Supplementary Material (SM) - Attribute Diffusion: Diffusion Driven Diverse Attribute Editing

Rishubh Parihar*,1 Prasanna Balaji*,1 Raghav Magazine[2] Sarthak Vora[3]
Varun Jampani[4] R. Venkatesh Babu[1]

[1]Indian Institute of Science, Bangalore [2]IIT Dharward [3]UCLA [4]Stability AI

A. Organization of supplementary

This document presents additional comparisons and results. First, we compare with text-based diverse editing using a text-to-image editing method built on the Stable Diffusion model [12] in Sec. B. Next, we provide details for dataset creation in Sec. C and implementation details in Sec. E and F. We provide results analysis of adding the same direction for multiple images in Sec. G and additional comparisons with existing attribute editing methods in Sec. H. Finally, we present additional results for diverse editing, coarse-to-fine editing, and in 3D face attribute variations. Please check the attached website (

B. Comparison with text-to-image editing [10]

We compare our proposed method for diverse attribute edits with a recent text-to-image editing method [10]. The approach involves generating localized object shape variations using a pre-trained text-to-image Stable Diffusion [12] model. Specifically, the prompt-mixing technique allows the exploration of plausible variations for a given object shape that switches text prompts at different stages of denoising. To perform localized variations, they segment out the object using internal self-attention and cross-attention layers.

To compare for attribute variations, we first perform an eyeglass edit on the source image using the single-direction StyleGAN editing method InterFaceGAN [13] and obtain one version of the edit. Next, we invert this edited image in the Stable Diffusion using null-text inversion [8]. We generate diverse eyeglass variations by providing the following eyeglass types: *brownline, cat-eye, double-bridge, mirror, narrow, oval, retro, rimless, round, square*. The results are present in Fig. 1. We can observe that the text-based editing method is ineffective in generating diverse eyeglass variations compared to our proposed approach. Moreover, getting a fine-grained control for the exploration of diverse attribute shapes is not possible with a text-conditioned editing method, suggesting the importance of our proposed *coarse-to-fine* sampling for attribute exploration.

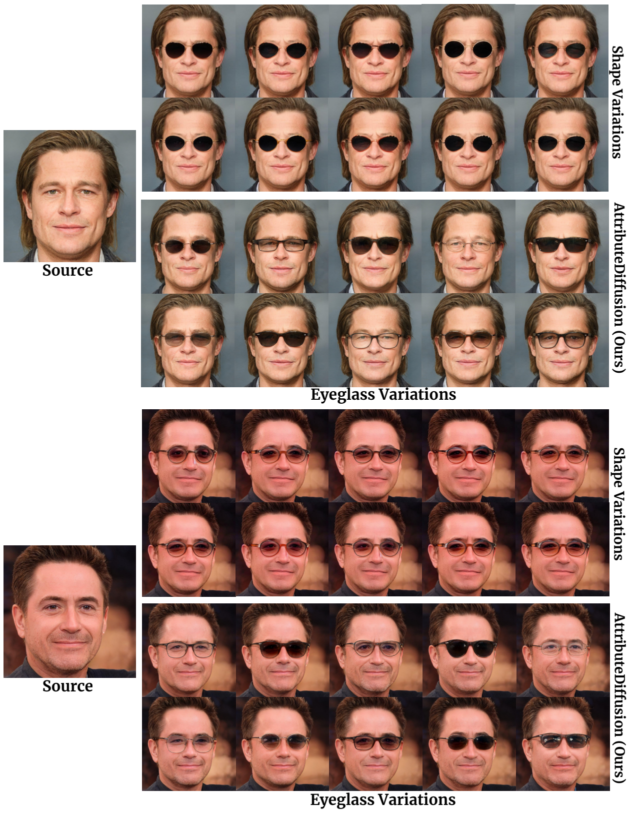

Figure 1. **Comparison with text-to-image editing method** [10] for diverse eyeglass editing that generates localized object shape variations by giving multiple prompts. We first embed a source image and edit it using InterFaceGAN [13] for eyeglasses. Next, we embed the edited image using null-inversion in stable diffusion and generate diverse eyeglass styles with prompt-mixing as proposed in [10]. This approach generates limited eyeglass shape variations, whereas AttributeDiffusion generates a large variety of eyeglasses. Moreover, our *coarse-to-fine* sampling enables us to explore these variations in a principled manner, which is challenging with text-based control.

*Equal contribution

C. Dataset Details

Our method requires a paired dataset of images with and without attribute edits for training. These image pairs can be easily obtained using any existing attribute editing methods, generating a single edit for the source image. To validate the robustness of our proposed method to the method used for generating paired data, we tried various off-the-shelf editing methods for data generation. Specifically, we used GAN latent space-based editing methods StyleCLIP [11] for hairstyles/cars edits, and encoder-decoder based editing methods - SAM [3] for aging, Latent-composition [4] for smile and eyeglasses and Swapping autoencoders [9] for church style editing. Our method performs good quality diverse edits when trained on the paired dataset obtained by all the approaches, proving the generalizability of our approach. Next, we will provide details of these methods.

