

3D-CLFusion: Fast Text-to-3D Rendering with Contrastive Latent Diffusion

Yu-Jhe Li

Kris Kitani

Carnegie Mellon University

{yujheli, kkitani}@cs.cmu.edu



Figure 1: Multi-view images generated from text prompts by 3D-CLFusion.

Abstract

We tackle the task of text-to-3D creation with pre-trained latent-based NeRFs (NeRFs that generate 3D objects given input latent code). Recent works such as DreamFusion and Magic3D have shown great success in generating 3D content using NeRFs and text prompts, but the current approach of optimizing a NeRF for every text prompt is 1) extremely time-consuming and 2) often leads to low-resolution outputs. To address these challenges, we propose a novel method named 3D-CLFusion which leverages the pre-trained latent-based NeRFs and performs fast 3D content creation in less than a minute. In particular, we introduce a latent diffusion prior network for learning the w latent from the input CLIP text/image embeddings. This pipeline allows us to produce the w latent without further optimization during inference and the pre-trained NeRF is able to perform multi-view high-resolution 3D synthesis based on the latent. We note that the novelty of our model lies in that we introduce contrastive learning during training the diffusion prior which enables the generation of the valid view-invariant latent code. We demonstrate through experiments the effectiveness of our proposed view-invariant diffusion process for fast text-to-3D creation, e.g., 100 times faster than DreamFusion. We note that our model is able to serve as the role of a plug-and-play tool for text-to-3D with pre-trained NeRFs.

1. Introduction

We aim the tackling the task of text-to-3D domain-specific content creation. 3D content can be represented with neural radiance field (NeRF) [19] in a photorealistic way. Currently, text-to-3D with NeRFs have been explored in DreamField [11], DreamFusion [26], or Magic3D [17]. By leveraging the pre-trained models from CLIP [27] or diffusion priors [30] as the learning objective, these works are capable of producing 3D content given the input text prompt through training a NeRF from scratch. However, training one individual model for each text prompt would cause the first issue: slow inference (*i.e.*, around one hour) for the above models. Due to no real image guidance during updating, the model would lead to the second issue: low-resolution multi-view rendering.

Recently, 3D latent-based models using radiance fields (*i.e.*, NeRFs [19]), such as EG3D [2] or StyleNeRF [8], have been proposed for unsupervised generation of multi-view consistent images. Similar to StyleGANs [13, 14], these NeRFs learn a controllable $w \in \mathcal{W}$ space and enable explicit 3D camera control, using only single-view training images. To achieve fast text-to-3D creation, one straightforward way is to produce the w latent from the input text prompt without further updating the NeRF model. This idea has been explored in clip2latent [25] for text-to-2D creation, which takes less than one minute to generate high-resolution realistic images from the input text prompt

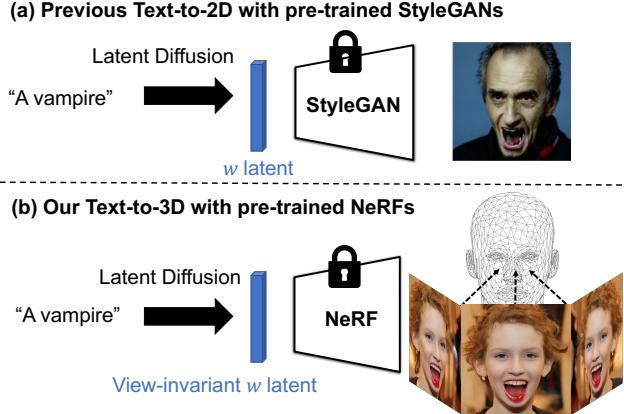


Figure 2: **Comparison of text to content creation with pre-trained GANs.** Compared with existing effective text-to-2D with StyleGANs [25], it is more challenging to perform text-to-3D with latent-based NeRFs via diffusion process since the denoised latent is assumed to be view-invariant.

with a trained diffusion prior. However, there are two main challenges for directly applying the clip2latent on 3D latent-based NeRF. First, the latent space $w \in \mathcal{W}$ is assumed to be view-invariant for 3D NeRFs which is different from 2D StyleGANs. That is, if only single-view prompts (image CLIP embeddings) are used to train the diffusion prior, the produced latent code may be only valid when generating images of the same view (camera pose) where this issue is also noted in NeRF-3DE [16]. To address this issue, we can use multi-view CLIP prompts from the same latent w to train the diffusion prior. However, this leads to our second challenge: how to ensure the diffusion process to produce view-invariant latent from multi-view CLIP embeddings, as shown in Figure 2.

To properly achieve fast and realistic text-to-3D creation with pre-trained NeRFs, we propose a framework named 3D-CLFusion to produce the view-invariant latent code from an input text prompt. Our 3D-CLFusion is composed of a diffusion prior network, a pre-trained clip model, and a pre-trained generator (*i.e.*, NeRF). First, in order to produce the latent code w for the generator to render from the input text prompt, we introduce the diffusion prior which takes the CLIP text embedding as a condition and produces w latent in each diffusion step. Since we do not have labeled text embeddings, we leverage the image CLIP embeddings for training the diffusion prior. This training strategy is inspired by clip2latent [25] where we believe CLIP text and image embeddings share the closed latent space. Second, in order to learn the view-invariant latent in \mathcal{W} space, we leverage the multi-view images generated by the model itself and contrastive learning to ensure the produced latent code in \mathcal{W} are the same given different CLIP embeddings

from a different view. Later in the experiments, we will show the significance of our introduced contrastive learning in the diffusion process. We have demonstrated the effectiveness of the proposed framework for the fast text-to-3D using StyleNeRF [8] and EG3D [2] as the pre-trained generators in Figure 1 and the experiments. Compared with DreamFusion [26] and Magic3D [17] which take around one hour for inference with NeRF, our model only takes less than 30 seconds for each 3D object. The contributions of this paper are summarized as follows:

