# Final Report: Enhancing 3D Character Generation with ContolNet and LoRA

Congrong Xu[1,2], Zhanhe Shi[1,2], Minshen Zhang[1,2], and Qingcheng Zhao[1,2]

[1] *UC Berkeley*
[2] *ShanghaiTech University*

December 28, 2023

**Abstract**

In the rapidly advancing domain of digital 3D content creation, the demand for efficient and sophisticated generation tools is increasingly crucial. This paper presents an innovative solution to augment 3D character generation by seamlessly integrating ControlNet and Low-Rank Adaptation (LoRA) into pre-existing text-to-image diffusion models. Traditional systems often grapple with issues such as lack of spatial consistency and the occurrence of multi-headed artifacts due to poor quality in multi-view image synthesis.

Our approach leverages ControlNet for refined pose control and adapts 3D Gaussian Splatting for effective spatial optimization and pruning. In addition, we utilize LoRA for the fine-tuning of pre-trained text-to-3D models, facilitating the creation of personalized and high-fidelity 3D characters that meet specific user requirements. A notable enhancement in our methodology is the application of Noise-Free Score Distillation (NFSD), which significantly elevates model performance at reduced CFG scales. This strategy enables the production of detailed, high-resolution 3D avatars from textual descriptions, while assuring feature consistency across diverse views.

To validate the effectiveness of our proposed method, we carried out comprehensive ablation studies and user evaluations. These assessments involved comparing our approach with existing baselines to showcase its superiority in generating photo-realistic 3D models that accurately reflect user inputs. Our research represents a significant advancement in AI-assisted 3D character generation, opening new avenues in industries such as gaming, animation, and virtual reality. It contributes a notable innovation to the burgeoning field of text-to-3D transformation.

# 1 Introduction

## 1.1 Background

The evolution of digital content creation, particularly in the 3D domain, is pivotal for industries like gaming, advertising, films, and the burgeoning MetaVerse. Traditionally, creating intricate 3D models has been both time-consuming and resource-intensive, demanding thousands of hours of work from skilled artists. This scenario underscores the need for innovative approaches that reduce manual labor while enabling both professionals and amateurs to produce 3D assets efficiently.

Recent breakthroughs in 2D content generation (Rombach et al., 2022) have sparked significant advancements in 3D content creation. These advancements can be broadly categorized into two streams: inference-only 3D native methods and optimization-based 2D lifting methods. While 3D native methods (e.g., Jun Nichol, 2023; Nichol et al., 2022; Gupta et al., 2023) show promise in swiftly generating 3D-consistent assets, they are hindered by the need for extensive 3D dataset training, which is labor-intensive and often lacks diversity and realism (Deitke et al., 2023b; a; Wu et al., 2023).

One notable method, Dreamfusion (Poole et al., 2022), uses Score Distillation Sampling (SDS) to circumvent the 3D data scarcity, inspiring the development of 2D lifting methods (Lin et al., 2023; Wang et al., 2023b; Chen et al., 2023b). Despite progress, these methods suffer from long optimization times due to the computationally intensive Neural Radiance Fields (NeRF) (Mildenhall et al., 2020), rendering them impractical for large-scale deployment. Furthermore, existing methods to accelerate

NeRF, such as occupancy pruning (Muller et al., 2022; Sara Fridovich-Keil and Alex Yu et al., 2022), are less effective in generative settings, especially when supervised by the ambiguous SDS loss.

### 1.2 Objectives

In response to these challenges, we propose a novel approach to personalize text-to-3D diffusion models for user-specific 3D generation needs. Our method aims to enrich the model's language-vision dictionary, allowing it to associate new words with specific subjects as defined by the user. This embedded dictionary enables the model to generate photo-realistic 3D models of subjects in various scenes and conditions, maintaining their distinctive features

We leverage ControlNet, an architecture that introduces spatially localized input conditions to pretrained text-to-image diffusion models. This integration allows for precise pose control in the generated 3D models. Additionally, we fine-tune the text-to-3D model with input images and text prompts that combine a unique identifier with the subject's class name (e.g., "A Darth Vader [V]"). This approach enables the model to apply its pre-existing knowledge of the subject class, tailored by the specific instance linked with the unique identifier.

We adapt 3D Gaussian Splatting (Kerbl et al., 2023) for generative settings. This approach, in contrast to NeRF-based methods, efficiently prunes empty space and simplifies the optimization landscape. The progressive densification of Gaussian splatting aligns with the generative settings' optimization progress, enhancing generation efficiency.

