Unsupervised Text-Guided Face Image Generation using Pre-Trained StyleGANv2

Mehmet Bayık, Erdem Eren Çağlar, Kaan Çakıroğlu

***Abstract*—This paper presents a novel approach to generating realistic human face images guided by textual descriptions using a pre-trained StyleGANv2. By employing the advanced generative adversarial networks (GANs) and language-image pre-training models' abilities, we designed a system that interprets textual input to provide high-quality images that adhere to the described attributes. We incorporate the BLIP and CLIP models to improve the understanding and generation process, enabling seamless integration of textual data into the visual generation pipeline.**

***Index Terms*—StyleGANv2, BLIP, CLIP, text-to-image, text-guided, image generation.**

I. Introduction

COMBINING text descriptions with image generation is a significant area of research, blending human language with the power of generative models. Our study focuses on using a pre-trained StyleGANv2, a model known for creating realistic images, on generating face images guided by text. We aim to refine this process by incorporating various components to control specific features in the output images. We aim to elevate and optimize the image generation process to respond to the textual inputs dynamically. These enhancements are aimed at enhancing the fidelity and accuracy of the images to the textual input, addressing common challenges in the domain. Through this perspective, we seek not only to improve the realism of the generated images but also to optimize the generation process to respond more effectively.

II. Related Work

Generative adversarial networks have demonstrated impressive improvements recently. StyleGAN [1], created by Karras and his team, is one such generative adversarial network known for its ability to create high-quality images of all kinds and at the disentanglement of latent factors compared to other models on similar tasks. In the field of text-to-image generation, several notable approaches have been developed, including the above-mentioned GAN models. One example is the "Generative Adversarial Text-to-Image Synthesis (AttnGAN)" [2], a model that creates images from textual descriptions using attention mechanisms to focus on different parts of the image, thereby producing more detailed and realistic results. Another significant model is "Generative Adversarial Networks for Image-to-Image Translation (CycleGAN)" [3], which translates images from one domain to another, such as converting images of horses to zebras and vice versa. Furthermore, there are popular pre-trained models like Dall-E 2 [4] and Craiyon (formerly known as Dall-E mini) [5], an open-source alternative to Dall-E 2. Dall-E 2 utilizes a VQ-VAE-2 model and a transformer model, while Craiyon comprises a transformer and a generator. Both models are capable of generating images from textual descriptions. Recently, it has also become possible to use the power of StyleGAN to generate images from textual descriptions, further advancing the capabilities in this field.

In recent studies, researchers have been combining generative adversarial networks with natural language processing (NLP), which integrates natural language into the process for different tasks. Accompanying these endeavors, various research groups work on forming connections between images and text using deep neural networks. One approach is to use StyleGAN along with techniques like "Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP)" [7]. BLIP uses the "Bootstrapped Language-Image Pre-training method," as its name suggests. The model initially employs a supervised approach via pre-training on paired data. In this stage, the alignment between image and text is aimed to be achieved. To enhance the model, after the supervised approach concluded, BLIP employs an unsupervised approach via self-training generating pseudo-labels for unpaired data. By increasing the total number of images it has learned, the model becomes knowledgeable in a broad range of visuals and text descriptions [7].

"Contrastive Language-Image Pretraining (CLIP)" developed by OpenAI [8], on the other hand, is another significant component and a big step in this direction. It bridges the gap between visual and textual data using natural language descriptions. It is trained using a contrastive learning objective that aligns images with corresponding descriptions in a shared multi-modal embedding space. In other words, CLIP learns a shared space where images and text descriptions can be understood simultaneously. Unlike traditional models trained on specific datasets for particular tasks, CLIP generalizes across a wide range of visual concepts described in language. Therefore, the model performs successfully on various vision tasks without the need for task-specific training data. Some researchers have used StyleGAN along CLIP, such as Zhang et al. [6]. In addition, other CLIP-based text-to-image models like "CLIPDraw" [9] generate images from textual descriptions by leveraging CLIP's shared embedding space. CLIPDraw is trained using a contrastive learning objective that aligns images with their corresponding descriptions in a shared multi-modal embedding space.