- We demonstrate the challenges of the task text-to-3D directly using latent-based NeRFs and the limitations of the current models for this task.
- We propose a framework named 3D-CLFusion, which aims to produce view-invariant latent for rendering 3D objects with NeRFs from the input text prompt.
- Compared with the existing models leveraging diffusion priors for text-to-2D with pre-trained StyleGANs, our designed model achieves more effective text-to-3D with pre-trained NeRFs.
- Though the 3D object created by our model is limited to the class of pre-trained model, the inference time is at least 100 times faster than the existing text-to-3D creation from neural rendering.

2. Related Works

Text-to-image generation. There are several text-to-image generative models that have been proposed in recent months. The majority of these models are trained using large amounts of text-image paired data where image generation is conditioned on the text input. To achieve high-quality and accurate image generation from text, several of the existing approaches make use of pre-trained CLIP to produce text embedding. One line of research is to leverage diffusion models in [28, 29, 30], where the models directly learn the mapping between text and image with diffusion priors (*i.e.*, Dalle 2 [28] and Glide[20]) or sample from a low-resolution latent space and decode the latent into high-resolution images (*i.e.*, Stable Diffusion [29]). Another line of work [24, 7, 1, 15] is to perform text-guided generation or editing relying on CLIP and the pre-trained StyleGANs [13, 14]. StyleGANs have achieved high image quality and support different levels of semantic manipulation in the $w \in \mathcal{W}$ latent space. Recently, clip2latent [25] has been proposed to produce the w latent in StyleGAN from the input text prompt with the diffusion model. However, most of these works focus on the rendering of 2D images and are not capable of manipulating camera poses easily as NeRF [19].

3D image synthesis with NeRFs. Methods built on implicit functions, *e.g.*, NeRF [19], have been proposed in

[3, 31, 23, 21]. To generate high-resolution images conditioned on the input style latent code, EG3D [2], StyleNeRF [8], VolumeGAN [33], StyleSDF [22], and GMPI [35] have been developed. In addition, some works such as Sofgan [4] and Sem2NeRF [5] are able to perform multi-view synthesis with NeRF by taking into multi-view or single-view semantic masks. However, most of these works are not capable of generating 3D objects given purely input text. We will demonstrate the effectiveness of our proposed diffusion prior to serving as a text-to-3D plug-and-play tool into EG3D [2] and StyleNeRF [8] in this work.

Text-to-3D generation with NeRFs. In recent years, several models [11, 26, 17] for the task of text-to-3D generation have been proposed using NeRFs [19]. Dream Fields [11] leverage the pre-trained image-text encoder (*i.e.*, CLIP [27]) as the image guidance to optimize the neural implicit scene representations (*i.e.*, NeRFs) through online training. Since the pre-trained CLIP models may not be effective image-level generation objectives, some works such as DreamFusion [26] and Magic3D [17] turn to train the NeRF using pre-trained diffusion prior [30] instead of CLIP. Though DreamFusion [26] and Magic3D [17] are capable of performing satisfactory 3D content creation and rendering with an open-vocabulary input text prompt, the inference takes 1.5 hours and 40 minutes for each model to train a NeRF from scratch. In this work, we aim to resolve the issue by leveraging the pre-trained NeRFs and diffusion prior, and producing the latent in less than a minute.

3. The Proposed Method

3.1. Problem Formulation and Overview

Given the input text prompt, our goal is to generate the conditioned multi-view images with the pre-trained latent-based NeRF generator as our text-to-3D task. Specifically, given the input text embedding with CLIP, denoted as e_t , we aim to produce the corresponding w latent¹ as output. The pre-trained NeRF generator denoted as G , is able to synthesize images x given different camera poses p : $x = G(w, p)$.

In order to achieve text-to-3D creation by producing the w latent for fast NeRF rendering, we propose a generative framework named 3D-CLFusion, which is presented in Figure 3. First, in order to produce the latent w for the generator to render from the input text prompt, we introduce the diffusion prior network (f_θ where θ denotes the network parameters) which is able to take the CLIP text embedding e_{txt} as an input condition and produces w_0 latent in the last diffusion step. More details of the diffusion/denoising process will be presented later. Since we do not have labeled

text embeddings for the ground-truth latent w_0 , we leverage the image CLIP embeddings for training the diffusion prior. We assume CLIP text and image embeddings share the closed latent space. Second, in order to learn the view-invariant latent w in \mathcal{W} space, we leverage the multi-view CLIP image embeddings from the images generated by the model and apply the contrastive loss to ensure the produced latent in \mathcal{W} are view-invariant given different CLIP embeddings from a different view. After the training stage, we can sample text CLIP embedding from a given text input prompt and generate the corresponding 3D multi-view images with the pre-trained NeRF.

