [35] and StyleGAN-NADA [8], as well as diffusion-based approaches like Prompt2Prompt [11]. (5) Our method demonstrates versatility in semantic transfer, drawing from diverse GAN techniques such as StyleSpace [35] and StyleGAN-NADA [8], as well as diffusion-based approaches like Prompt2Prompt [11]. (6) We share our source code along with the discovered directions to enable further research in this area.

2 Method

Our proposed method `GAN2Diff` aims to learn a latent direction d formulated as a conditional embedding, that represents the edit done by a GAN model such as StyleGAN. In order to achieve this task, we initially populate a small dataset consisting N image pairs corresponding to images generated by StyleGAN using style channels Δs, $\mathcal{G}(s)$, and their edited co-variants, $\mathcal{G}(s + \Delta s)$. Throughout our framework we label these image sets as $\mathcal{X}_{input} = \{x_1, \cdots, x_N\}$ and $\mathcal{X}_{edited} = \{x'_1, \cdots, x'_N\}$. Using these image sets, we formulate our overall loss function to learn the latent direction d with two objectives, which targets both the semantic level differences and latent level differences between the image pairs. These two objectives are described below.

Semantic Alignment Loss. To learn a latent direction that semantically aligns with the difference between the difference between the image sets $\mathcal{X}_{input}$ and $\mathcal{X}_{edited}$, we introduce a semantic alignment loss $\mathcal{L}_{sem}$. Using the CLIP [26] Image Encoder, we define our objective that aims to learn the embedding for d from the contrast between the two image sets. Fundamentally, for a latent direction that represents the difference between the image sets $\mathcal{X}_{input}$ and $\mathcal{X}_{edited}$, the similarity between $x' \in \mathcal{X}_{edited}$ and the direction d should be maximized, whereas it should be minimized for $x \in \mathcal{X}_{input}$. To reflect this behavior, we use the difference between the similarity values of x' and x with direction d, where we maximize the similarity value for x' and minimize for x. Furthermore, we enforce our method to learn the desired semantic change only from the paired images generated by the StyleGAN generator. We formulate the semantic alignment loss as $\mathcal{L}_{sem} = 1 - cossim(E_I(x'), d) + cossim(E_I(x), d)$

Direction Alignment Loss. Complementary to $\mathcal{L}_{sem}$, we introduce the second optimization term $\mathcal{L}_{dir}$ by utilizing the information encoded by the noise predictions across timesteps, which can capture fine-grained details for the given input-edit pair (x, x') where $x \in \mathcal{X}_{input}$ and $x' \in \mathcal{X}_{edited}$. Since our training data involves images that are identical, expect the semantic that we would like to learn with direction d, we expect the optimized direction d to give a strong response to the difference in noise predictions $||\epsilon_\theta(x'_t, d) - \epsilon_\theta(x_t, d)||_2^2$, as the optimized semantic reflects the difference between x' and x, where x_t denotes the noised variant of image x at timestep $t \in \mathcal{U}(1, T)$. With the motivation to use the understanding capabilities of the diffusion model complementary to CLIP [26], we formulate $\mathcal{L}_{dir}$ as the difference of noise predictions w.r.t. d with a negative sign, to maximize it. We formulate the direction alignment loss as $\mathcal{L}_{latent} = -E_{x_0, \epsilon^t \sim \mathcal{N}(0,1), t}[||\epsilon_\theta(x'_t, d) - \epsilon_\theta(x_t, d)||_2^2]$

The final loss to optimize the direction d, is the sum of semantic and direction alignment terms, $\mathcal{L} = \mathcal{L}_{sem} + \mathcal{L}_{latent}$.

2.1 Image Editing

Given a latent direction d, we perform image editing in a way that it successfully reflects the desired semantic to the input image in a disentangled manner. In order to achieve this, we expand the classifier-free guidance provided in Eq. 3 with the direction we optimize. Following the editing schemes provided by [6, 1] we adopt the classifier-free guidance with an editing term

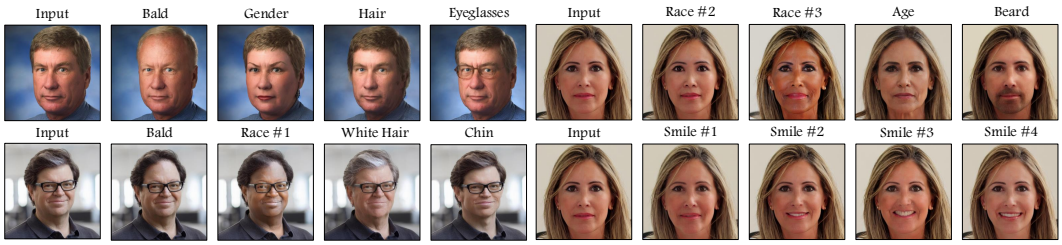

Figure 1: **Qualitative Results**. GAN2Diff successfully transfers editing directions that modify the overall look, including changes in *race* or *aging*, as well as more detailed edits that target specific facial attributes, such as *eyeglasses* or a *beard*. GAN2Diff can also distinguish among various edits for the same feature underlines the versatility of our approach, providing users with an extensive selection of editing options for individual characteristics, like multiple smile designs (see row 2) or styles of baldness (as shown in Rows 1 and 2).

based on d, which is the latent direction we learn during the optimization process. Briefly, we add an additional classifier-free guidance term corresponding to editing in which we formulate as $\bar{\epsilon_\theta}(x_t, c, d) = \tilde{\epsilon_\theta}(x_t, c) + \lambda_e(\epsilon_\theta(x_t, d) - \epsilon_\theta(x_t, \phi))$ for the timesteps where the edit is going to be applied. We denote the editing scale with λ_e, the image condition with c and the editing direction with d. Note that our editing scheme does not involve adding any additional input other than the conditional embedding optimized for the editing task, which is used as the input to the text embedding for the noise prediction residual applied on the edit.

Editing with multiple directions. To extend our editing scheme for multiple edits, we perform a sum over multiple editing directions with their corresponding editing scales. For a set of directions $D = \{d_1, \cdots, d_k\}$, which are going to be applied at a given timestep, we expand our editing formulation with a summation term over direction set D as $\tilde{\epsilon_\theta}(x_t, c) + \sum_{i=1}^{|D|} \lambda_{e_i}(\epsilon_\theta(x_t, d_i) - \epsilon_\theta(x_t, \phi))$

Real Image Editing. In addition to performing edits on generated images, the directions learned by GAN2Diff can also be used to edit real images. To do so, we adopt the DDPM Inversion [15] to obtain the series of noisy latents, which eventually gives x_T as the initial latent. After inverting x_T unconditionally, we reformulate the overall noise prediction $\epsilon_\theta(x_t, d)$ for real image editing, where d denotes the editing direction: $\bar{\epsilon_\theta}(x_t, d) = \epsilon_\theta(x_t, \phi) + \lambda_e(\epsilon_\theta(x_t, d) - \epsilon_\theta(x_t, \phi))$

Note that the approach for editing with multiple directions is also applicable to real image editing, by replacing the edit guidance term $\lambda_e(\epsilon_\theta(x_t, d) - \epsilon_\theta(x_t, \phi))$ with the summation provided as our multiple direction editing formulation.

3 Experiments

To evaluate the effectiveness of GAN2Diff in transferring semantically meaningful latent directions and to demonstrate the generalizability of our method, we evaluate our method in various domains, such as human faces, cats, and dogs. We also demonstrate the generalizability behavior of our framework on different images edited with representations different than style space $\mathcal{S}$.

Experimental Setup. We used Stable Diffusion-v1.5 for all of our experiments. We used StyleGAN2 [17] trained on various datasets, including FFHQ [16], AFHQ-Cats [5], and AFHQ-Dogs [5]. In our default setting, we train GAN2Diff with $N = 1000$ samples for each direction that we would like to transfer. To optimize, we use the AdamW optimizer [21] for 1000 iterations. Training our method on a single domain requires approximately 30 minutes, and once trained, zero-shot image editing takes about 5 seconds using a single NVIDIA L40 GPU.