Combining the above-mentioned models helps generate realistic images and describes them in text. For example, Zhang et al., as mentioned above, made a system where people can guide image creation just by typing, demonstrating how GANs and NLP can team up for state-of-the-art advancements.

Currently, some researchers are training encoders to connect StyleGAN's image-making numbers to CLIP's shared space. This approach enables the control of image creation in detail using input natural language descriptions. These advancements convey how combining GANs, NLP, and vision models can make more interactive and flexible creative tools and contribute to tasks within this scope.

III. Experiments

This section details our proposed text-guided face image generation model's method; the data used and its acquirement, and our model's further details and configuration.

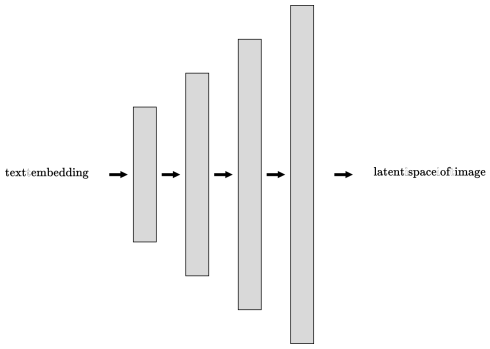

Fig. 1: Network architecture

Our approach involves image generation using the pre-trained StyleGANv2 model [12] on the FFHQ dataset [10] with 512x512-sized images. This significant model is employed to obtain the latent codes, which can also be described as the w vectors of the generated images. Having experimented with various datasets, it is determined that the proposed architecture is adequately tuned for this specific task. After acquiring the StyleGAN-generated images, these images are captioned – constructing a text description for the image by using the BLIP model [7]. This model provides text descriptions that are aligned with the input image. The obtained descriptions undergo the embedding process via "Contrastive Language-Image Pretraining (CLIP)" [8], a valuable approach to learning from natural language supervision [13].

A network is trained using the cumulative outcome of the above-mentioned sequence of operations with various models, which is the text embeddings as the input. The network generates the corresponding W vectors that recreate the images after feeding them back into StyleGANv2. The network structure is provided in Fig. 1. The network consists of four blocks with an input size of 512. The network output has a size of 8192, enabling us to feed this result into StyleGANv2, as mentioned above. It utilizes dropout and Leaky ReLU in each layer to mitigate the risk of overfitting. By employing this structure, we generated images that are responsive to the input texts without utilizing a static dataset, and we generated the images to be used for training on the fly, embracing an unsupervised approach.

Consequently, our approach does not rely on a pre-existing dataset; instead, we utilize a dynamic, on-the-fly data generation process leveraging the capabilities of the StyleGANv2 model. This method allows for the creation of diverse, high-quality facial images directly influenced by textual descriptions generated in real-time. The benefits of this approach include the ability to generate a vast array of unique and varied images that are not limited by the constraints of a fixed dataset. Nevertheless, having less control over the input images and their paired textual descriptions compared to a static dataset constitutes a constraint in some cases, as we cannot check the descriptions and what the model learns. Additionally, the sample size is not fixed but is determined based on the requirements of each experimental setup. This flexibility enables us to increase the quantity of training data to enhance model performance or for testing to evaluate model robustness comprehensively. Our model differs from the common approach, including a static dataset, so we do not employ a static train-test split ratio. We generate distinct batches of data for training and testing. This ensures that the model is evaluated on images and text descriptions it has not encountered during training.