3.2. Pre-trained Latent-based NeRF

Latent-based Neural Implicit Representation. Following StyleGANs [13, 14], Latent-based NeRFs [8] also introduce the mapping network f which maps noise vectors from a spherical Gaussian space \mathcal{Z} to the style space \mathcal{W} . f consists of several MLP layers and the input style vector $w \in \mathcal{W}$ can be derived by $w = f(z), z \in \mathcal{Z}$. Following the neural rendering mechanism in NeRF [19], all of our pre-trained latent-based NeRF also takes the position $u \in \mathbb{R}^3$ and viewing direction $d \in \mathbb{S}^2$ as inputs, and predicts the density $\sigma(u) \in \mathbb{R}$ and view-dependent color $c(u, d) \in \mathbb{R}^3$.

Volume Rendering with Radiance Fields. Once we have the color and density for each coordinate and view direction, we render the color $C(r)$ for each pixel along that camera ray $r(t) = o + td$ passing through the camera center o with volume rendering [12]:

$$C_w(r) = \int_{t_n}^{t_f} T(t)\sigma_w(r(t))c_w(r(t), d)dt, \quad (1)$$

where $T(t) = \exp\left(-\int_{t_n}^t \sigma_w(r(s))ds\right).$

The function $T(t)$ denotes the accumulated transmittance along the ray from t_n to t . In practice, the continuous integration is discretized by accumulating sampled points along the ray. More details can be obtained in [19, 8, 2].

3.3. Latent Diffusion with CLIP embedding

In order to produce the corresponding w from the input text prompt e_{txt} or image prompt e_{img} , we choose to employ a latent diffusion process to learn the mapping given the success of latent diffusion models in [29, 25, 28]. Now we present the overview of the latent diffusion model which contains: forward (diffusion) and backward (denoising) processing.

As stated in DDPM [9] and Latent Diffusion [29], we can formulate the diffusion process of w latent diffusion for our task as:

¹latent $w \in \mathcal{W}$ [32] or the extended latent $w \in \mathcal{W}+$ [14]

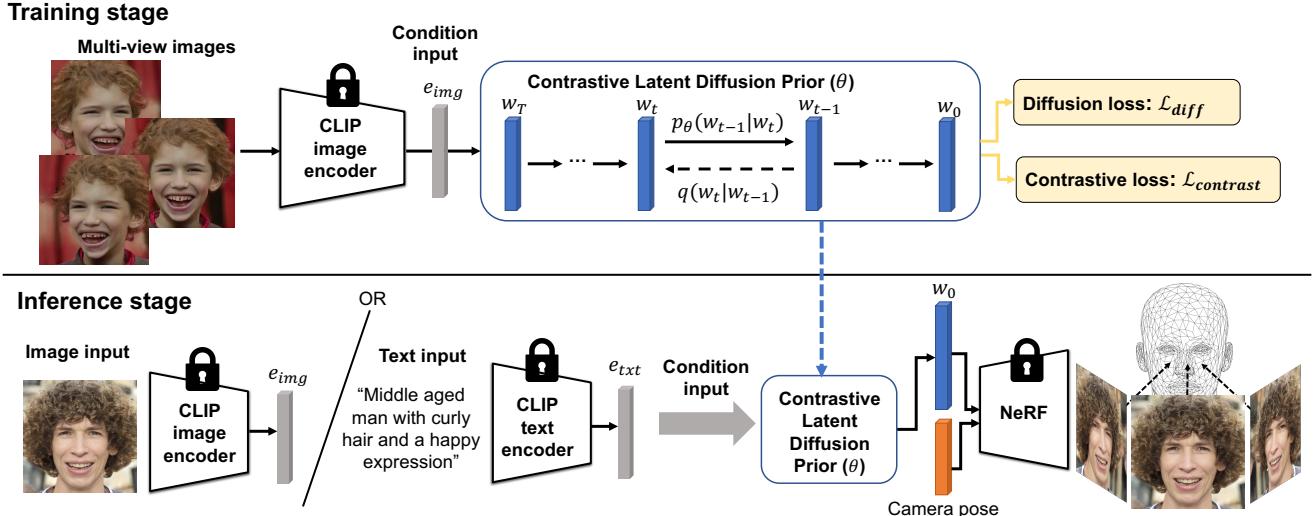


Figure 3: **Overview of our proposed 3D-CLFusion.** It consists of three modules: CLIP text/image encoders, contrastive latent diffusion prior, and pre-trained latent-based NeRF. More details can be referred to the section 3.

$$q(w_{1:T}|w_0) = \prod_{t=1}^T q(w_t|w_{t-1}), \quad (2)$$

$$q(w_t|w_{t-1}) = \mathcal{N}(w_t; \sqrt{1 - \beta_t} w_{t-1}, \beta_t \mathbf{I}),$$

where $w_1 \dots w_T$ are the latent of the same dimensionality of w_0 and $\beta_1 < \dots < \beta_T$ are variance schedule. The reverse process can also be formulated as:

$$p_\theta(w_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(w_{t-1}|w_t), \quad (3)$$

$$p_\theta(w_{t-1}|w_t) = \mathcal{N}(w_{t-1}; \mu_\theta(w_t, t), \Sigma_\theta(w_t, t)),$$

where θ is the learnable parameters. If we set $\Sigma_\theta(x_t, t) = \sigma_t^2 \mathbf{I}$ as fixed variance, we only need to learn $\mu_\theta(w_t, t)$. Since we can denote $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, we can estimate the forward process posteriors conditioned on w_0 as:

$$q(w_{t-1}|w_t, w_0) = \mathcal{N}(w_{t-1}; \tilde{\mu}_t(w_t, w_0), \tilde{\beta}_t \mathbf{I}),$$

$$\tilde{\mu}_t(w_t, w_0) = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} w_0 + \frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} w_t, \quad (4)$$