### 1.3 Significance

Our approach opens avenues for a variety of text-based 3D generation applications, including subject recontextualization, property modification, original art renditions, and more. To demonstrate the efficacy and versatility of our method, we conduct ablation studies, comparing our approach with alternative baselines and related work. Our method outperforms existing approaches in both quality and visual appeal. Additionally, we carry out a user study to evaluate the fidelity of subjects and prompts in our synthesized images, positioning our method as a significant advancement in the field of 3D content creation.

## 2 Related Work

### 2.1 Fine-Tuning Neural Networks

Fine-tuning neural networks is a pivotal process in adapting pretrained models to specific tasks. Traditional fine-tuning methods, which involve additional training with new data, often confront challenges such as overfitting, mode collapse, and catastrophic forgetting. To circumvent these issues, adapter methods have been introduced, particularly in NLP, for customizing pretrained transformer models. These adapters, embedded as new module layers, have shown promising results in various domains, including computer vision for tasks like incremental learning and domain adaptation.

Adapter variants, such as those designed by Houlsby et al. (2019) and Lin et al. (2020), offer different configurations in the Transformer blocks, balancing between efficiency and performance. Despite their compact design, adapters introduce additional computational load, especially in scenarios lacking model parallelism. This computational overhead becomes significant in online inference settings with small batch sizes. Our work integrates Low-Rank Adaptation (LoRA), which innovatively addresses the issue of catastrophic forgetting and computational efficiency by learning parameter offsets with low rank matrices.

### 2.2 3D Representations

Neural Radiance Fields (NeRF) [1] employs a volumetric rendering and has been popular for enabling 3D optimization with only 2D supervision. Although NeRF has become widely used in both 3D reconstruction [2][3][4][5][6][7][8], optimizing NeRF can be time-consuming. Various attempts have been made to accelerate the training of NeRF [11][12], but these works only focus on the reconstruction setting. The common technique of spatial pruning fails to accelerate the generation setting. Recently, 3D Gaussian splatting [9] has been proposed as an alternative 3D representation to NeRF, which

has demonstrated impressive quality and speed in 3D reconstruction [10]. The efficient differentiable rendering implementation and model design enables fast training without relying on spatial pruning. In this work, we for the first time adapt 3D Gaussian splatting into generation tasks to release the potential of optimization-based methods.

### 2.3 Text-To-2D Generation

The evolution of image diffusion models, originating from Sohl-Dickstein et al., has revolutionized image generation. Latent Diffusion Models (LDM) stand out by performing diffusion in the latent space, thus reducing computational demands. Text-to-image models like Glide and Stable Diffusion harness pretrained language models (e.g., CLIP) to encode textual inputs into latent vectors, achieving remarkable image generation results. Our work builds on these advancements, tailoring them for 3D character generation.

### 2.4 Text-To-3D Generation

The objective of Text-to-3D generation is to create three-dimensional assets based on text prompts. Recent advancements in 2D diffusion models have significantly impacted text-to-image generation. Yet, adapting these models for 3D generation presents substantial challenges, notably in assembling extensive 3D datasets. Traditional 3D diffusion models are typically restricted to a single object category, leading to a lack of variety. To facilitate the creation of diverse 3D content, several techniques have been developed to adapt 2D image models for 3D generation. These approaches involve fine-tuning a 3D model to align with the probabilities of pretrained 2D diffusion models when viewed from various angles, ensuring both three-dimensional consistency and realism. Subsequent research has focused on improving aspects like the fidelity of generation and the stability of training. Nonetheless, these methods based on 2D model adaptation often face lengthy optimization times for each case. Specifically, using Neural Radiance Fields (NeRF) for 3D representation results in high computational costs in both the rendering passes. In this study, we opt for 3D Gaussians as our differentiable 3D representation, demonstrating through empirical evidence its more streamlined optimization process.

### 2.5 Text-to-3D Character Generation

Adapting 2D diffusion models for 3D generation involves significant challenges, primarily due to the complexity of assembling comprehensive 3D datasets and ensuring multi-view consistency. Traditional 3D models, often limited to singular object categories, lack diversity. To address this, our method fine-tunes a 3D model to align with the probabilities of pretrained 2D diffusion models from various perspectives, ensuring both three-dimensional consistency and realism. This approach significantly improves upon the fidelity and stability of traditional 3D generation methods.