StyleCLIP [11]. We used StyleCLIP to generate an edited paired dataset for hairstyle and car edits. Specifically, we trained StyleCLIP [11] mapper with text prompts “bangs hairstyle”, “mohawk hairstyle”, “curly hairs”, “afro hairstyle”, “buzz cut”, and “bob cut” for hairstyles. Post-training, we use the trained StyleCLIP models to perform hairstyle edits on a subset of CelebA-HQ [6]. Examples of source and edited images are shown in Fig. 2. We create a dataset of $30K$ image pairs (edited and source) for each text prompt. We obtain a combined set for hairstyle variations consisting of $160K$ synthetic image pairs each. For car edits, we trained StyleCLIP models with the text prompts ‘sports car’ and ‘classic car’ and used them to edit $10K$ synthetic images generated by StyleGAN2 trained on cars. The obtained dataset is used to train the DDPM model to learn diverse edit variations.

SAM [3]. For the age attribute, we used SAM [3], a state-of-the-art age editing method, to generate image pairs with age editing. SAM is trained with an additional encoder model on StyleGAN2, conditioned on a given target age, to achieve fine-grained control over the age edit. The encoder is guided by the age regression and reconstruction losses to obtain precise age editing. To obtain a paired dataset, we use age 60 as the target value and perform edits on $30K$ source images from CelebA-HQ.

Swapping autoencoders [9] is an autoencoder-based method specifically designed for image manipulation. The core idea is to project the input image into two disentangled latent components controlling structure and texture. The latent components of the source images can be swapped with other images’ texture and structure components to obtain the desired edits. We swap the texture code with a randomly sampled image to generate texture variations for churches while preserving the structure of the image. We first generate $30K$ images from StyleGAN2 trained on churches and edit them using swapping encoders.

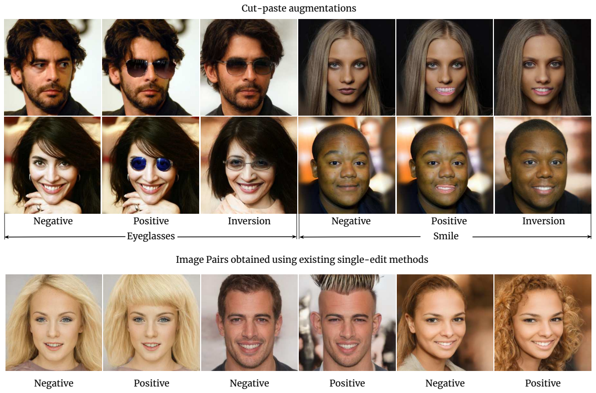

Figure 2. **Examples of synthetic image pairs.** (Top) We present positive and negative image pairs generated by cut-pasting the attribute region leveraging smooth StyleGAN priors [4]. The augmented positive image is passed through the encoder and StyleGAN2 to obtain smooth inversion. (Bottom) Example image pairs generated by single direction editing method StyleCLIP, given the text prompts ”bangs hairstyle”, ”mohawk hairstyle”, and ”curly hairs.” Note that the latent encodings of the negative and positive images are used to obtain the edit directions in the $\mathcal{W}+$ latent space.

Latent composition [4]. For generating images for smile and eyeglass attributes, we used a simple editing method based on cut-pasting the desired attribute region on the source image, leveraging the latent composition properties of GANs [4]. The augmentations generated by cut-pasting can be realistically composed by passing them through a pretrained encoder and StyleGAN2 model (Fig. 2) sequentially. Specifically, we sample a set of ‘positive’ images X_p with attribute a from the CelebA-HQ dataset using the attribute annotations from CelebA. We sample a set of ‘negative’ images X_n that does not have the attribute a. Then we sample an image from $x_n \in X_n$ and $x_p \in X_p$, mask the region of interest/part of the face containing a from x_p, and paste it onto x_n. For example, we used the mouth region for the smile attribute, while for eyeglasses, we used eyeglass regions to perform cut-paste augmentation (Fig. 2). When passed through the e4e encoder and StyleGAN2 generator, these augmented images are blended due to smoothness prior to the StyleGAN2 generator. To obtain part segmentation masks, we use a few-shot segmentation network [15], which uses StyleGAN features to perform segmentation. Specifically, we used five ground truth segmentations from CelebAHQ-Mask [7] to train the few shot segmentation models.