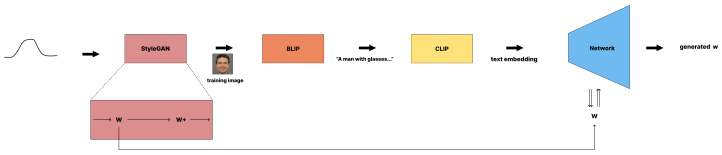

Fig. 2: Architecture design of the proposed model

A. Initial model based on AFHQ cat dataset

Initially, the architecture was built on a pre-trained StyleGANv2 model with AFHQ Cat 512x512 [11], using the previously explained approach. Despite giving a meaningful result, it has been realized that BLIP captions on this dataset are unsuitable for this task due to the poor quality of the captions. Therefore, it is decided to utilize the pre-trained StyleGANv2 model trained on the FFHQ dataset [10].

B. BLIP specification via keyword constraints

In order to address the subpar outputs from BLIP, such as generic captions like "a man looking at a camera," we implemented a keyword constraint system to enhance the quality of results. This system ensures that BLIP captions are evaluated at each iteration for the presence of specific adjectives, such as "blue," "long," or "dark" - depending on the objective. If the generated caption lacks these keywords, the BLIP model parameters are adjusted, and a new caption is generated for up to five attempts. These adjusted BLIP parameters are the search method, number of beams in beam search, top p value in nucleus sampling, as well as the maximum and minimum length constraints for captions in both beam search and nucleus sampling techniques. This iterative approach improved our training process, leading to more accurate and descriptive image captions generated on the fly.

C. Curated latent vector-image-caption dataset

To enhance the training process, dataset of 2,000 images is created with corresponding captions as well as their latent

vectors. This dataset was generated by using the StyleGANv2 model pretrained on the FFHQ dataset to produce images, and then generating captions for these images with the BLIP model. Afterward, images with low-quality captions are filtered out to ensure consistency and reliability. This approach reduced the computational cost of generating captions on the fly and provided greater control over caption quality, allowing us to manually exclude unsuitable captions.

D. Using Latent Directions in domain of StyleGANv2 to Gain Control Over Training Samples

In our pursuit of generating high-quality facial images with critical attributes accurately described by the text prompt, we adopt a dynamic approach to enrich the training process by introducing additional latent direction vectors into StyleGANv2's latent space using the pre-computed w's representing the directions to promote specific attributes. These vectors are computed employing logistic regression on attribute-labeled human face images generated by StyleGANv2.

Randomly generated face images typically do not inherit obvious appearance features that BLIP can recognize. BLIP ends up providing generic and often useless text descriptions of these generated images. Such generic descriptions are inadequate for training our encoder, as they do not help in learning the relationship between text embeddings and attributes in StyleGANv2's latent space. Introducing direction vectors for attributes such as smile, gender, or age mitigates this problem. These direction vectors steer the latent space representations toward more recognizable and distinct features, enhancing the utility of BLIP's descriptions and facilitating better encoder training.

(a) A man looking at the camera

(b) A man with a smile

(c) A man with beard

(d) A smiling man with beard

Fig. 3: Example Usage of Pre-computed w Directions to Obtain Smiling Faces and make BLIP generate quality descriptions.

Our revised training strategy begins by initially training the encoder with random latent vectors (w) sampled from StyleGAN2's latent space. Even though the text descriptions provided by BLIP are more generic at this stage, this initial phase allows the encoder to gain familiarity with the latent space structure of StyleGAN2. This enables it to learn the mapping from textual embeddings to manifold latent vectors without specific attribute manipulation.

After the initial phase, we introduce a dynamic adjustment to the latent direction vectors corresponding to critical attributes, such as age, gender, and smile, during the training process. We add the latent directions corresponding to these critical attributes to the randomly sampled w in each iteration. This dynamic adjustment allows for greater exploration of the latent space and promotes diversity in the synthesized images. Consequently, the increase in diversity and the sharpness of the attributes in the images helps BLIP generate more precise text descriptions, enabling the encoder to embed more diverse prompts into StyleGANv2's w space.