3.1 Qualitative Results

Our method leverages a single pre-trained diffusion model to transfer latent directions across different domains. Given the significant variability in facial features and the popularity of face editing in both GAN and diffusion-based models, we initially explore the potential for face editing in directions uncovered by GAN2Diff. As illustrated in Fig. 1, our technique showcases a range of editing

capabilities, from comprehensive changes that alter the face's overall appearance, such as *race* or *aging*, to more fine-grained adjustments targeting specific facial features, such as *eyeglasses* or a *beard*. Our approach can transfer a variety of editing directions that belong to the same semantics, for example, different styles of bangs, hairstyles, or degrees of baldness. In particular, our method is capable of distinguishing between very detailed variations within the same editing task; for example, when provided with four distinct GAN directions for varying smile edits, `GAN2Diff` successfully learned to differentiate between them. This highlights our method's adaptability, offering users a wide range of options for a single attribute, such as various smile types (row 2) or baldness styles (refer to rows 1 and 2). We also note that a key feature of our edits is their high degree of disentanglement, ensuring that only the targeted modifications are made without altering other unrelated aspects. In addition, our method's effectiveness extends to domains beyond human faces, including cats and dogs. The qualitative results depicted in Fig. 10(b) demonstrate `GAN2Diff`'s ability to grasp and apply a wide array of semantic variations across different domains.

3.2 Quantitative Results

Method	LPIPS↓	CLIP-T↑	DINO↑	SigLIP-T↑	DreamSim↑
SEGA [1]	0.179±0.07	0.388±0.03	0.714±0.13	0.134±0.02	0.757±0.09
Prompt2Prompt [11]	0.074±0.05	**0.408±0.29**	0.867±0.09	0.143±0.02	0.869±0.07
InstructPix2Pix [3]	0.059±0.05	0.403±0.03	0.851±0.13	**0.145±0.02**	0.877±0.10
Concept Sliders [9]	0.121±0.06	0.325±0.06	0.842±0.10	0.097±0.03	0.844±0.08
Ours	**0.030±0.01**	0.407±0.03	**0.929±0.05**	0.139±0.02	**0.905±0.05**

Table 1: **Quantitative Comparisons with Diffusion-based Editing Methods.** We compare the editing performance of the directions learned by `GAN2Diff` with diffusion based methods using LPIPS [40], DINO [22], DreamSim [7], CLIP-T [26] and SigLIP-T [38] metrics. In our experiments, we compare our framework with SEGA, Prompt2Prompt, InstructPix2Pix and Concept Sliders. In addition to these metrics, we also conduct a user study on the disentanglement properties of our framework.

Quantitative Comparisons with Diffusion-based Editing Methods. To quantitatively evaluate the directions learned by `GAN2Diff` against diffusion-based editing methods, we compare our method on images edited with "Asian" and "Smile" semantics, over 200 input-edit pairs. For each edit, we generate a set of 100 input-edit pairs, where the editing performance is assessed over synthetic images. For each method, we use their default editing parameters provided in their publicly released code, to not create any unfair advantage with extensive parameter fine-tuning. In our benchmark, we follow the same evaluation protocol as our ablation studies and evaluate the faithfulness to the input image with LPIPS [40], DINO [22] and DreamSim [7] metrics, and alignment with the target semantic with CLIP-T [26] and SigLIP-T [38] text-to-image similarity scores. Note that our edits do not have access to a target editing prompt or concept and apply the edits based on the direction transferred from GANs. We present the quantitative results in Tab. 1. Our evaluations demonstrate that `GAN2Diff` outperforms the competing methods in terms of preservation of the input characteristics, and is on par with diffusion-based editing methods in terms of alignment to the target editing prompt. For qualitative results regarding comparisons with diffusion-based editing methods, see the supplementary material.

4 Discussion

Conclusion In this paper, we introduce a novel approach that capitalizes on the strengths of GANs, known for their disentangled latent spaces and powerful manipulation capabilities, and harmonizes them with the exceptional image generation abilities of diffusion models. Our method aims to bring the best of both worlds, transferring directions from GAN models to exploit the generative capacity of text-to-image diffusion models like Stable Diffusion. This strategic combination not only delivers editing capabilities that are competitive in both diffusion-based and GAN-based image editing techniques, but also significantly refines the precision of the image generation process.

References

[1] Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: SEGA: Instructing text-to-image models using semantic guidance. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=KIPAIy329j

[2] Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. arXiv preprint arXiv:2311.16711 (2023)

[3] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)

[4] Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22560–22570 (October 2023)

[5] Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)

[6] Dalva, Y., Yanardag, P.: Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. arXiv preprint arXiv:2312.05390 (2023)

[7] Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)

[8] Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators (2021)

[9] Gandikota, R., Materzyńska, J., Zhou, T., Torralba, A., Bau, D.: Concept sliders: Lora adaptors for precise control in diffusion models. arXiv preprint arXiv:2311.12092 (2023)

[10] Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. arXiv preprint arXiv:2004.02546 (2020)

[11] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)

[12] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[13] Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

[14] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

[15] Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. arXiv preprint arXiv:2304.06140 (2023)

[16] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)

[17] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)

[18] Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)

[19] Li, X., Hou, X., Loy, C.C.: When stylegan meets stable diffusion: a $\mathcal{W}_+$ adapter for personalized image generation. arXiv preprint arXiv:2311.17461 (2023)

[20] Liu, N., Du, Y., Li, S., Tenenbaum, J.B., Torralba, A.: Unsupervised compositional concepts discovery with text-to-image generative models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2085–2095 (October 2023)

[21] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[22] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[23] Park, Y.H., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. arXiv preprint arXiv:2307.12868 (2023)

[24] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. arXiv preprint arXiv:2103.17249 (2021)

[25] Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[26] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

[27] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[28] Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)

[29] Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1532–1540 (2021)

[30] Simsar, E., Kocasari, U., Er, E.G., Yanardag, P.: Fantastic style channels and where to find them: A submodular framework for discovering diverse directions in gans. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4731–4740 (2023)

[31] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[32] Song, K., Han, L., Liu, B., Metaxas, D., Elgammal, A.: Stylegan-fusion: Diffusion guided domain adaptation of image generators. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5453–5463 (2024)

[33] Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1921–1930 (June 2023)

[34] Wu, C.H., De la Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7378–7387 (2023)

[35] Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. arXiv preprint arXiv:2011.12799 (2020)

[36] Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3121–3138 (2022)

[37] Yüksel, O.K., Simsar, E., Er, E.G., Yanardag, P.: Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14263–14272 (October 2021)

[38] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)

[39] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

[40] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

A Related Work

Latent Space Exploration of Diffusion Models. Empowered by pre-trained text encoders as the conditioning approach, text-to-image diffusion models encode semantically rich information in their latent spaces, which is utilized to generate high-quality images, given a text condition. With the motivation to uncover the semantics represented by this feature space, exploration of the latent space of diffusion models experienced significant advances by recent research [18, 34]. Among the proposed methods, [18] proposes an approach focused on the representations encoded in the bottleneck block of the denoising network, where [34] proposes modifying the latent representations across translation between different imaging domains. In addition to these previous efforts, [23] proposed a method to identify latent directions targeted by an input semantic, inspired by latent space exploration studies in GANs. Despite succeeding in domain-specific diffusion-based models such as DDPMs, such approaches cannot generalize well over large-scale diffusion models such as Stable Diffusion. Addressing this gap, [20] proposes decomposing latent variables into energy functions of semantic significance, whereas the decomposition represents concepts such as lighting and camera position. In a different approach that tackles the latent space of large-scale diffusion models, [6] proposes a contrastive learning objective to discover disentangled latent directions. However, these methods still operate within the complex latent space of diffusion models, making it challenging to achieve fine-grained interpretability and precise control over generated outputs. Our method aims to bridge this gap by transferring semantics from GAN models that are celebrated by their well-understood latent spaces.

Combining GAN and Diffusion Models. Recent studies such as [32] aims to use a pre-trained text-to-image diffusion model as a training objective to adapt a GAN generator to another domain. However, they do not transfer directions between GAN models and diffusion models. Studies such as $\mathcal{W}+$ Adapter [19] and Concept Sliders [9], have also explored leveraging the disentangled image editing capabilities of the StyleGAN model. However, these approaches differ significantly from ours in terms of both their main goals and their methodologies. Concept Sliders [9] focus on finetuning the Stable Diffusion model using LoRAs [14] in order to learn specific concepts where one use case was demonstrated as using a pre-trained StyleGAN to generate paired images. Unlike this method, our approach involves transferring directions from StyleGAN to a single pre-trained Stable Diffusion model, eliminating the need for fine-tuning or training separate LoRA models. The $\mathcal{W}+$ Adapter [19] seeks to utilize the StyleGAN model for image editing on individual images. Unlike our approach, which learns directions applicable to any given image, the $\mathcal{W}+$ Adapter finetunes the Stable Diffusion so that it can edit an image on $w+$ vectors of StyleGAN and *transfer the edit per image* to diffusion model. In contrast, our method can *transfer directions* from StyleGAN to the diffusion model, which can perform edits within the diffusion model.

