Supplementary Materials for GANFusion: Feed-Forward Text-to-3D with Diffusion in GAN Space

Souhaib Attaiki[1], Paul Guerrero[2], Duygu Ceylan[2], Niloy J. Mitra[2,3], Maks Ovsjanikov[1]

[1]LIX, École Polytechnique, IPP Paris
[2]Adobe Research
[3]University College London (UCL)

https://ganfusion.github.io/

Abstract

In the supplementary document, we provide a detailed description of the BLIP [3]-based captioning process. Additionally, we include further qualitative examples and evaluate on additional datasets to illustrate the performance of our method in comparison to other baselines.

Alongside this document, the project webpage presents a qualitative comparison encompassing a broader set of randomly selected results from our method and all baselines. This webpage also includes videos for several examples.

1. Additional details for BLIP-based captioning

We ask BLIP [3] the following questions:

- What is the gender of this person?
- What is this person wearing in top?
- What is the person wearing on their lower half?
- What color is this person wearing in top?
- What color is the clothing the person is wearing on their lower half?
- Is the person barefoot?
- What is the person wearing on their feet?
- What is the color of the thing the person is wearing on their feet?

and define a set of possible answers for each question:

- **`Gender:`** 'man', 'woman'
- **`Color:`** 'Red', 'Blue', 'Green', 'Yellow', 'Purple', 'Orange', 'Black', 'White', 'Gray', 'Pink', 'Brown', 'Gold', 'Silver', 'Beige', 'Maroon', 'Teal', 'Olive', 'Navy', 'Coral', 'Turquoise', 'Indigo', 'Khaki'
- **`Upper body clothing:`** 'T-shirt', 'Polo shirt', 'Dress shirt', 'Tank top', 'Crop top', 'Blouse', 'Sweater', 'Hoodie', 'Jacket', 'Coat', 'Blazer', 'Vest', 'Sweatshirt', 'Pullover', 'Cardigan', 'Tunic', 'dress', 'Jumpsuit', 'Romper', 'Suit', 'Pajamas', 'Raincoats', 'Windbreakers', 'Parkas', 'Puffer jackets', 'Trench coats', 'Pea coats', 'Duffle coats'
- **`Lower body clothing:`** 'Jean', 'pant', 'Sweatpant', 'Legging', 'Short', 'Skirt', 'Capri', 'Chino', 'Jogger'
- **`Footwear:`** 'Sneakers', 'Loafers', 'boots', 'heels', 'Flats', 'Sandals', 'Flip flops', 'Espadrilles', 'Oxfords', 'Clogs', 'socks'

We use the answers chosen by BLIP to fill in the following prompt template:

> `[Prefix] [Gender]` wearing a `[Upper Body Color] [Upper body clothing], [Lower body color] [Lower body clothing], [Footwear color] [Footwear].`

`[Prefix]` is randomly chosen from either *"a photo of a"* or *"a full-body photo of a"*, and `[Gender]` is selected randomly between the response of BLIP and *"person."* At test time, we randomly drop some of the items in the prompt template, such as the upper body color and clothing, or the footwear color and clothing.

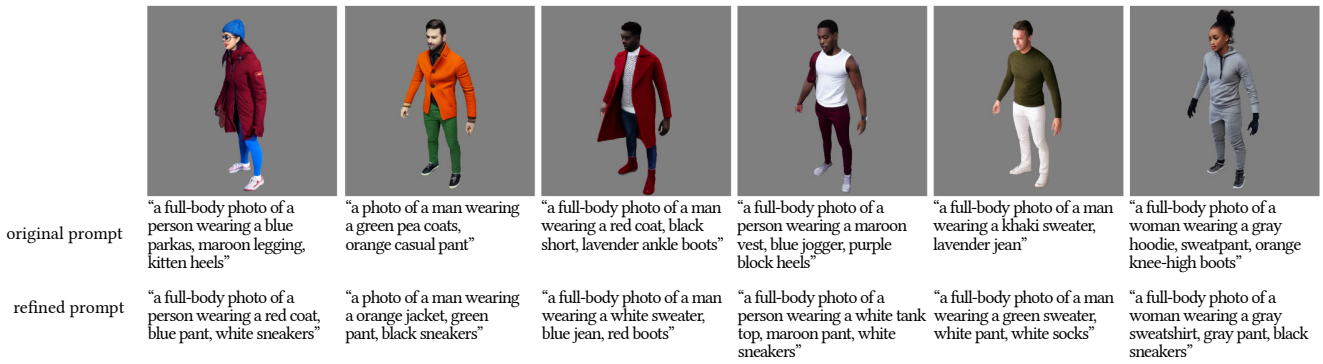

Figure 5. **Dataset examples.** We show a few images and corresponding prompts from our single-view image dataset. Prompts that were initially used to generate the images are often not accurate, we refine them with our BLIP-based captioning approach.