In conclusion, the adoption of a dynamic approach to incorporating latent direction vectors into the training process represents a significant refinement in our methodology. By dynamically adjusting the latent direction vectors in each iteration, we enable the encoder to learn the complex relationships between textual embeddings and attribute variations in StyleGAN2's latent space in a more versatile and adaptive manner. This approach holds promise for further enhancing the fidelity and diversity of the generated facial images, ultimately advancing the capabilities of text-to-image synthesis techniques and opening up new avenues for creative content generation and facial attribute manipulation.

E. Achieving Diversity in Inference Using a Guidance Scale

In our architecture design, the inference process is deterministic. Specifically, our encoder maps the text embeddings of a given prompt to a consistent w in StyleGANv2's latent space without introducing randomness. For instance, our architecture will generate an identical image of the same individual each time a prompt such as "a man" is used. Additionally, our architecture tends to produce variations of the same individual for different prompts. For example, when provided with the prompt "a man with glasses," the generated image is highly likely to depict the same individual from the "a man" prompt, but with glasses. This characteristic limits the diversity of the outputs generated by our architecture. To mitigate this limitation, we introduce a parameter for randomness, referred to as the "guidance scale." This parameter controls the trade-off between adherence to the text prompt and output diversity. In practice, we add a stochastic component to our encoder's output, scaled by the guidance scale. Let λ be the guidance scale. Then our generated image would be computed as:

$$G(Encoder(CLIP(textprompt)) + \lambda \times w_r)$$

Where w_r is a random w obtained by feeding a random Gaussian sample z through the StyleGANv2's mapping network.

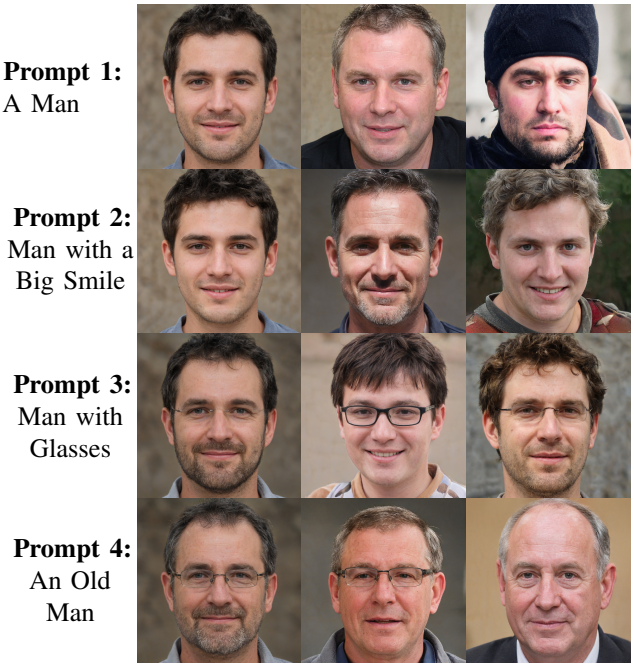

Fig. 4: Generated images for different prompts. First column: Guidance Scale = 0 (No randomness), Rest of the columns: Guidance Scale = 0.4

F. Baselines

Our model is benchmarked against three leading text-to-image synthesis models: Stable Diffusion [14], DALL-E mini [5], and DALL-E 2 [4]. These models represent significant advances in the field and are robust comparators due to their state-of-the-art performance in generating high-quality images from textual descriptions. Stable Diffusion is a latent diffusion model known for its ability to generate detailed images across various styles and subjects. DALL-E mini is widely used due to its accessibility and speed in text-to-image generation. DALL-E 2 is an advanced model developed by OpenAI that provides highly realistic images based on textual descriptions using newer techniques for better text and image alignment.

We utilize three metrics: the Fréchet inception distance (FID), Kernel inception distance (KID), and CLIP score. FID measures the distance between feature vectors calculated for real and generated images. Lower values indicate that the generated images are more similar to the real images, suggesting better model performance. KID computes the distance between subsets of images in Inception space. Different from FID, KID offers an objective similarity estimate between the output and real image distributions. CLIP Score, on the other hand, considers the semantic consistency between the generated output image and the text prompts that were used for guidance. Higher scores show greater alignment between text and image, demonstrating the model's ability to interpret textual inputs effectively.