B Background

GANs The StyleGAN2 generation process consists of several latent spaces, namely $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{W}+$, and $\mathcal{S}$. More formally, let $\mathcal{G}$ denote a generator acting as a mapping function $\mathcal{G} : \mathcal{Z} \rightarrow \mathcal{X}$ where $\mathcal{X}$ is the target image domain. The latent code $\mathbf{z} \in \mathcal{Z}$ is drawn from a prior distribution $p(\mathbf{z})$, typically chosen as Gaussian. The $\mathbf{z}$ vectors are transformed into an intermediate latent space $\mathcal{W}$ using a mapper function consisting of 8 fully connected layers. The latent vectors $\mathbf{w} \in \mathcal{W}$ are then transformed into channel-wise style parameters, forming the *style space*, denoted $\mathcal{S}$, which is the latent space that determines the style parameters of the image. This particular space provides extensive editing possibilities, with each style channel governing a specific attribute modification, such as *smile, eye color,* or *hair type*. Essentially, this means that targeted adjustments to the channel-wise style parameters can facilitate precise and disentangled alterations to an image. In our work, we use the directions in the style space $\mathcal{S}$ identified by previous work [35, 30].

Diffusion Models Diffusion models [12, 31, 27], create data samples using an iterative denoising procedure, commonly referred to as the reverse process. This process operates on a sequence of noise levels $t \in \{1, ..., T\}$, with $\epsilon^t = \alpha^t \epsilon$ where ϵ is drawn from a normal distribution $\mathcal{N}(0, 1)$. The role of the denoising network, denoted as ϵ_θ, is to predict the noise component ϵ in the noised image x_t during the reverse process. Here, x_t symbolizes the noised variant of the original image x_0, subjected to a noise level of ϵ^t. The training of such a denoising network revolves around an objective function that is structured as follows:

$$\mathcal{L}_{DM} = E_{x_0,\epsilon^t\sim\mathcal{N}(0,1),t}[||\epsilon^t - \epsilon_\theta(x_t, t)||_2^2] \tag{1}$$

To produce an image with the denoising network ϵ_θ, the reverse process begins with an initial input x_T, which is sampled from a normal distribution $\mathcal{N}(0, 1)$. During the reverse diffusion procedure, the variable x_t undergoes a series of iterative denoising steps to gradually approach x_0, for each noise level t ranging from 1 to T. This iterative denoising process is mathematically represented by Eq. 2, which is defined for a given step size γ and a specific timestep t.

$$x_{t-1} = x_t - \gamma\epsilon_\theta(x_t, t) + \xi,\ \xi \sim \mathcal{N}(0, \sigma_t^2 I) \tag{2}$$

Classifier-free guidance, introduced by [13], facilitates conditioned sampling by making nuanced modifications to both the forward and reverse diffusion processes based on a specific condition c. By adapting the training of the denoising network ϵ_θ to be compatible with classifier-free guidance, it becomes feasible to generate images conditionally. This is achieved by adjusting the standard noise prediction $\epsilon_\theta(x_t)$ to incorporate the condition, resulting in a conditional noise prediction denoted as $\tilde{\epsilon_\theta}(x_t, c)$. For the sake of clarity, the notation $\epsilon_\theta(x_t)$ is used here to indicate the predicted noise at timestep t, with the understanding that t is implicitly indicated by the variable x_t. The formulation for the noise prediction under classifier-free guidance, $\tilde{\epsilon_\theta}(x_t, c)$, is given by:

$$\tilde{\epsilon_\theta}(x_t, c) = \epsilon_\theta(x_t, \phi) + \lambda_g(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)) \tag{3}$$

where λ_g is guidance scale and ϕ is null-text.

C Ablation Studies

We conducted ablation studies that focus on the number of time steps, the number of samples from the GAN direction intended for transfer, and the distinct components of our loss function. In accordance with our ablation study, we provide qualitative and quantitative analyses in Fig. 2 and Tab. 2. For our quantitative analysis, we report results using the "Beard" semantic, and report our metrics on 100 input-edit pairs generated by Stable Diffusion. In our experiments, we use LPIPS [40], DINO [22] and DreamSim [7] metrics to assess the faithfulness to the input image and use CLIP-T [26] and SigLIP-T [38] text-to-image similarity metrics to assess the faithfulness to the target semantic.

Ablation on sample size. We perform ablations to assess the impact of the number of images sampled from the GAN directions during the learning phase (see Fig. 2 (b)). Our findings reveal that `GAN2Diff` can successfully learn directions with as few as $N = 10$ samples. Moreover, increasing the sample size appears to yield slightly more disentangled results, as we also verify with our quantitative analyses presented in Tab. 2.

Ablation of loss terms. We perform ablations focusing on the individual loss terms (see Fig. 2 (c)). Employing only the semantic alignment loss (indicated as 'w/o $\mathcal{L}_{dir}$') demonstrates the capability to learn the desired edit. However, this approach results in entangled outcomes that alter facial structure. However, utilizing both loss terms in conjunction (indicated as 'with $\mathcal{L}_{dir}$') leads to highly disentangled edits, ensuring that the edits are consistent with the facial structure and background. We refer to our quantitative ablations to assess the impact of $\mathcal{L}_{dir}$ on the disentanglement of the edits. As shown in Tab. 2, absence of this loss term results in deteriorated disentanglement properties (as validated by LPIPS and DINO metrics).

D Qualitative Comparisons with Diffusion-based Editing Methods

To qualitatively compare the edits performed with diffusion-based editing methods and the edits performed by `GAN2Diff`, we provide additional results here. In addition, we further demonstrate the editing capabilities of our framework, such as edit strength interpolation and generalization over out-of-domain images (e.g. images with style components) to demonstrate how the edit strength can be controlled and applied to a diverse set of imaging settings.

To compare the edits performed by `GAN2Diff` with diffusion-based editing methods, we benchmark our approach against several recent approaches including Concept Sliders [9], SEGA [1],

Method	LPIPS↓	CLIP-T↑	DINO↑	SigLIP-T↑	DreamSim↑
$N = 10$	0.098±0.038	0.403±0.033	0.869±0.092	0.143±0.020	**0.872±0.066**
$N = 100$	0.136±0.051	**0.425±0.019**	0.842±0.119	**0.148±0.011**	0.771±0.085
w/o $\mathcal{L}_{dir}$	0.121±0.046	0.409±0.020	0.832±0.117	0.143±0.013	0.860±0.080
Ours	**0.093±0.034**	0.406±0.021	**0.891±0.090**	0.147±0.013	0.861±0.065

Table 2: **Quantitative Results for the Ablation Studies.** We perform ablation studies quantitatively on the number of samples N used in training, and the loss components included during optimization. We evaluate each variant with LPIPS [40], DINO [22], DreamSim [7], CLIP-T [26] and SigLIP-T [38] metrics. We label the final direction as "Ours", where we set $N = 1000$ and use both $\mathcal{L}_{sem}$ and $\mathcal{L}_{dir}$.

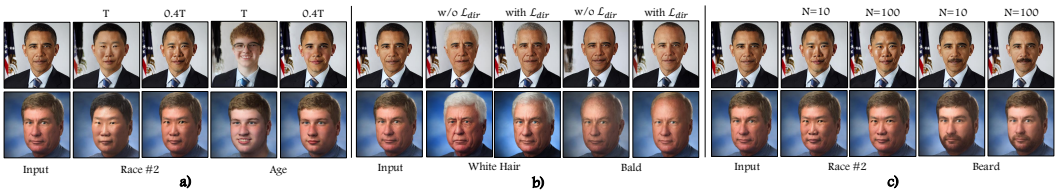

Figure 2: **Ablation Study.** We perform ablations to assess the effectiveness of three different variables, which are the selection of editing timesteps (a), effect of proposed loss terms (b) and the number of sample used for training (c). For each of our ablations we present qualitative results on two different edits, where their corresponding labels are provided for easy understanding.

InstructPix2Pix [3] and Prompt2Prompt (P2P) [11]. Here we use Null-text inversion (NTI) to adapt Prompt2Prompt for real image editing task (NTI + P2P). In particular, SEGA, Prompt2Prompt, and InstructPix2Pix often result in substantial alterations to the input image for edits like 'Asian'. In addition, Prompt2Prompt and SEGA lead to entangled edits for edits such as 'Gender', where alterations on attributes such as age and race are also visible, whereas InstructPix2Pix produces results that are deteriorated in terms of image quality. As illustrated in Fig. 13, `GAN2Diff` surpasses these alternatives in maintaining semantic accuracy and in its ability to execute disentangled edits. Concept Sliders face challenges in applying multiple edits simultaneously, such as combining Race and Beard modifications, resulting in significant deviations from the original input image, whereas `GAN2Diff` can successfully combine multiple edits. We provide additional examples of combining multiple edits in Fig. 7.