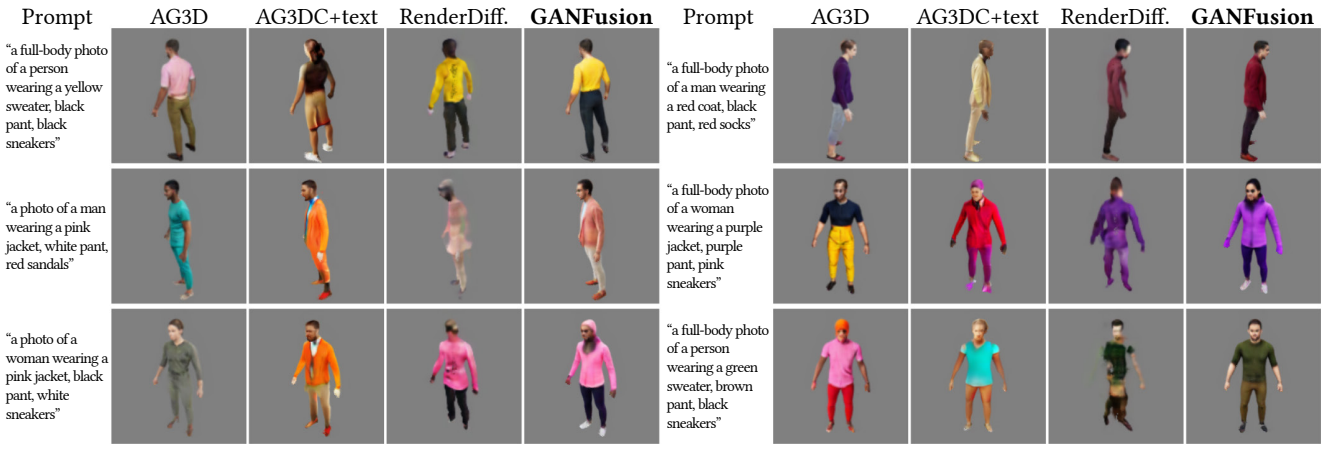

Figure 6. **Qualitative comparison without upsampling.** We provide visual results from our method as well as the baselines. We disable the 2D upsampler for AG3D and our method and provide renderings of the triplane features directly for all methods.

2. Additional Results

In this section, additional qualitative and quantitative results are provided.

In Figure 4 of the main document, we present qualitative results obtained by training our model on the FFHQ dataset [1]. We also evaluate our model quantitatively by measuring the FID of the generated samples and comparing our results to those of EG3D [1]. Our model achieves an FID score of 49.4, compared to 26.7 achieved by EG3D. While our FID score is higher than that of EG3D, we emphasize that our model is conditioned on text, a capability that EG3D does not possess. We attribute the increased FID to the automatic labeling process used in stage 2. Specifically, the VQA model employed produces a limited set of labels, which restricts the variety of faces learned by our model compared to the full FFHQ dataset. This limitation could be addressed by utilizing a more powerful labeling algorithm.

Beyond learning human faces, we also trained our model to generate realistic cat faces from the AFHQ dataset [1] using the EG3D [1] backbone in stage 1, and realistic 3D human figures from the DeepFashion dataset [4] using the AG3D [2] backbone in stage 1. Qualitative results for these experiments are provided in Figure 7 and Figure 8, respectively. These results demonstrate that our model generates high-quality, realistic images while closely following the provided prompts. This highlights the adaptability of our model to multiple domains and its ability to leverage different backbones in stage 1.

Figure 5 presents a selection of example images from the single-view image dataset we created, alongside their original captions and the refined captions, which were used to train our network.

Since RenderDiffusion does not employ an image upsampler, we also compared the results of the three main baselines without upsampling, as shown in Figure 6. Although omitting the upsampler reduces image detail in both AG3D and our method, both still significantly outperform RenderDiffusion.

Figure 9 showcases a qualitative comparison with all baselines, excluding the unsuccessful text-conditioned version of

Figure 7. **Qualitative results on the AFHQ dataset [1].** We replace AG3D [2] with EG3D [1] as the generator in our first stage to effectively enable text-conditioning on real-world cat data.

Figure 8. **Qualitative results on the DeepFashion dataset [4].** We use AG3D [2] as the generator in our first stage to effectively enable text-conditioning on realistic 3D human figures.

AG3D, which is presented in Figure 10. Due to training instability, experiments involving text-conditioning consistently led to mode collapse or other forms of training instabilities. The figure illustrates a case of severe mode collapse.

In Figure 11 (top row and bottom left), we showcase images produced by our method using identical prompts but different seeds, demonstrating the generation of diverse samples for the same prompt.