IV. Results

This section provides our model's evaluation results and outcomes and its comparison with the baseline models. The

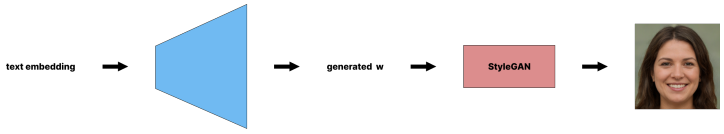

Fig. 5: Overall process of model inference

models used in this section for comparison are Stable Diffusion [14], Dall-E 2 [4], and Dall-E mini [5].

The evaluation results, summarized in I, demonstrate that our model achieves a high CLIP score that is close to other models, meaning that our model was sufficiently consistent with the given text prompts according to the CLIP score evaluations. As can be seen from the table, our FID score was higher than the other models. This leads to the result that our model was not able to generate images that are as realistic as Stable Diffusion, Dall-E 2, and Dall-E mini – solely based on the FID scores. For Dall-E 2's FID score, the FID calculation required 2000 images at minimum, and due to its fee and financial constraints, the FID score for Dall-E 2 could not be provided. Regarding KID, our model outputs a high score, which is to the detriment of our model. Nevertheless, Table I provides us insights into the quality of the outcomes of our model and its performance compared to well-known and successful models.

Evaluation Metrics			
Model	FID ↓	CLIP Score ↑	KID ↓
Ours	99.89	0.928	0.080
SD [14]	**63.01**	0.949	**0.038**
Dall-E mini [5]	66.96	**0.956**	0.043
Dall-E 2 [4]	-	0.944	0.041

TABLE I: Evaluation Metrics (FID, KID, CLIP Score)

Regarding the qualitative comparisons, Fig 6 illustrates sample outputs from our model alongside those from the baseline models detailed above, demonstrating the qualitative improvements in image relevance to the text descriptions. As seen from these images, our model successfully generates images that are consistent with the textual descriptions. Compared with the Stable Diffusion model, our outcomes could be constituted as less diverse, as all samples follow a similar pattern in terms of appearance. As we employed an unsupervised and on-the-fly approach, the output quality and correspondence with the input prompts rely heavily on the data generation process. If StyleGANv2-generated images did not consist of the described features or BLIP could not generate a relevant prompt, the desired features may not be covered by our model. In other words, our reliance on these components forms constraints for our model. Furthermore, comparing the Dall-E mini results with ours, it is apparent that our outcomes are far more realistic than the ones generated by the Dall-E mini model. Although the images of the Dall-E mini consist of features that convey better alignment with the text, the realism of the images is compromised. In addition, as shown in the figure, the Dall-E 2 model's outcomes result in diverse and realistic images compared to all other models, including ours.

Comparing the time, budget, and computation powers of the teams releasing the above-mentioned baseline models, we believe our study constitutes a successful approach to human

face generation based on textual descriptions and suggests a vast potential for future improvements.