In addition, we provide additional comparisons with LEDITS++ [2], which requires text prompts to perform edits within the Stable Diffusion model. We perform these comparisons on Beard, Gender, and Race semantics, which we provide in Figs. 3, 4 and 5. As can be seen from the results, our method performs more disentangled edits compared to LEDITS++. Furthermore, we provide comparisons with the DDPM-based direction discovery methods Asyrp [18] and DiffAE [25] in Fig. 14, along with the diffusion-based editing methods PnP-Diffusion [33] and MasaCtrl [4].

Lastly, since our method performs real image editing with DDPM inversion [15], we provide ablations on the impact of the inversion method in Fig. 6, where we compare with DDIM inversion [31]. As demonstrated qualitatively, our method is able to perform edits with both inversion methods, but can preserve nuanced details with DDPM inversion better, compared to DDIM inversion.

E Qualitative Comparison with Diffusion-based Latent Direction Methods.

We conduct comparisons with methods that identify latent directions within Stable Diffusion. Diffusion Pullback [23] introduces an unsupervised approach to discover latent directions in diffusion-based models, employing the pullback metric for this purpose. Another recent method, NoiseCLR [6], employs a contrastive learning framework to uncover directions without supervision. Given the unsupervised nature of both methods, we selected three overlapping directions for comparison: 'Gender', 'Old', and 'Race'. Fig. 9 shows a comparative analysis between our approach, Diffusion Pullback (referred to as D-Pullback), and NoiseCLR. The comparison reveals that Diffusion Pullback often results in significant alterations to the input image (such as race edit), as acknowledged in their study

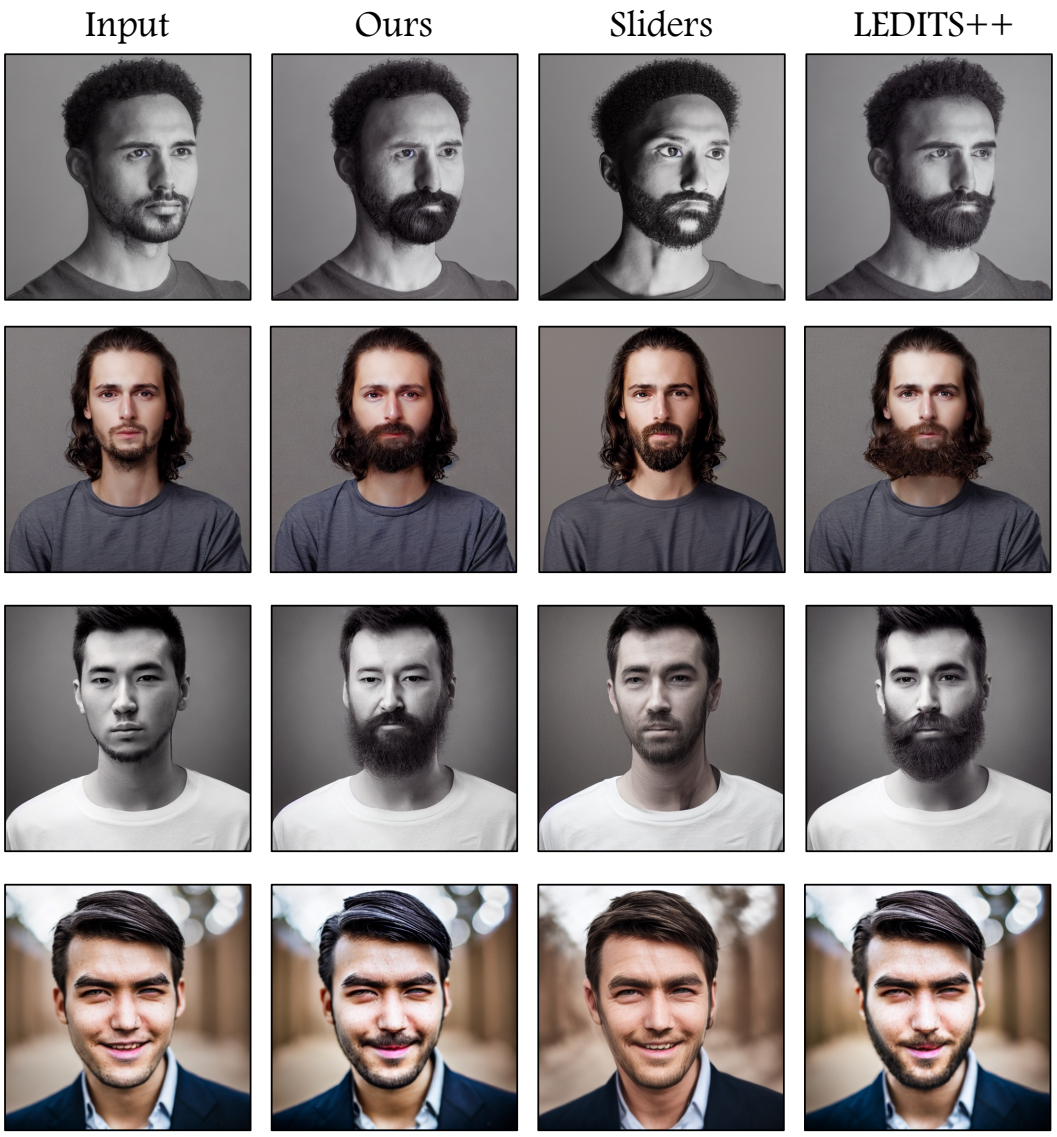

Figure 3: **Comparisons on Beard Edit.** We provide qualitative comparisons on beard editing task with Concept Sliders [9] and LEDITS++ [2] where we use image sliders in [9] for a fair comparison. Our editing results succeeds over the competing approaches both in terms of editing quality and content preservation. Note that [9], which learns the edit based on reference images, struggle when it attempts to add a beard to sample without any traces of the attribute.

[23]. Although NoiseCLR shows performance on par with our method, its unsupervised approach inherently restricts it to a limited number of discoverable directions.

F Qualitative Comparison with GAN-based Latent Direction Methods.

GAN-based editing techniques are recognized for their exceptional editing abilities, attributed to their disentangled latent spaces [36]. In our comparative analysis, we evaluate `GAN2Diff` alongside leading GAN-based methods capable of identifying directions within the latent space, including StyleCLIP [24] that finds directions using text prompts, and unsupervised methods LatentCLR [37], GANSpace [10] and SeFa [29] (see Fig. 8). The comparison reveals that our diffusion-based approach delivers results that are on par with those of GAN-based models.