D. Ablation Study

Diversity parameter γ. We performed ablation over the diversity parameter γ, for a set of edit directions in Fig. 3. However, for large γ values, the edits also affect other attributes. As we increase γ, we can observe that the diversity

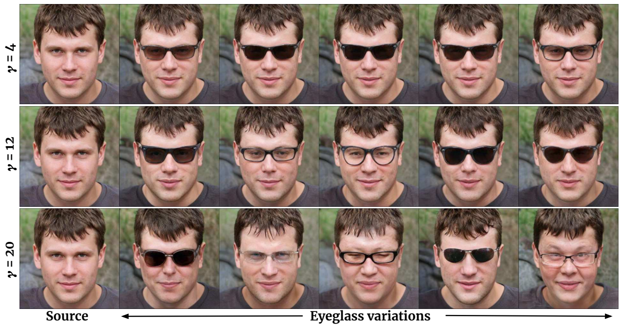

Figure 3. Eyeglass edits for three different values of diversity parameter γ. For a lower value of the parameter ($\gamma = 4$), the variation in eyeglasses is lesser. A higher diversity parameter ($\gamma = 20$) distorts the person's identity. A moderate value of $\gamma = 12$ generates diverse eyeglasses and maintains the face identity.

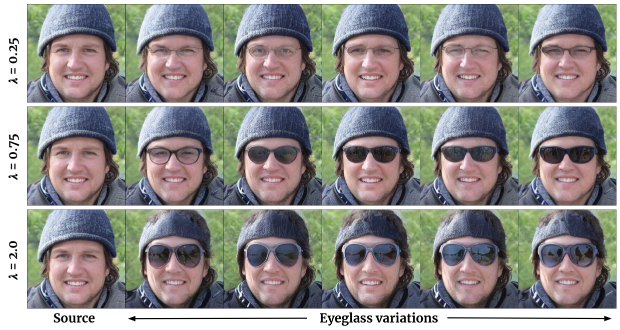

Figure 4. Eyeglass edits for three different strength parameter values λ. For a lower value of $\lambda = 0.25$, the eyeglasses added are transparent with thin frames, while for a higher $\lambda = 2.0$, the frames are thicker with a dark tint, but the identity of the person is not preserved. For $\lambda = 0.75$, the eyeglasses added are prominent, and the person's identity is unaffected.

of eyeglass shapes and lenses increases.

Strength parameter λ. We also performed ablation over the strength parameter λ on a set of edit directions, which is visualized in Fig. 4. It can be observed that the attribute becomes more prominent with an increase in λ until a threshold, beyond which the person's identity starts getting modified.

E. Model architecture

We implemented our denoising network as an MLP network with 10 fully connected layers with 2048 neurons in each layer. Additionally, we added a time conditioning by first encoding the timestamp with 128 dimensional positional encoding. The encoded time embeddings are passed through another MLP network of two fully connected layers with 256 neurons each. Time conditioning is added to all hidden layers of the base MLP network. A skip connection is added after each linear layer, followed by a layer norm layer. Layer norm is not added in the final layer.

F. Computational requirement

We performed all our experiments on a single NVIDIA A5000 GPU. The training time of the DDPM model on a dataset for latent directions is 1 hour on a single GPU for a batch size of 64, although 3-4 such models can be trained on a single A5000 at a time since the runs require less memory.

G. Adding same edit direction on multiple inputs

To analyze the generalization capability of the edit directions for different editing, we perform editing with a single direction on multiple source images. Specifically, we sample a set of edit directions from a trained diffusion model for hairstyles and eyeglasses and perform edits on six source images as shown in Fig. 5. Editing with a direction generates similar styles in all the source images. For example, for hairstyles, d_1 generates curly hair, d_2 generates bangs, and d_3 generates a mohawk hairstyle. Observe that similar frame shapes are generated for each eyeglass direction. For eyeglasses, d_1 generates ellipse-shaped frames, d_3 generates similarly shaped yet thicker frames, and $d2$ generates round-shaped frames.

H. Comparison with single-direction based editing methods

We compare edits generated by our method against three state-of-the-art single-direction based editing methods: InterfaceGAN [13], StyleCLIP [11], StyleFlow [2] and CLIP2StyleGAN [1]. Note that these methods can generate a single edit w.r.t. an attribute for a given image, whereas our method is trained to generate multiple edits w.r.t. an attribute. StyleFlow uses normalizing flows conditioned on the attribute classifier scores to map a source latent code to an edited latent code. CLIP2StyleGAN learns disentangled directions in the CLIP space and transfers them into StyleGAN latent space for attribute editing. We performed editing to generate a single output for eyeglasses, smile, and age attributes using all these methods. We present the comparison results in Fig. 7. For CLIP2StyleGAN, we have generated results for only smile and eyeglasses attributes as CLIP2StyleGAN could not find disentangled edit directions for age. Our method achieves edits with high realism and disentanglement with superior identity preservation ability even with a single-directional edit. StyleFLow also achieves good edits with attribute disentanglement. However, StyleFlow requires additional attribute classifiers to obtain attribute scores, which are required as input to edit any new input image.