Initiatives like Avatar-CLIP have explored the realm of 3D avatar generation, employing CLIP for shape sculpting and texture generation. However, these methods often result in oversimplified models. In contrast, DreamAvatar and AvatarCraft represent concurrent advancements in the field. While DreamAvatar generates static posed avatars, AvatarCraft excels in producing high-quality, animatable avatars through a combination of coarse-to-fine training and multi-box techniques, utilizing SMPL models for shape prior and local transformations.

## 3 Methods

### 3.1 Architecture

#### 3.1.1 Overview

The architecture of our proposed system integrates three cutting-edge technologies: Stable Diffusion, ControlNet, and LoRA (Low-Rank Adaptation). This integration aims to generate high-resolution, detailed 3D avatars from textual descriptions. The process involves fine-tuning the Stable Diffusion model using LoRA for personalized character generation, employing ControlNet for ensuring multi-view consistency, and leveraging DreamGaussian for 3D model generation. Our whole pipeline is shown in 1.

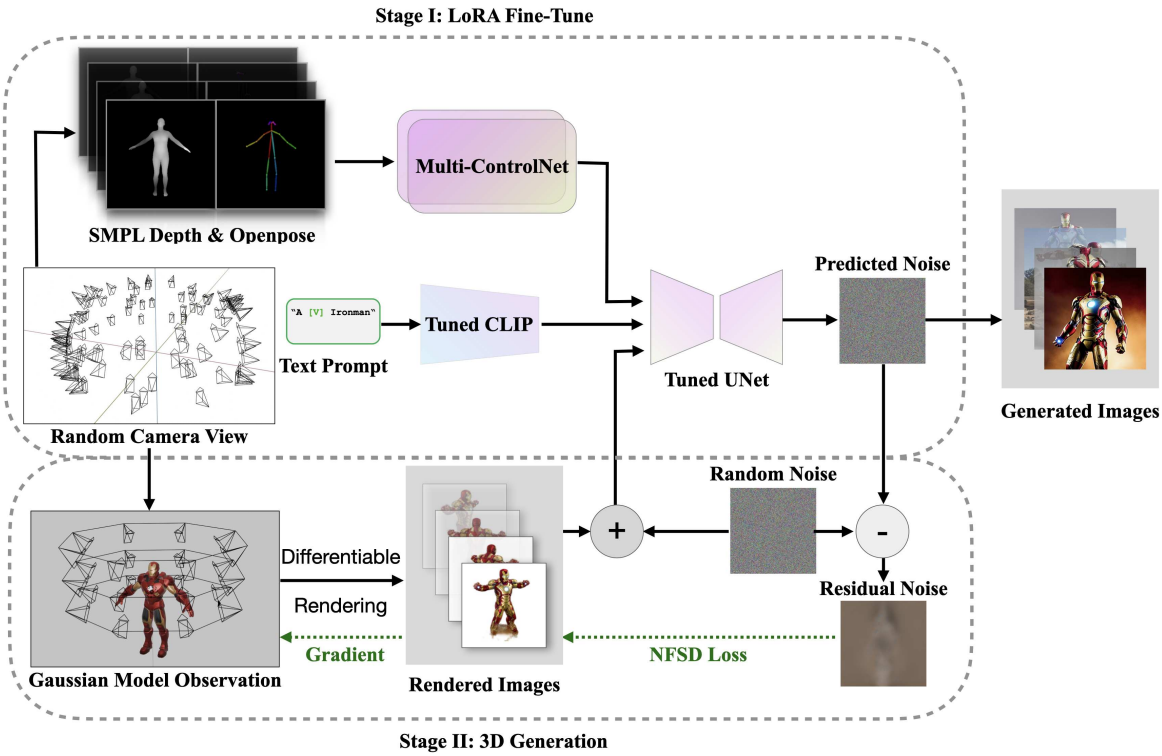


Figure 1: Our Two-Stage Training Methodology

### 3.1.2 Stable Diffusion Fine-Tuned with LoRA

Stable Diffusion serves as the backbone of our architecture, renowned for its capability to generate detailed images from textual inputs. To tailor this model for personalized avatar creation, we employ LoRA, a fine-tuning technique that adjusts only a small fraction of the model's parameters. This approach enables the model to maintain its general capabilities while becoming specialized in generating specific character features as described in the input text. LoRA's low-rank matrix adaptation ensures that the personalization is efficient and does not require extensive retraining of the model.