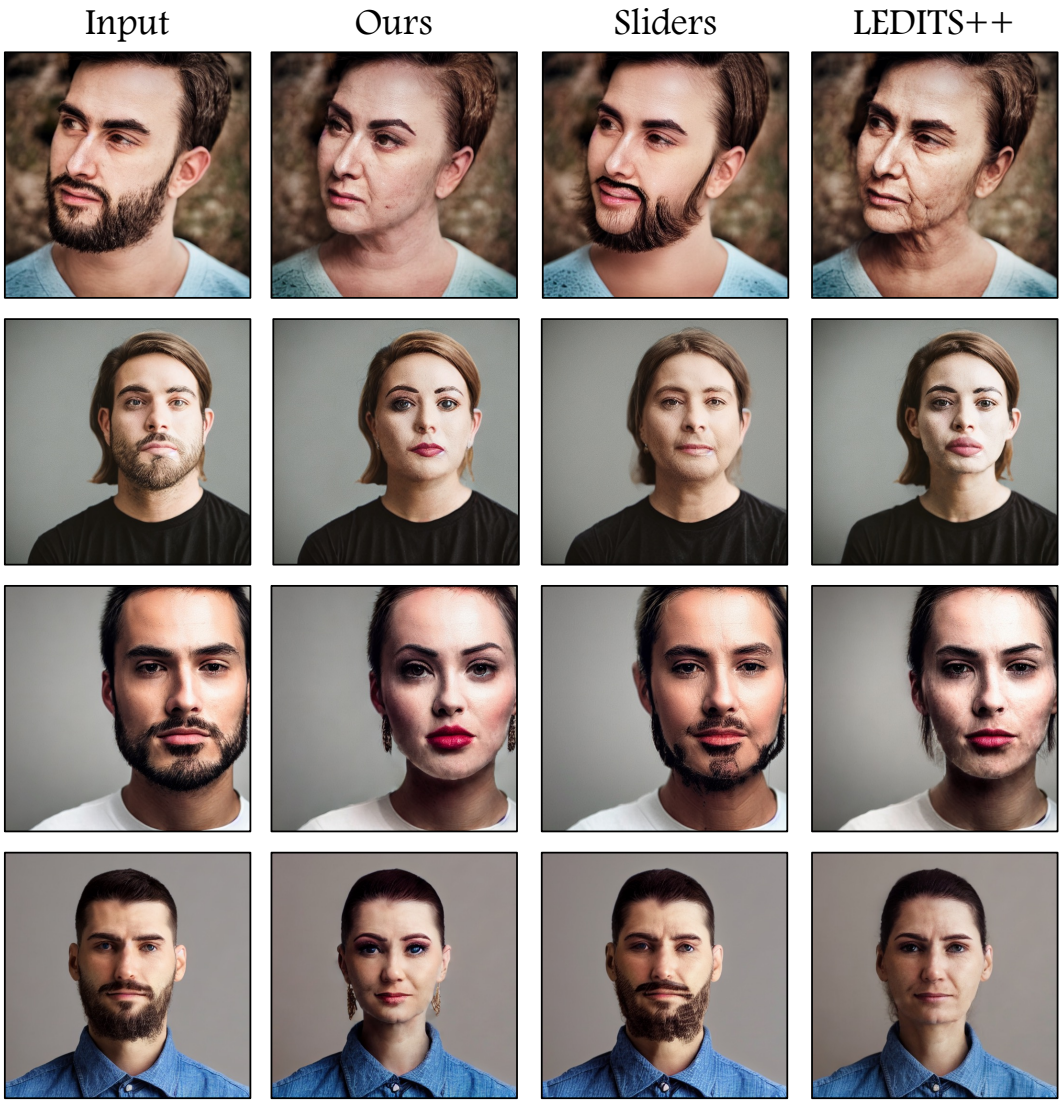

Figure 4: **Comparisons on Gender Edit.** To demonstrate the effectiveness of the gender editing direction learned by `GAN2Diff`, we provide qualitative comparisons with Concept Sliders [9] and LEDITS++ [2]. Notably, both [9] and [2] struggle with artifacts while performing such an edit that changes the overall appearance of the face, where [9] experiences it more severely. Directions learned by our method can both perform such edits without sacrificing from generation quality and in a disentangled manner.

Transferring Directions from Different Image Editing Methods. In addition to the style space of StyleGAN, `GAN2Diff` can also transfer latent directions from input-edit pairs obtained by different methods. To demonstrate this generalizability of our framework, we demonstrate qualitative results on transferring directions from image pairs generated by StyleGAN-NADA [8] and Prompt2Prompt [11], in Figure 10. `GAN2Diff` can both identify semantic changes generated by different image translation methods and can apply them in a way that is not limited to the output space of the model from which the direction is transferred (e.g., StyleGAN-NADA is specialized for face images, where `GAN2Diff` can also apply the edits learned on full-body images).

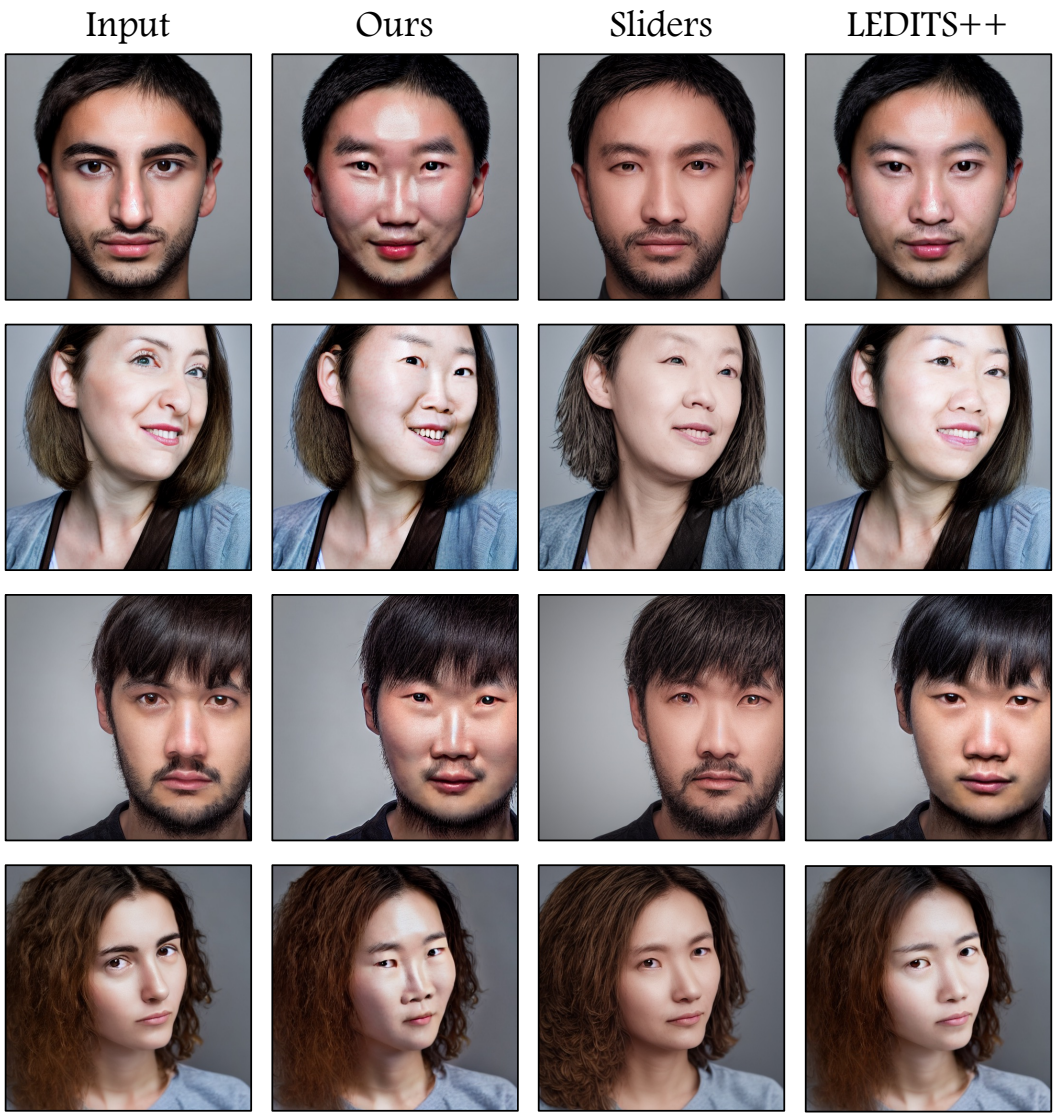

Figure 5: **Comparisons on Race #2 Edit.** We provide qualitative comparisons on the attribute Race #2 with Concept Sliders [9] and LEDITS++ [2]. As observed from the provided examples, our method successfully reflects the edit while preserving the identity of the input image. Note that with LoRA based approaches such as [9], image quality is sacrificed in order to apply the edit where significant changes to the input are present in the corresponding edits.

G Supplementary Editing Results and Details

In addition to the results we provided in the main paper, we provide additional editing results in the supplementary material. Specifically, editing with multiple directions in Fig. 7, editing results with images generated by Stable Diffusion in 16 and editing results on artistic paintings in Fig. 17.

H Failure Cases

Extending our discussion on the Limitations of `GAN2Diff`, we provide examples of failure cases of our method in Fig. 6. In case of edits performed by StyleGAN that involve very nuanced changes (e.g. minor modifications on the nose shape), our method fails to conduct the desired edits. As we demonstrate qualitatively, even though `GAN2Diff` fails to perform the exact edit, it can still recognize

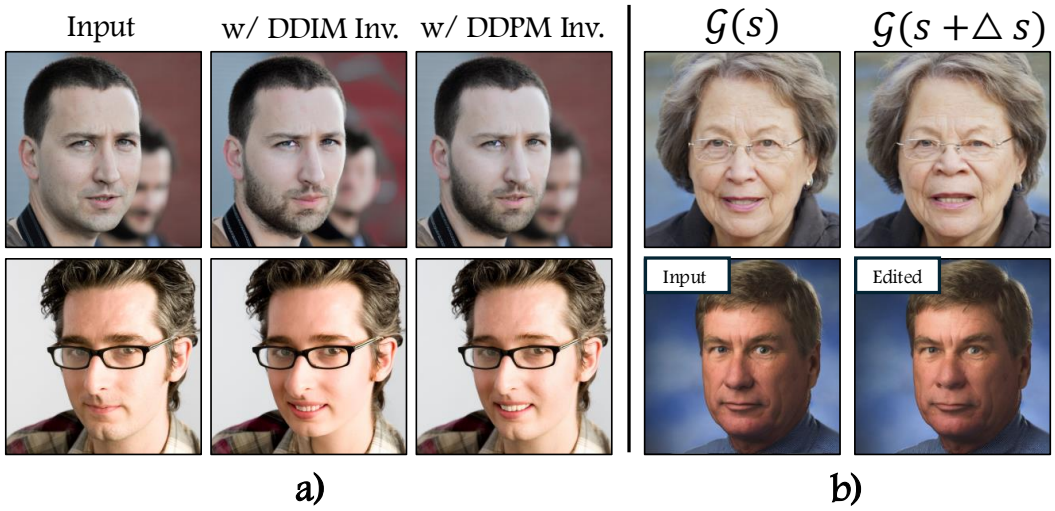

Figure 6: **Ablations on Inversion Method and Failure Cases.** We provide ablations on the inversion method used and the failure cases of the `GAN2Diff` framework. Regarding the inversion method, we observe that DDPM inversion results in more disentangled edits. However, note that our method is able to perform edits on both inversion methods. For the failure cases, we demonstrate an example direction that involves low amount of semantic difference between the input and edited image, where the direction d fails to capture such differences.