Additionally we compare with CLIP2StyleGAN and StyleFLow for generating multiple outputs by changing the strength of the edits in Fig. 6. We can observe that the

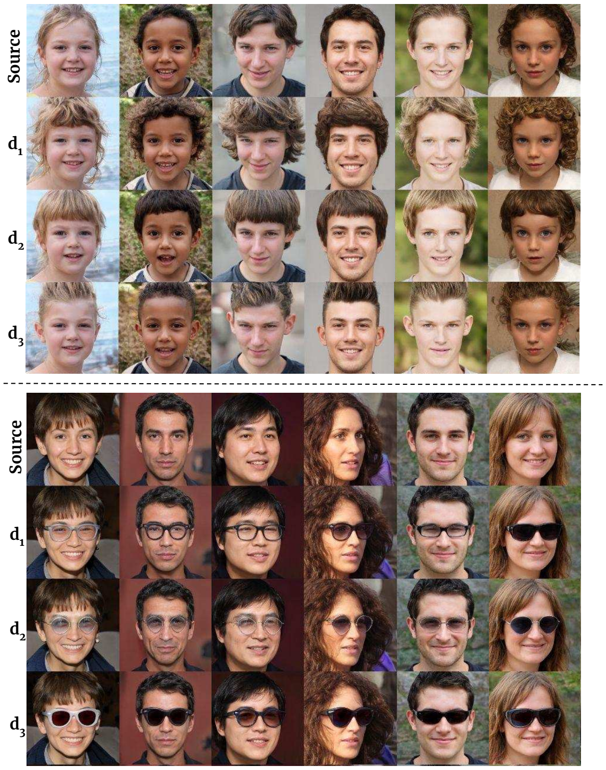

Figure 5. Editing multiple source images with the same direction. Each row, other than the source row, results from editing six different source images using the same edit direction d_i, sampled using the appropriate DDPM model. We present results for hairstyle edits on top and eyeglass edits at the bottom. Observe that for each edit direction, an attribute edit of a similar style is generated for all the source images.

edits generated by both baselines have limited diversity. As we increase the edit strength, the identity of the subject changes in CLIP2StyleGAN. StyleFlow achieves better identity preservation in both edits but requires attribute classifiers during inference. Our method generates highly diverse and realistic attribute edits with superior identity preservation. This explains the need for methods that can model diverse attribute edits.

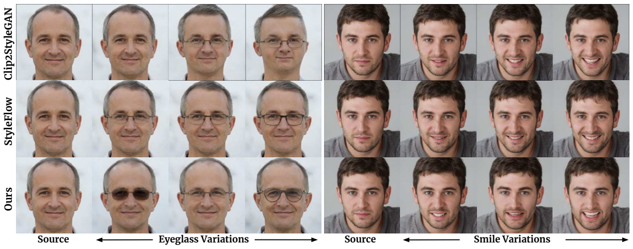

Figure 6. Comparison for diverse attribute edits with single direction-based editing methods - CLIP2StyleGAN [1] and StyleFlow [2]. The variations in CLIP2StyleGAN and StyleFlow are generated by increasing the edit strength.

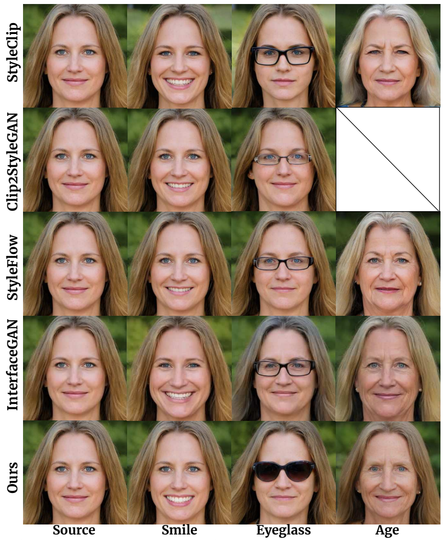

Figure 7. **Qualitative comparison with single direction methods.** StyleCLIP and CLIP2StyleGAN change the identity in age and eyeglass edits, respectively. InterFaceGAN entangles hair color with eyeglass attribute edits. StyleFlow changed the hair color while age editing. The proposed method can generate realistic edits without altering other attributes.

Table 1. **Quantitative comparison with single direction methods.** We compare Cosine Similarity (CS) and Euclidean Distance (ED) between the face embeddings of the source and edited images to measure identity preservation. FID is reported to evaluate the realism of edits. The proposed method outperforms state-of-the-art methods in all the metrics.