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t.$$

As stated in DDPM [9], we can choose to learn the $\mu_\theta(w_t, t)$ in Equ. 3 and reparameterize it as:

$$\mu_\theta(w_t, t) = \tilde{\mu}_t(w_t, w_0) = \frac{1}{\sqrt{\alpha_t}} (w_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(w_t, t, e)),$$

$$\text{and } w_0 = \frac{1}{\sqrt{\alpha_t}} (w_t, \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(w_t, t, e)) \quad (5)$$

where ϵ_θ is a function approximator intended to predict the noise ϵ from w_t , timestep embeddings t , and the CLIP embeddings e_{txt} or e_{img} . Or we can reparameterize $\mu_\theta(w_t, t)$ as:

$$\mu_\theta(w_t, t) = \tilde{\mu}_t(w_t, w_0) = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} f_\theta(w_t, t, e) + \frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} w_t, \quad (6)$$

where f_θ is a function approximator intended to predict the w_0 from w_t , timestep embeddings t , and the CLIP embeddings e_{txt} or e_{img} . Later we will show learning f_θ in Equ. 6 is preferable for contrastive learning compared with learning ϵ_θ in Equ. 5. The diffusion prior loss can be defined as:

$$\mathcal{L}_{diff} = \mathbb{E}[\|\epsilon - \epsilon_\theta(w_t, t, e)\|^2] \quad (7)$$

$$\text{or } \mathcal{L}_{diff} = \mathbb{E}[\|w_0 - f_\theta(w_t, t, e)\|^2]$$

3.4. View-invariant Latent Diffusion

For 3D latent-based NeRF diffusion, the latent w not only needs to match the input CLIP embedding e but is also assumed to be view-invariant. As stated in NeRF-3DE [16], only the latent w inside the valid manifold in the latent space \mathcal{W} for latent-based NeRF is capable to produce reasonable view-invariant 3D objects, while the latent outside the manifold would lead to severe 3D distortion.

Since the CLIP image embeddings generated from the same w with different camera poses p_i , e_{img}^i representing CLIP image embeddings from different views are also different. Hence, the produced w_t using Equ. 3 would not be view-invariant. We need to ensure the latent produce from

the diffusion prior in each step are view-invariant as:

$$\begin{aligned} f_{\theta}(w_t, t, E_{clip}^{img}(x_i)) &= f_{\theta}(w_t, t, E_{clip}^{img}(x_j)) \\ x_i &= G(w_0, p_i), i \neq j \end{aligned} \quad (8)$$

where $p_i, i = 0, 1, \dots, n$ represent camera poses. Minimizing the objective allows the model to learn the view-invariant latent code \hat{w} since it maps multi-view embeddings $e_i = E_{clip}^{img}(x_i)$ (controlled by the pose p_i) to the same latent code w for each set of the training sample. During inference, a single-view CLIP embedding (either text or image embeddings) is mapped to the latent code \hat{w} which can produce multi-view images of the same identity by changing the poses.

To ensure this, we can use contrastive learning to train the diffusion prior by applying constraints on $\tilde{w}_0 = f_{\theta}(w_t, t, e)$ in Equ. 6. Specifically, we perform contrastive learning with L2 loss \mathcal{L}_2 and triplet loss \mathcal{L}_{tri} on the produced w_0 in each diffusion step. We can formulate \mathcal{L}_2 loss as:

$$\mathcal{L}_2 = \|f_{\theta}(w_t, t, e^i) - f_{\theta}(w_t, t, e^j)\|^2, \quad (9)$$

where e_i and e_j are produced by the same ground-truth w . To maximize the inter-class discrepancy while minimizing intra-class distinctness, we introduce \mathcal{L}_{tri} . Specifically, for each input image embedding e , we sample a positive image embedding e_{pos} with the same identity label and a negative image e_{neg} with different identity labels to form a triplet tuple. Then, the following equations compute the distances between e and e_{pos}/e_{neg} :

$$\begin{aligned} d_{pos} &= \|f_{\theta}(w_t, t, e) - f_{\theta}(w_t, t, e^{pos})\|_2, \\ d_{neg} &= \|f_{\theta}(w_t, t, e) - f_{\theta}(w_t, t, e^{neg})\|_2, \end{aligned} \quad (10)$$

With the above definitions, we have the triplet loss \mathcal{L}_{tri} defined as:

$$\mathcal{L}_{tri} = \max(0, m + d_{pos} - d_{neg}), \quad (11)$$

where $m > 0$ is the margin used to define the distance difference between the positive image pair d_{pos} and the negative image pair d_{neg} . The contrastive loss can be summed up as:

$$\mathcal{L}_{contrast} = \mathcal{L}_2 + \mathcal{L}_{tri} \quad (12)$$

We would like to note that, the contrastive loss can still be applied using Equ. 5 on the predicted $\tilde{w}_0 = \frac{1}{\sqrt{\alpha_t}}(w_t, \sqrt{1 - \bar{\alpha}_t}\epsilon_{\theta}(w_t, t, e))$. However, applying constraints on this \tilde{w}_0 would not be stable since it depends on both predicted ϵ and w_t in each step, where w_t is varying and sampled from Gaussian in each step.