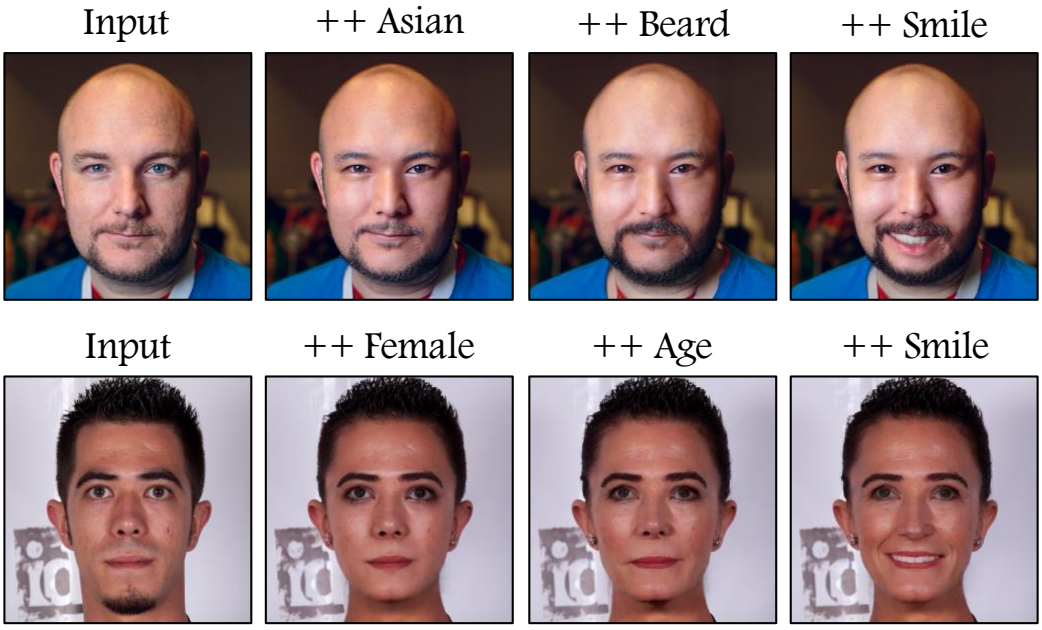

Figure 7: **Application of Multiple Editing Directions.** We provide qualitative results on the application of multiple editing directions over a given input image. From top to bottom, we apply the edit triplets ("race", "beard" "smile") and ("gender", "age" "smile") to each input. Note that we apply each of these edits in a cumulative manner, from left to right.

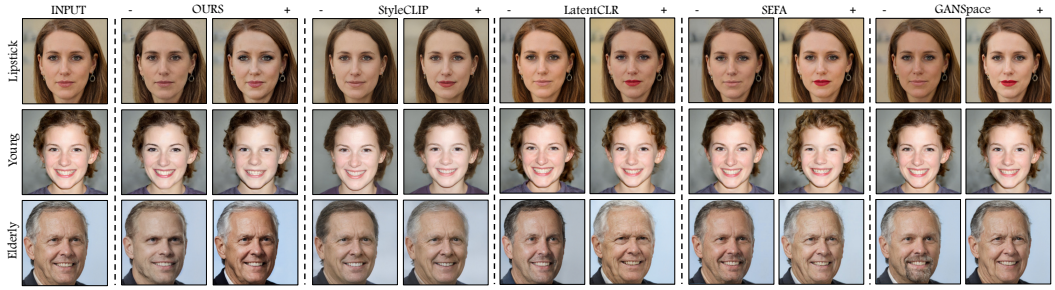

Figure 8: **Qualitative Comparisons with GAN-based Latent Direction Discovery.** `GAN2Diff` is evaluated against methods for discovering latent directions in GANs. Our findings show that the editing and direction discovery capabilities of `GAN2Diff` are on par with those of GAN-based approaches, especially in the context of detailed face editing.

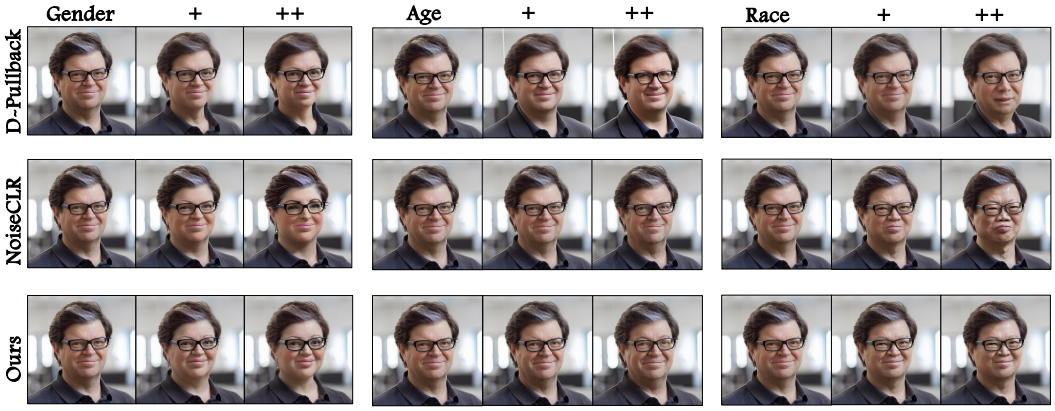

Figure 9: **Qualitative Comparison with Diffusion-based Latent Direction Methods** We compare our method with Diffusion Pullback [23] and NoiseCLR [6].

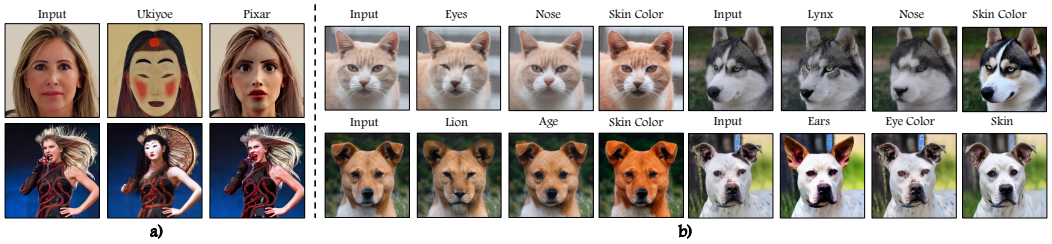

Figure 10: **Edits Transferred from additional methods & domains.** We demonstrate the generalizability of `GAN2Diff` by transferring edits using input-edit pairs generated by from StyleGAN-NADA [8], and Prompt2Prompt [11] (a). In addition, we also demonstrate the generalizability of our method to domains different than faces, such as cats and dogs (b).

the region to be edited (e.g., the shape of the nose). We suggest the cause of this behavior the capabilities of the text encoder and the amount of detail that can be encoded by the diffusion model (Stable Diffusion for our case).