	Metric	InterFaceGAN	StyleCLIP	CLIP2StyleGAN	Ours
Eyeglasses	FID	80.49	69.66	73.41	**66.65**
	CS	0.953	0.950	0.944	**0.958**
	ED	0.47	0.49	0.51	**0.41**
Smile	FID	78.59	65.42	72.25	**64.68**
	CS	0.883	0.949	0.920	**0.960**
	ED	0.736	0.439	0.616	**0.428**
Age	FID	99.86	77.83	-	**71.62**
	CS	0.905	0.878	-	**0.921**
	ED	0.58	0.75	-	**0.54**

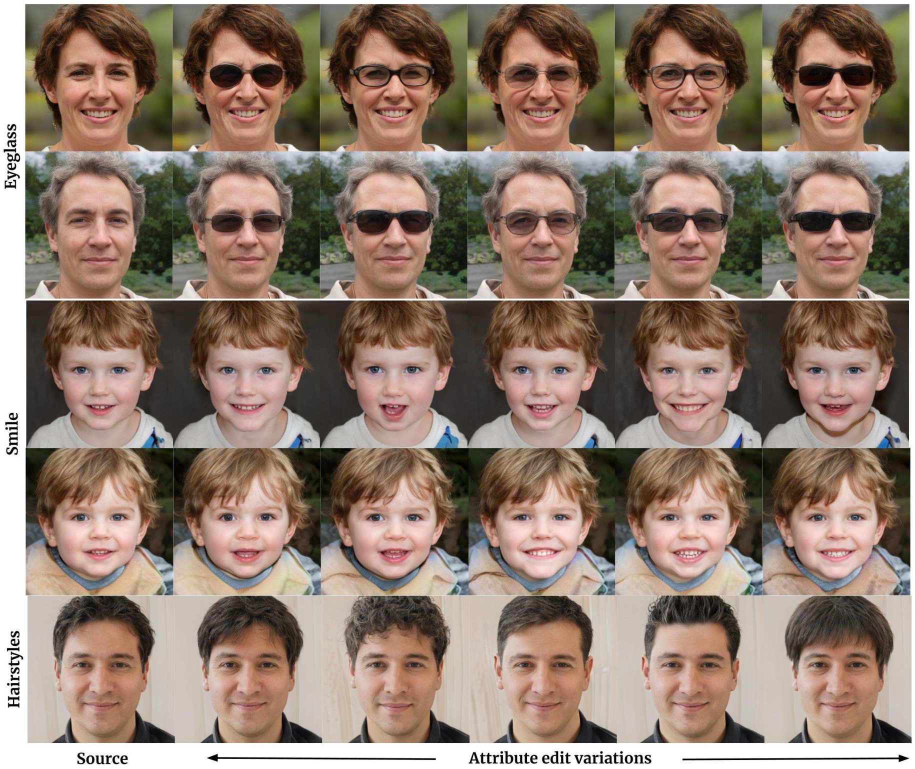

Figure 8. Diverse attribute edits for eyeglasses, smile, and hairstyles.

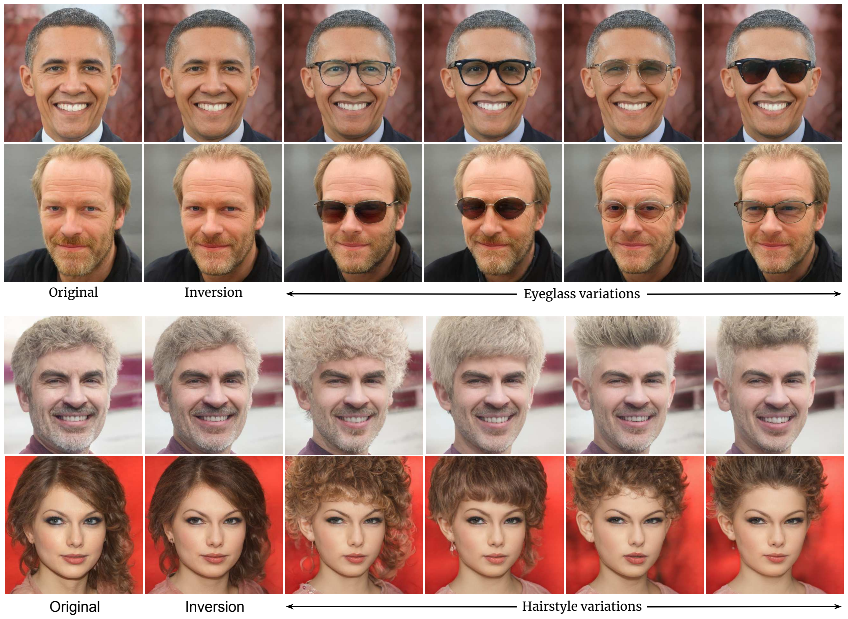

Figure 9. **Diverse attribute edits on real images.** Given a real image, we encode it using e4e [14] encoder and perform diverse attribute editing on the obtained latent code.

I. Additional Results

We present additional results for diverse attribute editing on real and synthetic images in Fig. 9 and Fig. 8, respectively. For real images, we first invert the input image into $\mathcal{W}+$ latent space using e4e [14] encoder model. The proposed method can generate diverse and realistic attribute variations for both real and synthetic datasets. Additionally, we present results for diverse editing on car styles and church styles in Fig. 10.