Furthermore, our primary objective is to demonstrate the feasibility of our approach using varied prompts that differ in properties such as colors and clothing items, rather than presenting a production-ready model. Generalizing to arbitrary prompts would require large-scale training on datasets like LAION [5], which is outside the scope of this study. Nonetheless, our model exhibits some generalization to out-of-distribution (OOD) prompts, facilitated by the pre-trained CLIP encoder. For instance, in Figure 11 (bottom right), our model correctly handles new colors like lavender,

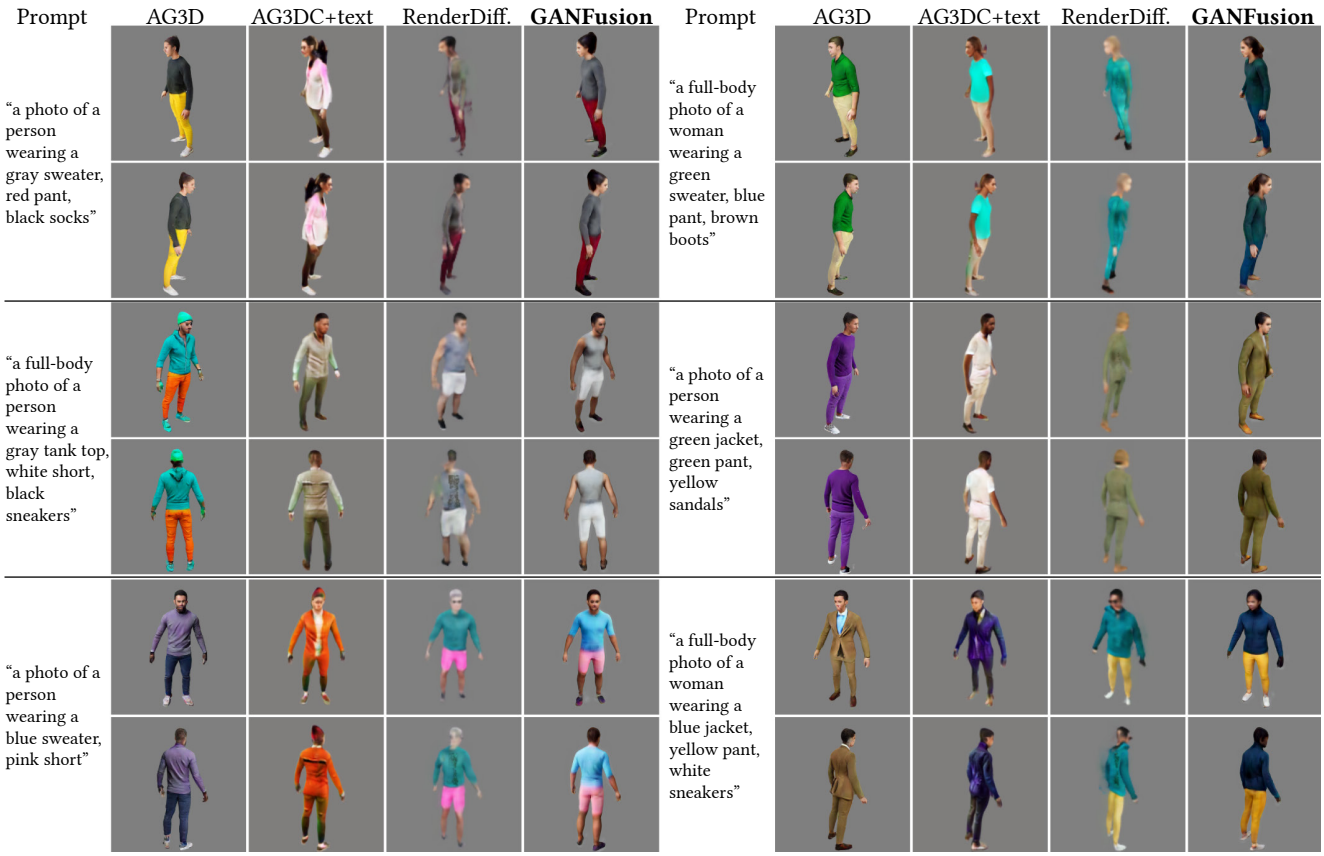

Figure 9. **Qualitative comparison.** We provide visual results from our method as well as the baselines.

Figure 10. **Text-conditional AG3D.** We attempt to add text conditioning to AG3D [2] by providing text embeddings as additional input both to the generator and the discriminator. However, we find that the training is not stable and does not converge.

which are absent from the training dataset, and supports some degree of grammatical rearrangement in sentences.

Acknowledgements The authors would like to thank the anonymous reviewers for their valuable suggestions. Parts of this work were supported by the ERC Consolidfator Grant No. 101087347 (VEGA) and the ANR AI Chair AIGRETTE.

References

[1] 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. 2, 3

[2] Zijian Dong, Xu Chen, Jinlong Yang, Michael J.Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *ICCV*, 2023. 2, 3, 4

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 1

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes,

Figure 11. **Top & Bottom left**: Diverse generations produced by our model using the same prompt but different seeds. **Bottom right**: Images showcasing our model's ability to generalize to some out-of-distribution prompts not encountered during training.

Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 3