I Comparisons with W+ Adapter and Concept Sliders

We provide a qualitative comparison between the $\mathcal{W}+$ Adapter and our proposed method in Fig. 12. Our approach effectively learns latent directions from StyleGAN, such as Gender or Age, and applies these directions to any given image, as demonstrated in Fig. 12. In contrast, the $\mathcal{W}+$ Adapter requires

Figure 11: **Qualitative Results from GAN2Diff.** GAN2Diff can successfully transfer edits from GANs to diffusion models, and facilitate them over edits in different imaging setups

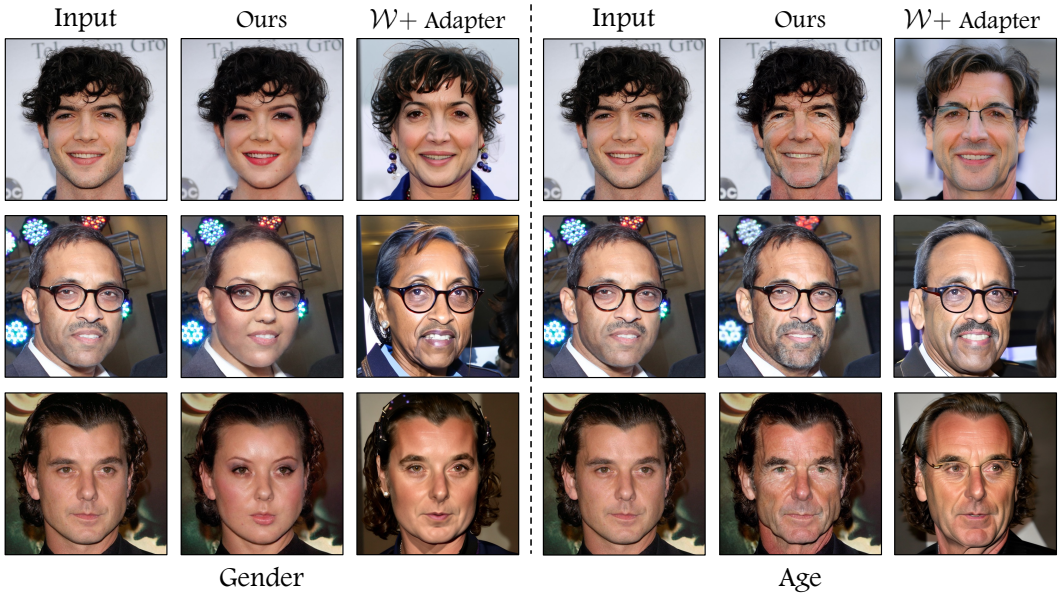

Figure 12: **Comparisons with $\mathcal{W}+$ Adapter [19].** We compare our method with [19] as a competing approach on face editing task. Apparent from the results we provide, our method succeeds over in terms of capabilities such as content preservation (preserving the identity and details irrelevant to the edit) and disentangled editing (such as disentangling attributes like eyeglasses and age).

fine-tuning the Stable Diffusion model by training a separate $\mathcal{W}+$ Adapter. Our results demonstrate that our edits more accurately preserve the integrity of the input image while implementing the desired edits, such as changes in Gender or Age.

Moreover, we provide qualitative comparisons with Concept Sliders [9] in Fig. 3, Fig. 4 and Fig. 5 where our method surpasses [9] both in terms of disentanglement capabilities and representation quality without the need of training separate LoRA models for each direction.

Figure 13: **Qualitative Comparison with Diffusion-based Image Editing Methods** We compare our approach with Concept Sliders [9], SEGA [1], InstructPix2Pix [3] and Prompt2Prompt (NTI + P2P) [11]. The qualitative outcomes demonstrate that `GAN2Diff` outperforms the aforementioned methods in achieving disentangled image edits both in single and multiple semantics (e.g. "Race" and "Beard"), and in identifying detailed latent directions.

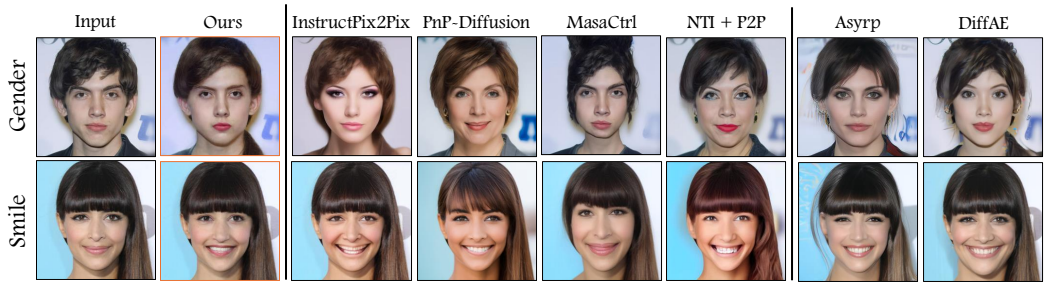

Figure 14: **Supplementary Comparisons with Diffusion-based Editing Methods.** In addition to the editing methods included in our evaluation benchmark such as InstructPix2Pix [3] and Prompt2Prompt [11] (NTI + P2P), we present qualitative results with PnP-Diffusion [33], MasaCtrl [4], Asyrp [18], and DiffAE [25]. As we demonstrate qualitatively, `GAN2Diff` achieves more disentangled edits compared to competing methods, while successfully reflecting the edited semantic.

J StyleGAN Edits vs. Transferred Edits

To demonstrate how the representations that `GAN2Diff` learn align with the latent space of StyleGAN, we demonstrate the edits performed by the directions learned by our framework and the input-edit pairs sampled from StyleGAN in Fig. 15.

K Training Algorithm

To further clarify our training procedure for learning a latent direction d, we provide the training algorithm of `GAN2Diff` in Alg. 1. We also provide our training code as a part of our supplementary material.

L Rescoring Analysis

To assess the disentanglement capabilities of the learned directions, we perform a rescoring analysis to assess how the CLIP classification probabilities for specific attributes change following an edit,

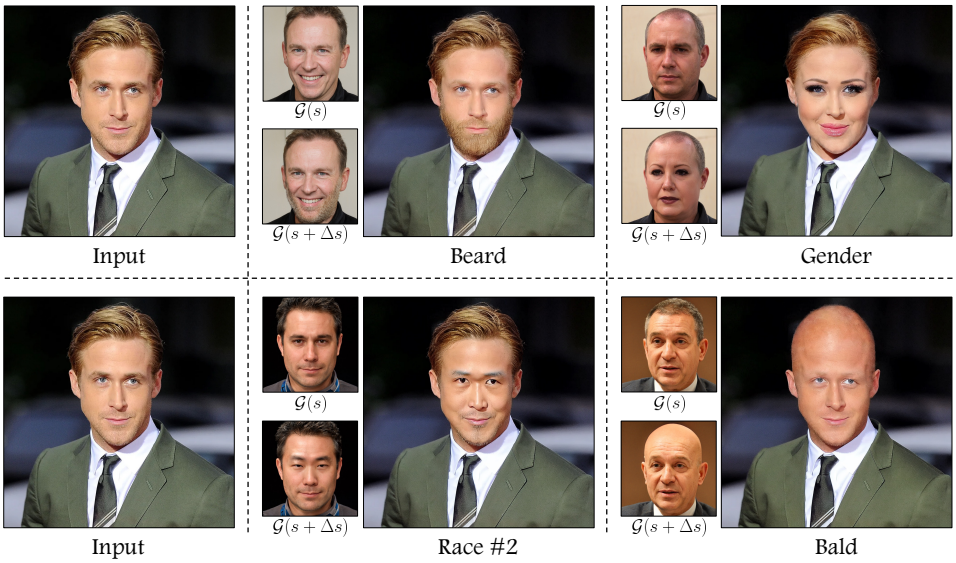

Figure 15: **Edits learned by `GAN2Diff`.** We demonstrate the beard, gender, race # 2 and baldness edits above, along with reference images from the training datasets constructed using StyleGAN. Above, we show the images generated by StyleGAN as $\mathcal{G}(s)$ and their edited counter-parts as $\mathcal{G}(s+\Delta s)$, respectively. As our qualitative results also show, edits learned by `GAN2Diff` successfully translates the disentangled directions performed on StyleGAN to Stable Diffusion.

following [28, 6]. The rows in Tab. 3 correspond to various editing directions—Asian, Smile, Gender, and Beard—applied to 100 images generated by Stable Diffusion, while the columns show the resulting shifts in CLIP scores. Consistent with our expectations, performing a targeted edit increases the probability of the image being classified under that attribute. For example, an Asian edit improves the likelihood that the image is identified as Asian by 53.6%, with similar increases observed for other attributes, as detailed in the diagonal entries of Tab. 3. Additionally, some edits naturally impact related attributes; for instance, enhancing the Gender attribute towards femininity notably reduces the Beard attribute probability. The interaction between Beard and Smile attributes also demonstrates a degree of interdependence, which can be attributed to inherent biases within the SD model, where adding beard diminishes the presence of smile attribute. Our approach notably supports disentangled editing as it minimally affects the scores of unrelated attributes when making specific edits.

M User Study

We conducted a user study with 40 participants on Prolific.com, compared to [1, 3, 11, 9]. Participants were shown a series of edits made using common semantics for each method in comparison. Unlike the previous experiments, we conduct this set of experiments with real images to also evaluate the real image editing capabilities of our method. They were then asked to judge whether they deemed the edit successful in conveying the intended semantics and if the edit was executed in a disentangled manner, for a set of 60 input-edit pair. The participants rated each question on a scale from 1 to 5, with 5 representing the highest level of satisfaction, where we provide additional details on our study in the Supplementary Material. We present the result for the user study in Tab. 1, where our method outperforms competing approaches in terms of user preference.