I.1. Hierarchical sampling for attribute variations

We present results for proposed coarse-to-fine sampling of diverse attribute edits in Fig. 12 for faces and in Fig. 13 for cars and churches. Observe the similarity and nuanced changes within the fine variations and structural diversity in the coarse variations of the outputs.

J. Editing on 3D faces

We present diverse attribute editing results on 3D aware GAN, EG3D [5]. We request the reader to check the accompanying website (*web.html*) to check out the edits from different viewpoints. The proposed method can generate diverse attribute edits that are 3D consistent and preserve the identity. We present results for editing in 3D along with the surface maps of the edited geometry in Fig. 11. We want to highlight that the surface maps are not perfect, as the EG3D model models the geometry in a smaller resolution than the image resolution. Observe that the shape of eyeglasses and hairstyles are changed with diverse edits as visible in the surface maps.

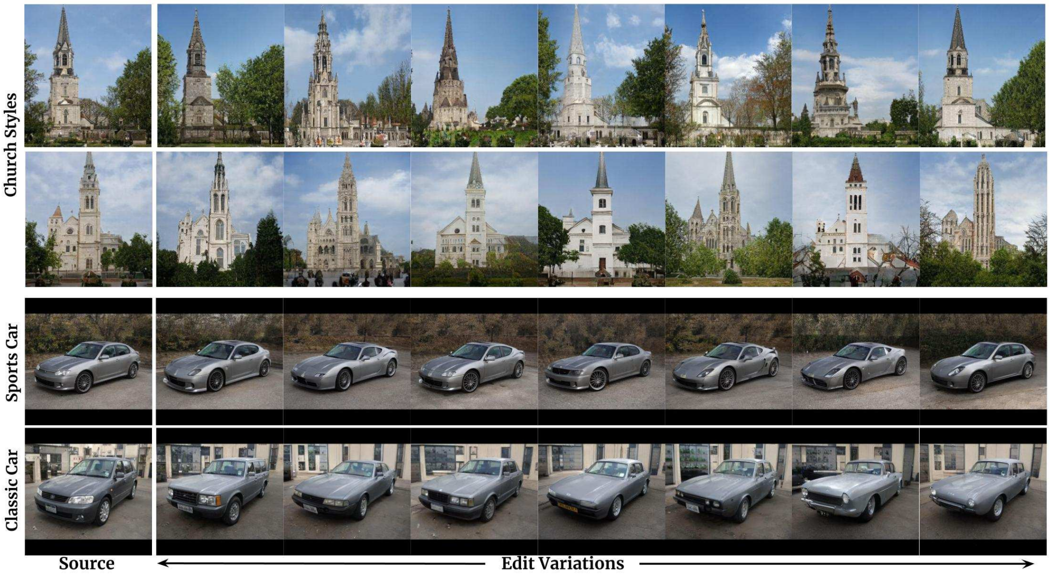

Figure 10. Diverse attribute edits for car styles - sports car and classic car and church styles.