The total loss \mathcal{L} for training our proposed NeRF-3DE is summarized as follows:

$$\mathcal{L}_{total} = \lambda_{diff} \cdot \mathcal{L}_{diff} + \lambda_{contrast} \cdot \mathcal{L}_{contrast}, \quad (13)$$

where λ_{diff} and $\lambda_{contrast}$ are the hyper-parameters used to control the weighting of the corresponding losses.

4. Experiment

4.1. Datasets

Since our pipeline relies on the pre-trained latent-based NeRF, we train the latent-based NeRF on some datasets including FFHQ [13], AFHQ [6], and CompCar [34].

FFHQ [13] is a face dataset which contains 70,000 face images. We assume the captured faces are mostly in the center. Though the size of the images is provided as 1024 × 1024 in the dataset, all of the images are resized into 256 × 256 for training the checkpoints.

AFHQ [6] is an animal face dataset that contains 15,000 high-quality images at 512×512 resolution and includes three categories of animals which are cats, dogs, and wildlife. Each category has about 5000 images. For each category, the dataset split around 500 images as a test set and provide all remaining images as a training set. Only the training images are used in the experiments.

CompCars [34] is a car dataset that contains 136,726 images capturing the different vehicles with various styles. The original dataset contains images with different aspect ratios. All of the images in the dataset are center-cropped and resized into 256 × 256.

4.2. Experimental Settings

Pre-trained latent-based NeRFs. To train our diffusion prior in our 3D-CLFusion, we use the pre-trained StyleNeRF [8] and EG3D [2] models as the generators trained on FFHQ, AFHQ, and CompCars. To have a fair comparison, we use the pre-trained models that are provided by original papers online. Specifically, StyleNeRF [8] provides checkpoints on FFHQ in 256, 512, and 1024 dimensions. EG3D [2] provides models trained on FFHQ and AFHQ (only cat category). Since StyleNeRF and EG3D do not provide the checkpoints for realistic cars such as the images in CompCars, we additionally train the checkpoint of the generator on cars using StyleNeRF for evaluation.

Baselines. Since our 3D-CLFusion is the first diffusion prior to leveraging latent-based NeRFs for text-to-3D, we compare it with the most similar baseline: clip2latent [25]. clip2latent is also a framework for w latent diffusion while the main difference is their design model is for 2D StyleGAN and does not have the constraints on view-invariant learning for NeRFs. To further compare with the direct optimization method, we compare our model with latent vector optimization in \mathcal{W} [14].

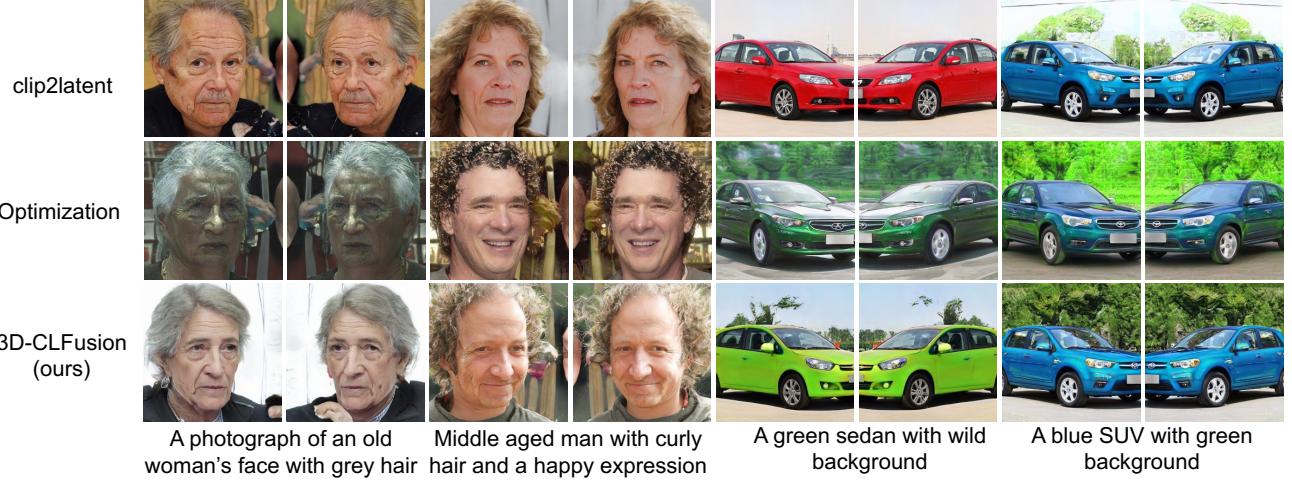


Figure 4: **Qualitative comparisons on text-to-3D with pre-trained latent-based NeRF: StyleNeRF [8]**. All of the output images are rendered using the same camera poses and the checkpoints from FFHQ and CompCars datasets.