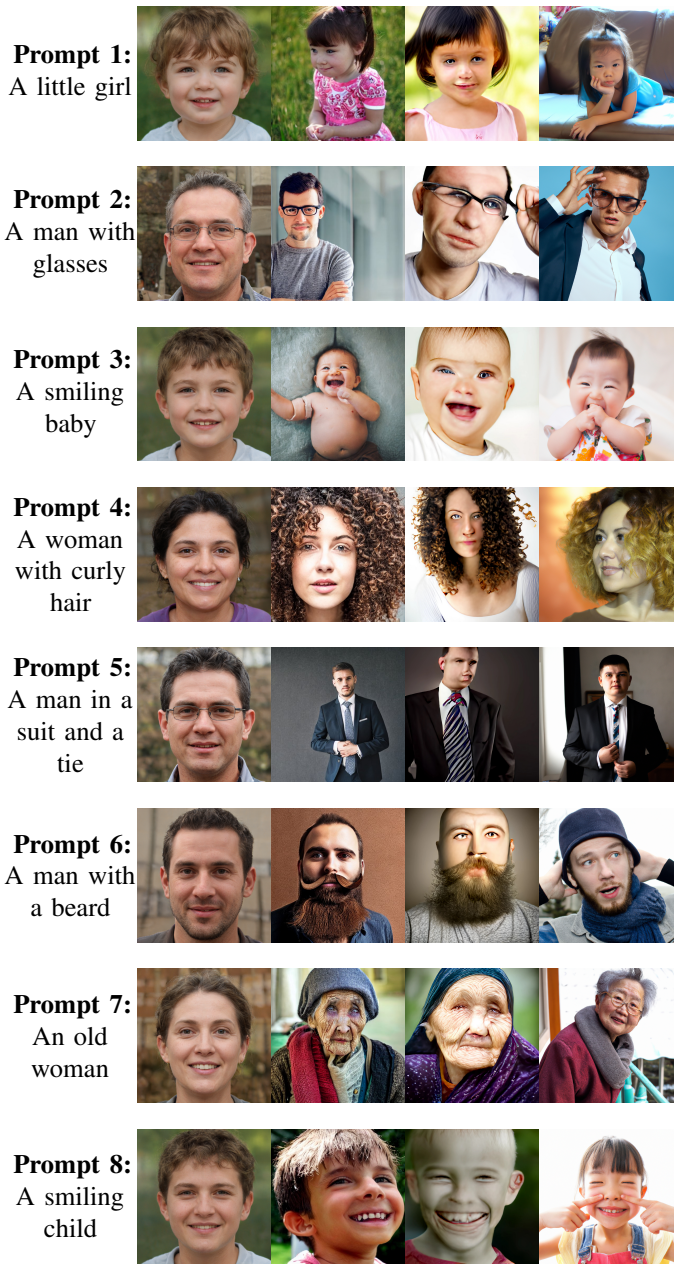

Fig. 6: Samples from predictions. First column: our model with a guidance scale of 0.3, second column: SD, third column: Dall-E mini, fourth column: Dall-E 2.

REFERENCES

[1] Karras, Tero, Laine, Samuli, and Aila, Timo. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[2] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks." Proceedings of the IEEE International Conference on Computer Vision, 2017.

[4] OpenAI, "DALL-E 2," OpenAI. [Online]. Available: https://openai.com/index/dall-e-2/. [Accessed: May 20, 2024].

[5] B. Dayma, "dalle-mini," GitHub. [Online]. Available: https://github.com/borisdayma/dalle-mini. [Accessed: May 20, 2024].

[6] You, X., & Zhang, J. (2023). "TextCLIP: Text-Guided Human Face Image Generation and Manipulation Without Adversarial Training." arXiv preprint arXiv:2309.11923.

[7] J. Li et al., "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn., 2022.

[8] OpenAI, "CLIP: Contrastive Language–Image Pre-training," GitHub repository, 2023. [Online]. Available: https://github.com/openai/CLIP. [Accessed: Mar. 10, 2024].

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, Ilya Sutskever. "CLIP-Draw: Text-to-Drawing Generation from Natural Language Descriptions." arXiv preprint arXiv:2106.14843, 2021.

[10] NVIDIA Corporation, "Flickr-Faces-HQ (FFHQ) Dataset," GitHub Repository, [Online]. Available: https://github.com/NVlabs/ffhq-dataset.

[11] Y. Choi, Y. Uh, J. Yoo, and J. W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8188-8197.

[12] NVIDIA, "StyleGAN2," NVIDIA NGC Catalog, 2023. [Online]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/research/models/stylegan2/files. [Accessed: Mar. 10, 2024].

[13] Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." Proceedings of the International Conference on Machine Learning, 2021.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 10684-10695, 2022.