Algorithm 1: Learning direction d with `GAN2Diff`

Input: Pre-trained diffusion model $\epsilon_\theta(x_t, c)$; CLIP image encoder $E_I(x)$; latent codes $\{s_1, \ldots, s_N\}$; editing direction Δs; StyleGAN generator $\mathcal{G}(s)$; random initial embedding d; learning rate λ

Output: Trained conditional embedding d

Procedure: LEARNDIRECTION$(\epsilon_\theta, E_I, \mathcal{G}, \{s_i\}, \Delta s, d, \lambda)$

begin

  while training **do**

    for $i \leftarrow 1$ **to** N **do**

      $x_i \leftarrow \mathcal{G}(s_i)$;

      $x_i' \leftarrow \mathcal{G}(s_i + \Delta s)$

      sample $\epsilon \sim \mathcal{N}(0, 1)$;

      sample $t \sim \mathcal{U}(1, T)$ and set $\epsilon^t \leftarrow \alpha^t \epsilon$

      $x_{i,t} \leftarrow x_i + \epsilon^t$;

      $x'_{i,t} \leftarrow x_i + \epsilon^t$

      $\epsilon_{\text{input}} \leftarrow \epsilon_\theta(x_{i,t}, d)$;

      $\epsilon_{\text{edited}} \leftarrow \epsilon_\theta(x'_{i,t}, d)$;

      $\mathcal{L}_{\text{latent}} \leftarrow -\left\|\epsilon_{\text{edited}} - \epsilon_{\text{input}}\right\|_2^2$

      $\mathcal{L}_{\text{sem}} \leftarrow 1 - \text{cossim}\big(E_I(x_i'), d\big) + \text{cossim}\big(E_I(x_i), d\big)$

      $\mathcal{L} \leftarrow \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{latent}}$

      $d \leftarrow d - \lambda \nabla_d \mathcal{L}$

    end

  end

  return d

end

	Asian	Smile	Gender	Beard
Asian	**53.6**	15.8	-12.1	-3.5
Smile	-20.4	**41.2**	11.8	-7.2
Gender	2.0	-4.3	**94.7**	-19.8
Beard	-2.6	-8.1	-0.05	**28.3**

Table 3: **Re-scoring Analysis.** `GAN2Diff` can perform edits efficiently on several attributes. The attributes edited are shown as rows, whereas the measured attributes are shown as columns.

Method	User Pref. ↑
SEGA	1.96
Prompt2Prompt	2.74
InstructPix2Pix	2.06
Concept Sliders	3.25
Ours	**3.36**

Table 4: **User Study Results.** Human preference scores on attribute disentanglement and edit quality.

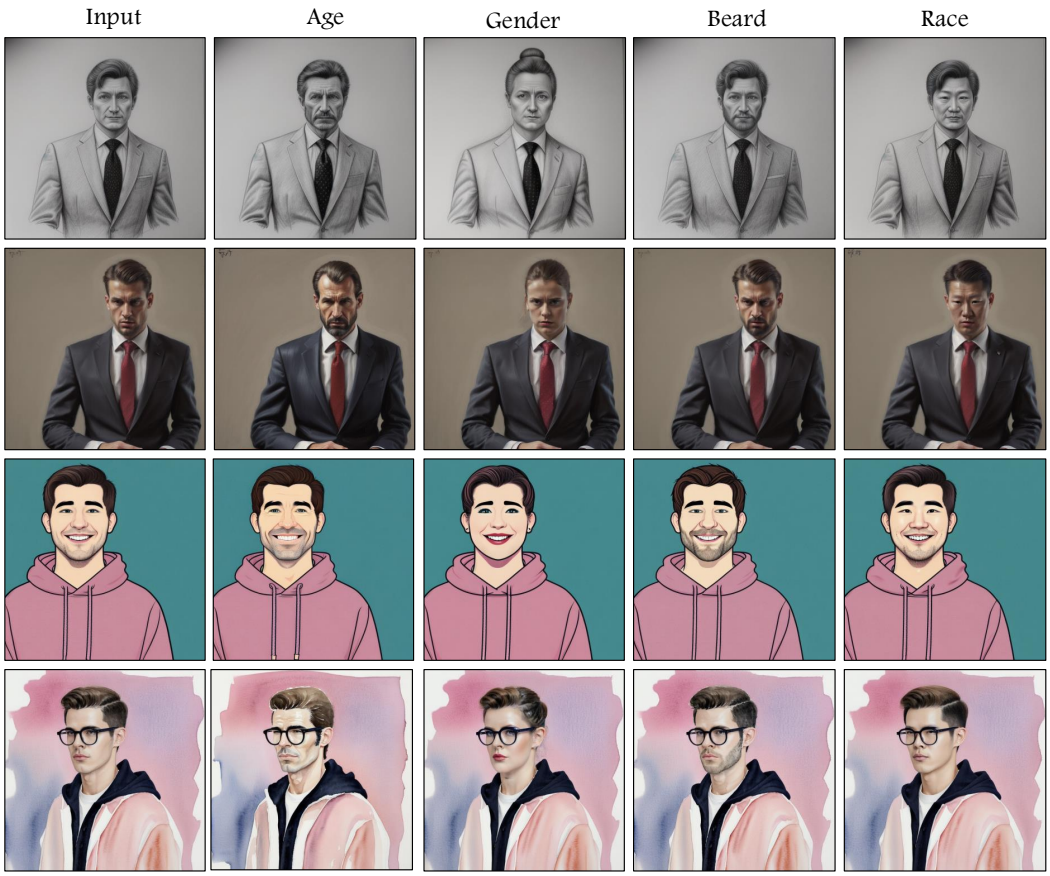

Figure 16: **Supplementary Editing Results on Images Generated by Stable Diffusion.** To show the generalization capabilities of the directions learned by `GAN2Diff`, we present qualitative results on images generated with Stable Diffusion, containing diverse style elements. Our directions can apply the desired semantic to the generated images, without the need of inversion. Additionally, our method can preserve the stylization characteristics of the edited images.

Figure 17: **Supplementary Editing Results on Artistic Paintings.** We provide additional editing results painting images, to further demonstrate our methods compatibility with stylized images. As we demonstrate qualitatively, our method is able to perform edits without changing the overall style of the image (e.g. oil painting).

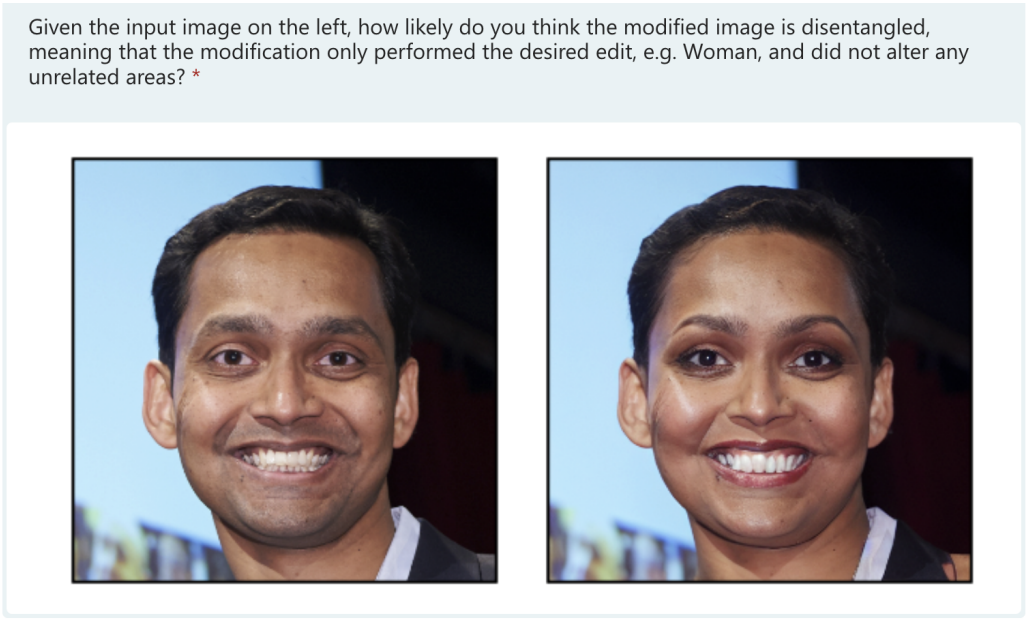

Figure 18: **User Study Setup.** To further clarify the experimental setup used for the user study, we provide an example question. In the conducted study, users are shown an input-edit pair for a corresponding method, and asked to assign a score from 1-to-5, based on their assessment of the edit.