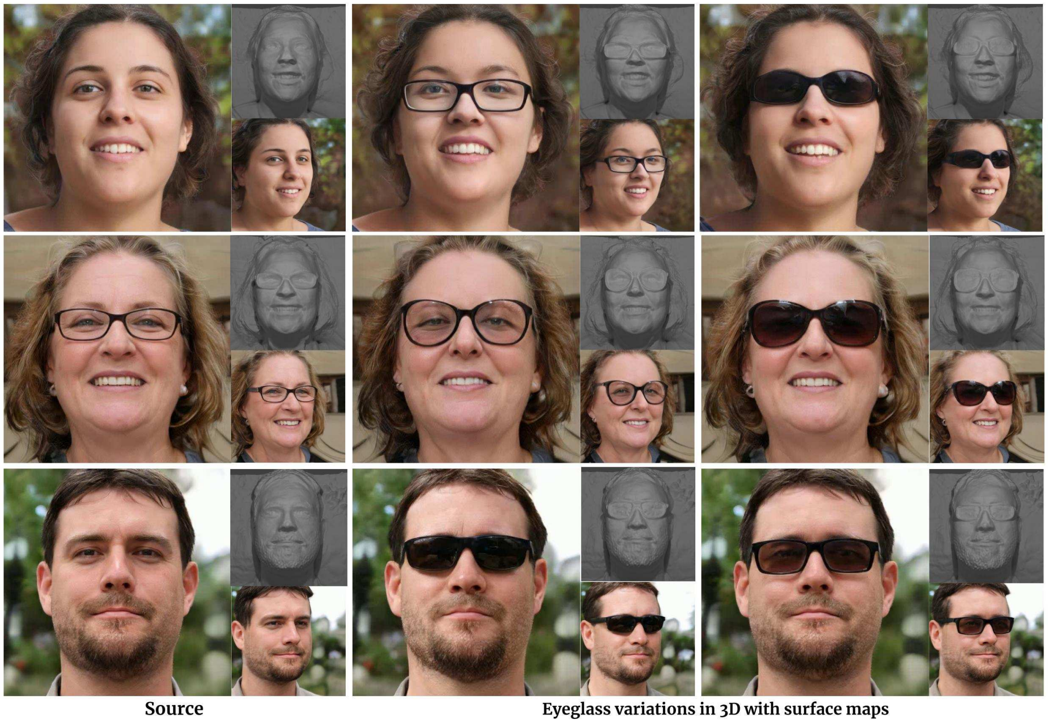

Figure 11. **Diverse eyeglass variations for 3D aware GANs.** Our method can generate 3D consistent, diverse variations for eyeglasses using the latent space of EG3D backbone. We show a side view and the obtained 3D geometry after editing the face in the inset. The edited image has high fidelity and preserves identity and other attributes. Notably, the edits result in a change in the geometry of the face (inset). Interestingly, in the second example, we generate variations for eyeglasses even if the subject is already wearing eyeglasses.

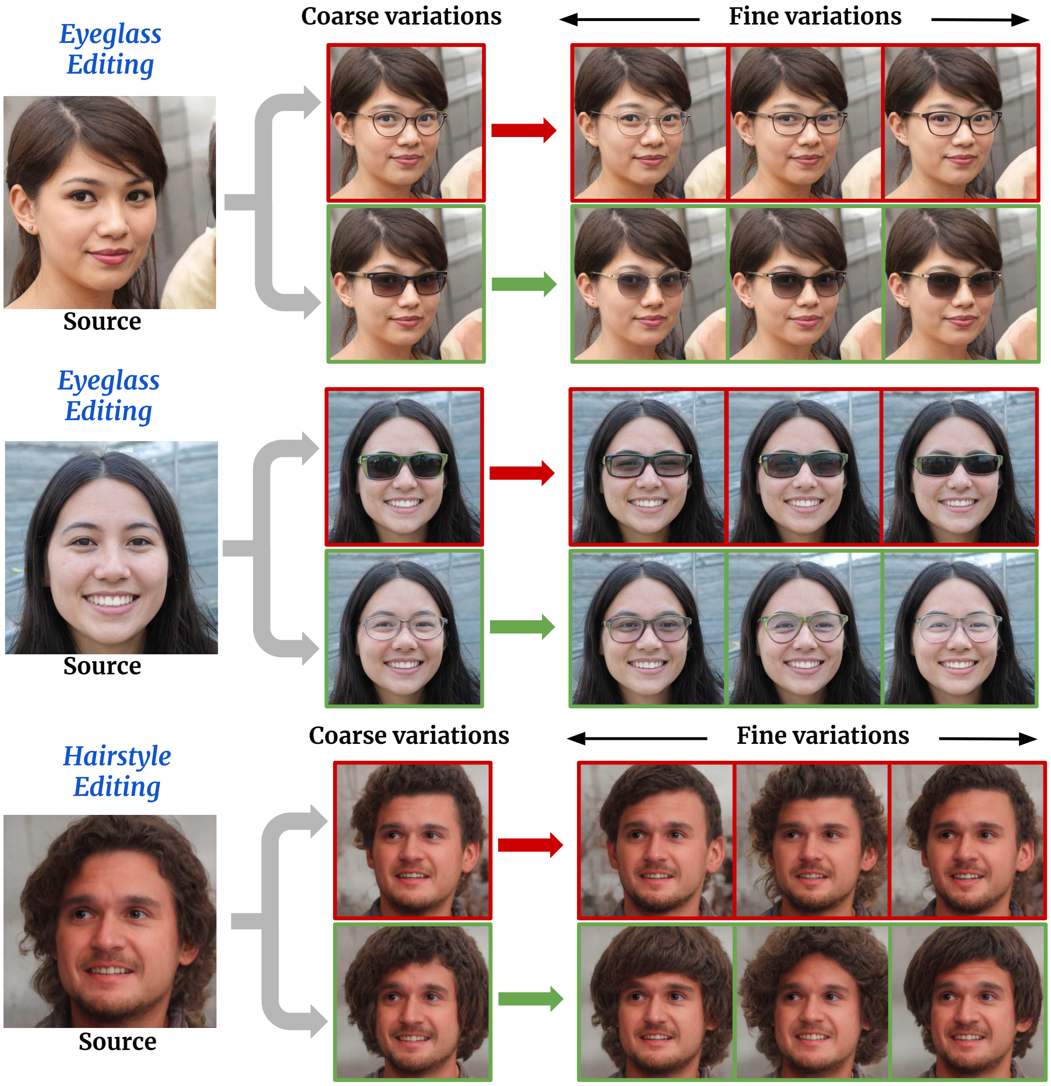

Figure 12. (Zoom in to appreciate fine variations) Hierarchical attribute editing in a coarse-to-fine manner. Given a source image, we first obtain two coarse edit directions for a given attribute (eyeglass/hairstyle). Next, we sample finer variations for a coarse style, preserving other facial attributes and the subject's identity. In the first example, the fine variations generate diverse transparent eyeglasses with variety in the frame shape. The second example shows diverse frame shapes and colors for a given reference coarse variation. In the third example, diverse hairstyles are generated.