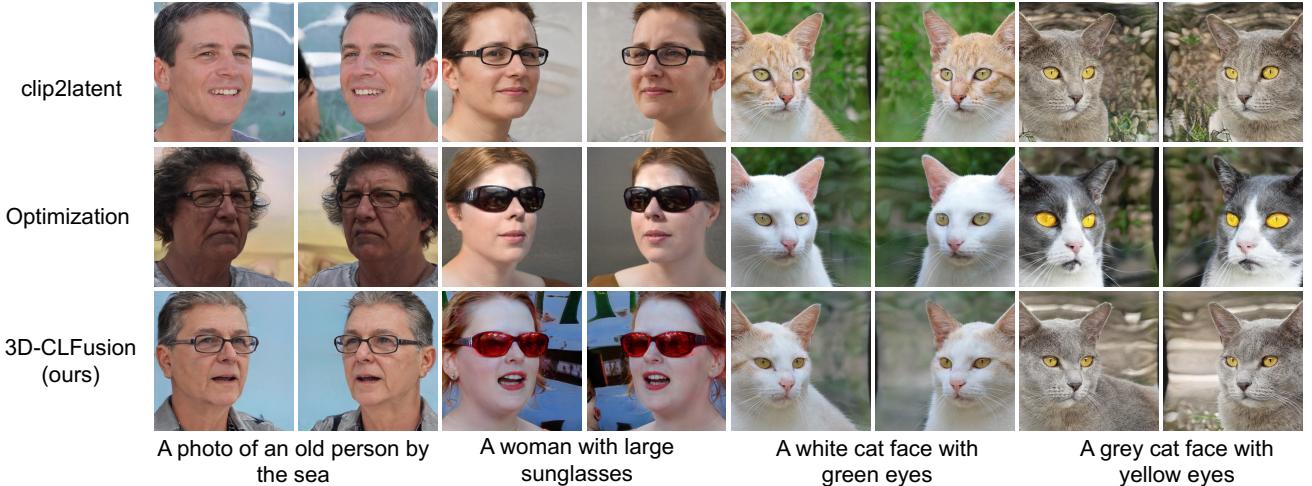


Figure 5: **Qualitative comparisons on text-to-3D with pre-trained latent-based NeRF: EG3D [2]**. All of the output images are rendered using the same camera poses and the checkpoints from FFHQ and AFHQ (Cat class) datasets.

Evaluation settings. For qualitative comparison, we compare the input text prompt and the multi-view images generated from its latent code. We would like to note that, the generators trained on FFHQ and AFHQ are trained using the frontal head yaw angle roughly ranging between -45° to 45° degrees for StyleNeRF and EG3D. Only the checkpoints on CompCars can have 360-degree rendered images. For quantitative results, we use the CLIP similarity score following clip2latent [25] and only measure the frontal view of the latent-based NeRFs generated by the models trained on FFHQ.

4.3. Implementation Details

To train our diffusion priors (either ϵ_θ or f_θ), we leverage the generated images from the generators to generate the ground-truth paired data. Specifically, we sample a latent $w \in \mathbb{R}^{512}$ from the generator and generate $k = 8$ views by manipulating camera poses for one w . Each generated image will be resized to 224×224 for the CLIP image encoder, *i.e.*, ViT-B/32 we use in the experiments, to generate the CLIP embeddings $e \in \mathbb{R}^{512}$. To have the same comparison with clip2latent [25], we also choose the same architecture as the diffusion prior in DALLE-2 [28] where the ϵ_θ or f_θ is implemented with causal transformer [18]. Following the training strategy in clip2latent, we also apply classifier-free

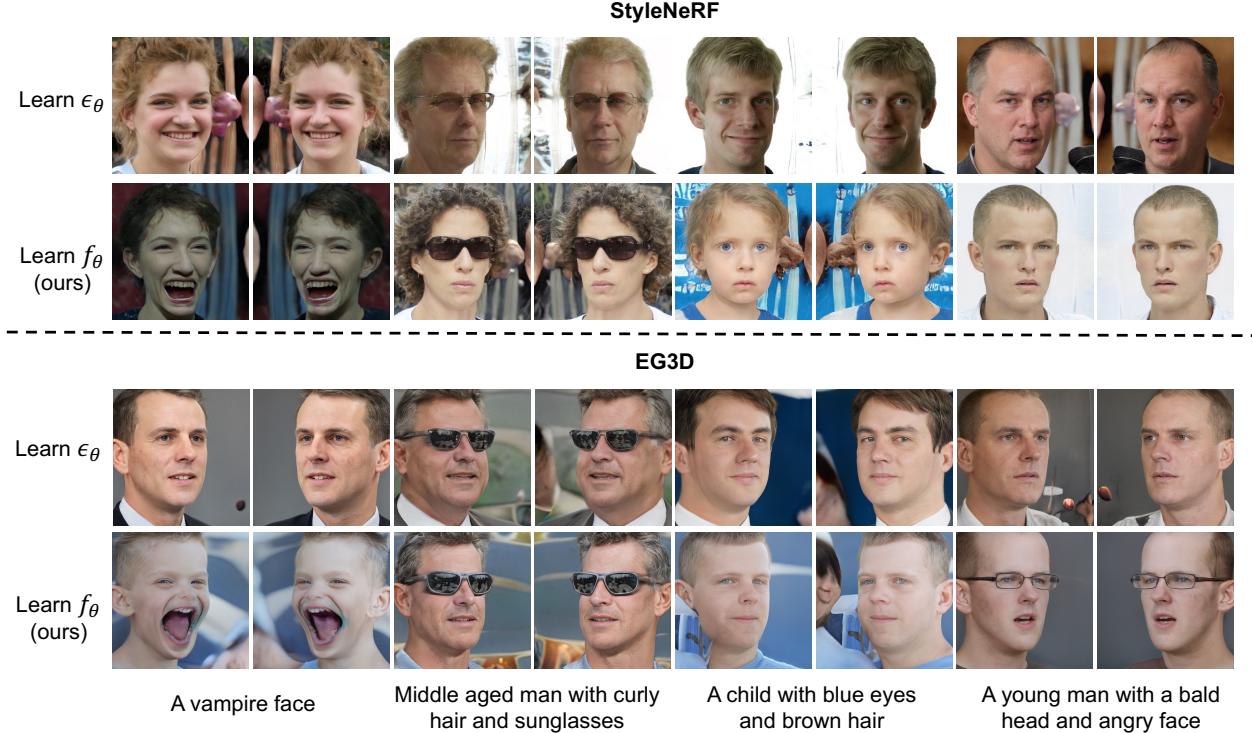


Figure 6: **Ablation studies on the learning of ϵ_θ and learning of f_θ in the diffusion prior of 3D-CLFusion.** The results are produced from the generators of both StyleNeRF and EG3D, which are trained on the FFHQ dataset.

Table 1: Quantitative results on CLIP score and the inference time. The CLIP score is measured only on the rendered frontal view for fair comparison. The results are measured on the StyleNeRF and Eg3D generators trained on FFHQ. The testing 64 prompts are from the ones provided by clip2latent [25]. The time is measured on one Nvidia 3090 GPU. The number in bold and underline indicate the best and the second-best results.

Method	CLIP score		Time
	StyleNeRF [8]	EG3D [2]	
clip2latent [25]	0.282	0.245	18.1 s
Optimization [14]	0.358	0.343	55.5 s
Ours (w/ f_θ)	<u>0.337</u>	<u>0.291</u>	18.4 s
Ours (w/ ϵ_θ)	0.287	0.254	18.2 s
Ours w/o \mathcal{L}_2	0.305	0.272	18.4 s
Ours w/o \mathcal{L}_{tri}	0.311	0.282	18.4 s

guidance [10] and pseudo-text embeddings [36] to enhance the diversity of the generated image and prevent overfitting on the image embeddings. The number of timesteps for the diffusion process is set as 1000. For the hyperparameter for all of the loss functions, all losses are equally weighted ($\lambda_{diff} = 1.0$ and $\lambda_{contrast} = 1.0$) for all the experiments.

The batch size is set as 512 where we ensure there are 64 ground-truth w latent each with 8 different CLIP image embeddings from different views for optimizing the loss. We optimize the network using Adam optimizer with a learning rate of 0.0001 and train the model for 1,000,000 iterations. Each experiment is conducted on 1 Nvidia GPU 3090 (24G) and implemented in PyTorch.

4.4. Results and Comparisons.

In this section, we compare our proposed model with the baseline approach: clip2latent [25] and the online optimization method [14] qualitatively (in Figure 4 and Figure 5) and quantitatively (in Table 1).

Qualitative results. As shown in Figure 4, we compare the quality of the generated latent w among all of the models on the generated images from StyleNeRF. The left part displays the results on the FFHQ dataset while the right part displays the results on CompCars. There are some phenomena that can be summarized. First, although clip2latent is able to generate a latent code that can generate close images, semantically the output images are not well matched. For example, the gender on the FFHQ and the color of the vehicle on CompCars are not well matched with the input text prompts. This infers that without the constraints to ensuring the latent w to be invariant, the generated latent code

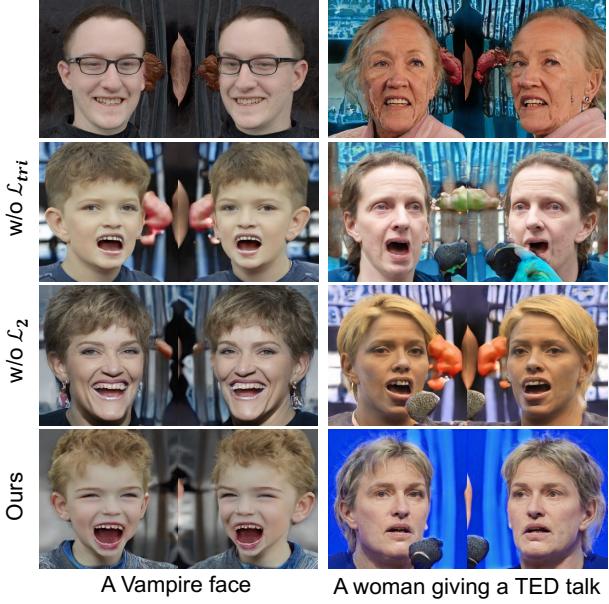


Figure 7: **Ablation studies on the proposed losses.** The results are from the StyleNeRF generator trained on FFHQ.

w from the text CLIP embeddings may not share close latent \mathcal{W} space. Second, the optimization is able to generate satisfactory results on FFHQ while failing to optimize the entire color of the vehicle. We credit the reason to that because the 360-degree vehicle is difficult to manipulate compared with faces in FFHQ. Thus, directly optimizing CLIP loss will be easy to bring artifacts and generate unnatural results (see the rightmost column of the blue SUV with green background).

We can also observe similar observations on EG3D [2] shown in Figure 5. First, clip2latent fails to generate the well-matched latent for EG3D to render (see the sunglasses in the second big column and the color of the cat in the third big column). The optimization method is also able to generate well-matched results while they are unnatural (see the yellow eyes in the rightmost column). Our proposed methods exhibit comparable results as direct optimization.

Quantitative results. As shown in the first three rows of Table 1, we compare the two methods with the CLIP similarity score using the 64 text prompts provided by clip2latent [25]. To measure the inference time and have a fair comparison, we train the optimization [14] for 200 iterations for each text prompt. Our proposed model achieves a superior score compared with clip2latent with a similar inference time. On the other hand, although the optimization method has artifacts in qualitative results, it has the best result in CLIP score. This is basically because the objective of the optimization is sorely CLIP similarity. Thus, online optimization is supposed to have the best score.