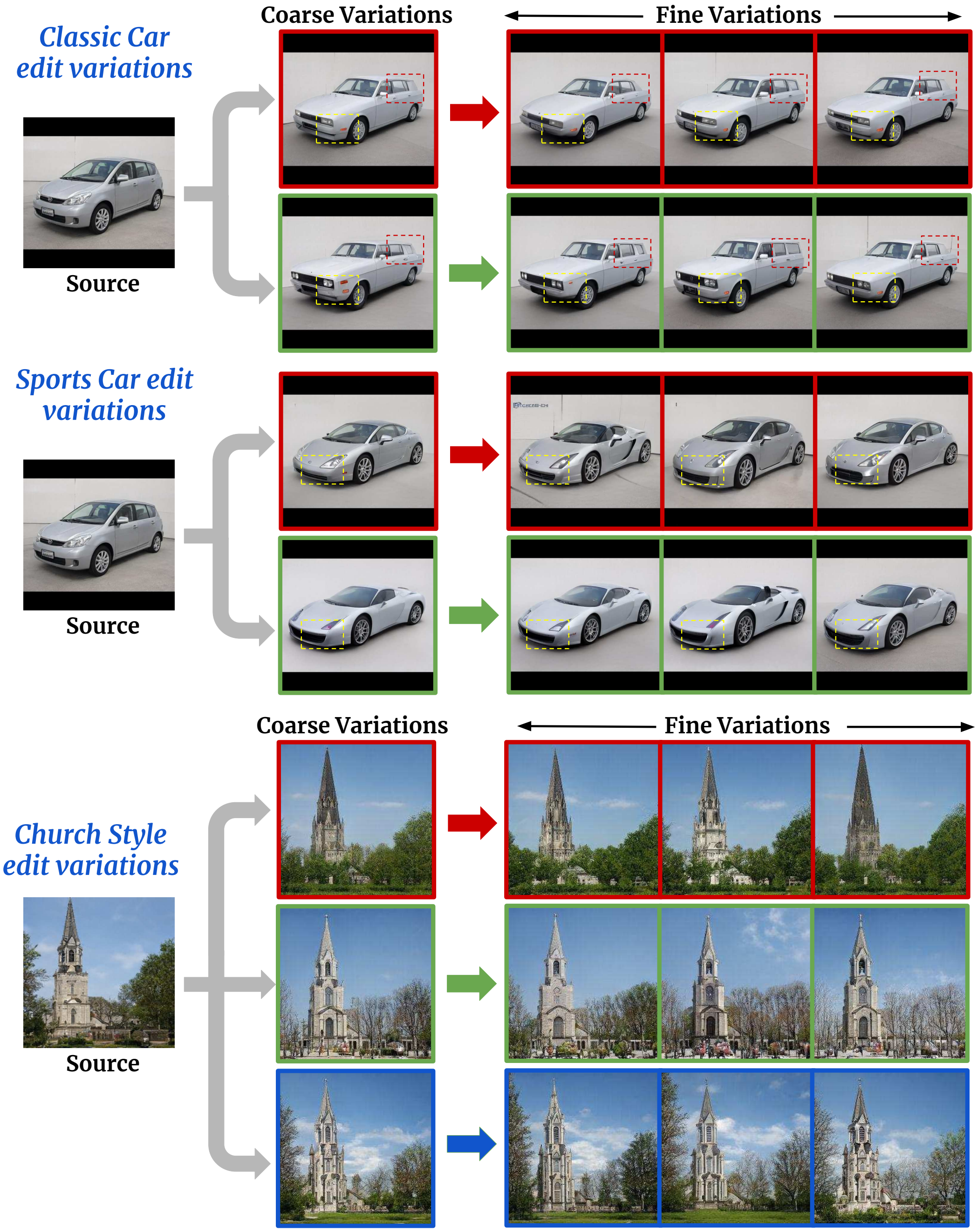

Figure 13. (Zoom to appreciate fine-variations). Hierarchical attribute editing for cars and churches in a coarse-to-fine manner. (Top) We present car style edit variations for 'Classic car' and 'Sports car' edits. Given a source image, we generate two diverse coarse variations for each car style. Next, finer style variations are generated, which can be observed as subtle changes in the headlamp and side windows of the cars, as highlighted in boxes. (Bottom) We first generate diverse coarse textures for churches with the same layout. Next, finer texture variations are generated for the selected coarse texture.

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3, 4

[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. 3, 4

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 2

[4] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*, 2021. 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 6

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[7] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[8] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 1

[9] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020. 2

[10] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. *arXiv preprint arXiv:2303.11306*, 2023. 1

[11] 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, 3

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[13] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 1, 3

[14] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 5, 6

[15] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021. 2