4.5. Ablation studies

Learning objectives: ϵ_θ vs f_θ . To further support the claim in Section 3 that learning f_θ is more stable than ϵ_θ when applying contrastive loss, we compare the models learning two different objectives qualitatively in Figure 6 and quantitatively in Table 1. As shown in Figure 6, it is obvious that learning ϵ_θ would lead to inferior results. For example, learning ϵ_θ can not generate good latent when feeding the input text prompt as "A vampire's face" in either StyleNeRF or EG3D. Table 1 also shows that learning ϵ_θ exhibits a worse CLIP score compared with f_θ . We credit the reason that by producing the \bar{w}_0 from ϵ_θ , the \bar{w}_0 will be far from consistency since it also depends on w_t in each step. The view-invariant learning will not work easily on the network in ϵ_θ .

Contrative loss function: $\mathcal{L}_{contrast}$. To further analyze the effectiveness of essential contrastive losses of our proposed 3D-CLFusion, we conduct the experiments with one of them excluded and present the qualitative result in Figure 7 and quantitatively in Table 1 as well. First, apparently without L2 loss \mathcal{L}_2 , the diffusion model is not able to well derive view-invariant latent code in each diffusion process while it can still ensure the multi-view CLIP embeddings will produce close w_0 with the remaining triple loss. Second, we can observe that in the model with the triplet loss \mathcal{L}_{tri} in our $\mathcal{L}_{contrast}$ excluded, the produced latent code w starts to semantically deviate the input text prompt, which infers that the inter-class distance learning is also important for learning the view-invariant latent code in \mathcal{W} latent space. Third, with the entire contrastive loss $\mathcal{L}_{contrast}$ excluded, the model is not able to have any guidance on view-invariant learning and would produce similar results as clip2latent [25].

5. Conclusion

We have unveiled the challenges of the task text-to-3D directly using latent-based NeRFs and the limitations of the current models for this task. We propose a framework named 3D-CLFusion, which aims to produce view-invariant latent for rendering 3D objects with NeRFs from the input text prompt. Compared with the existing baseline models in the experiments, our designed model achieves more effective text-to-3D with pre-trained NeRFs. Though the 3D object created by our model is domain-specific to the pre-trained model, the inference time is at least 100 times faster than the existing text-to-3D creation from neural rendering. We would like to note that, our model will be applicable in the real world if more pre-trained NeRFs with vast categories of 3D objects are available.

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. 2
- [2] 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 *CVPR*, pages 16123–16133, 2022. 1, 2, 3, 5, 6, 7, 8
- [3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 3
- [4] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)*, 41(1):1–26, 2022. 3
- [5] Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Sem2nerf: Converting single-view semantic masks to neural radiance fields. *ECCV*, 2022. 3
- [6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 5
- [7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2
- [8] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *ICLR*, 2021. 1, 2, 3, 5, 6, 7
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 4
- [10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7
- [11] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 1, 3
- [12] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 3
- [13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 2, 3, 5
- [14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 2, 3, 5, 7, 8
- [15] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 895–904, 2022. 2
- [16] Yu-Jhe Li, Tao Xu, Bichen Wu, Ningyuan Zheng, Xiaoliang Dai, Albert Pumarola, Peizhao Zhang, Peter Vajda, and Kris Kitani. 3d-aware encoding for style-based neural radiance fields. *arXiv preprint arXiv:2211.06583*, 2022. 2, 4
- [17] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. 1, 2, 3
- [18] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. *arXiv preprint arXiv:2204.07258*, 2022. 6
- [19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020. 1, 2, 3
- [20] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2
- [21] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 3
- [22] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13503–13513, 2022. 3
- [23] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. *NeurIPS*, 34:20002–20013, 2021. 3
- [24] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021. 2
- [25] Justin NM Pinkney and Chuan Li. clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip. *BMVC*, 2022. 1, 2, 3, 5, 6, 7, 8
- [26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2, 3
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3, 6
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3
- [30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2, 3

- [31] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 33:20154–20166, 2020. 3
- [32] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020. 3
- [33] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *CVPR*, pages 18430–18439, 2022. 3
- [34] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015. 5
- [35] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multi-plane images: Making a 2d gan 3d-aware. In *ECCV*, 2022. 3
- [36] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021. 7

A. More results on multi-modal manipulation

Since our 3d-CLFusion takes the input text embedding and the input gaussian noise as the starting latent for diffusion, we show examples of manipulating the random initial gaussian latent given the same input text prompt.

Generator employing StyleNeRF The results on FFHQ are presented in Fig 8 and the results on CompCars are presented in Fig 9.

Generator employing Eg3D The results on FFHQ are presented in Fig 10 and the results on AFHQ cats are presented in Fig 11.

Young man with black curly hair and sunglasses



Young woman with brown straight hair and a smiling face



Figure 8: Results on StyleNeRF generator trained on FFHQ. Each row indicates the same input text with different input sampled noise.

A red car



A blue car



Figure 9: Results on StyleNeRF generator trained on CompCars. Each row indicates the same input text with different input sampled noise.

An old man with gray hair and smiling face



A woman with glasses and a disappointing face



Figure 10: Results on Eg3D generator trained on FFHQ. Each row indicates the same input text with different input sampled noise.

A black cat with green eyes



A brown cat with closed eyes



Figure 11: Results on Eg3D generator trained on AFHQ cats. Each row indicates the same input text with different input sampled noise.