### 3.1.3 ControlNet for Multi-View Consistency

ControlNet is integrated into the pipeline to address the challenge of multi-view consistency in image generation. This module ensures that the generated images of the character from various angles are consistent in terms of appearance and pose. More specifically, we use FrankMocap to estimate 3D body poses and mesh from real images. These are then used to render 2D poses in randomly sampled viewpoints along with the depth map as illustrated in 2. Afterward, we send the projected pose and depth map to Multi-ControlNet. This guides the Stable Diffusion output to preserve feature consistency across various views, ensuring more accurate and consistent rendering of the images. This method effectively facilitates the creation of images that cohere when compiled into a 3D model.

### 3.1.4 Generative Gaussian Splatting

The final step in the pipeline is the transformation of the multi-view images into a 3D model. For this, we utilize DreamGaussian, an algorithm capable of synthesizing these images into a cohesive 3D avatar. DreamGaussian works by interpreting the spatial relations and depth cues from the set of images, thereby constructing a high-resolution, detailed 3D model that faithfully represents the character described in the input text.

Figure 2: Projected pose and depth map in front view

### 3.2 3D Representations

The foundation of our method lies in representing 3D information through a set of 3D Gaussians, as proposed by Kerbl et al. (2023). Each Gaussian is defined by its center ($\mathbf{x} \in \mathbb{R}^3$), scaling factor ($\mathbf{s} \in \mathbb{R}^3$), rotation quaternion ($\mathbf{q} \in \mathbb{R}^4$), opacity value ($\alpha \in \mathbb{R}$), and color feature ($\mathbf{c} \in \mathbb{R}^3$). Collectively, these parameters are denoted as $\Theta$, where $\Theta_i = \{\mathbf{x}_i, \mathbf{s}_i, \mathbf{q}_i, \alpha_i, \mathbf{c}_i\}$ represents the parameters of the $i$-th Gaussian. Our rendering process projects these 3D Gaussians onto a 2D plane, facilitating volumetric rendering in a front-to-back depth order.

### 3.3 Text-to-3D Generation

The core of our optimization process involves Score Distillation Sampling (SDS), a technique that minimizes the KL divergence between Gaussian distributions and the learned score functions of a pre-trained diffusion model. We initialize the 3D Gaussians with random positions and periodically densify them, aligning with the generation progress. The SDS optimizes these 3D Gaussians by sampling random camera poses and rendering the RGB and transparency of the current view. This process is guided by different 2D diffusion priors ($\phi$), which is back-propagated to optimize the 3D Gaussians.

In our text-to-3D task, a single text prompt is inputted and converted into CLIP embeddings ($e$). The SDS loss is formulated as:

$$\nabla_\Theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,p,\epsilon} \left[ \left( \epsilon_\phi \left( I^p_{\text{RGB}}; t, e \right) - \epsilon \right) \frac{\partial I^p_{\text{RGB}}}{\partial \Theta} \right]$$

This loss function effectively guides the optimization of the 3D Gaussians towards generating a 3D representation that is coherent with the textual description.

### 3.4 Noise-Free Score Distillation

First, consider the difference $\delta_C = \varepsilon_\phi(Z_t; y, t) - \varepsilon_\phi(Z_t; \mathcal{X}, t)$ in SDS loss. While $\varepsilon_\phi(Z_t; y, t)$ ideally points towards a local maximum in the probability density of noisy real images conditioned on $y$, $\varepsilon_\phi(Z_t; \mathcal{X}, t)$ points towards a denser region in the distribution of unconditioned noisy images. Thus, the difference $\delta_C$ between the two predictions may be thought of as the direction that steers the generated image towards alignment with the condition $y$, and we henceforth refer to it as the *condition direction*.

NFSD [**?**] focuses on isolating the distortion-related component ($\delta_{\text{D}}$) from the predicted noise in the diffusion process. We distinguish between smaller ($t < 200$) and larger ($t \geq 200$) timestep values. For smaller timesteps, the noise component ($\delta_{\text{N}}$) is negligible, and the score primarily consists of $\delta_{\text{D}}$ which is just original SDS. For larger timesteps, we approximate $\delta_{\text{D}}$ as the difference between the predicted noise under null-condition and a negative-condition prompt (described as "unrealistic, blurry, low quality," etc.).

We can reform the original SDS loss as:

$$\nabla_\theta L_{SDS} = w(t)(\varepsilon_\phi(Z_t; y, t) - \epsilon)\frac{\partial \chi}{\partial \theta} = w(t)(\delta_D + \delta_N + s\delta_C - \epsilon)\frac{\partial \chi}{\partial \theta}. \tag{1}$$

Figure 3: Ours LoRA v.s. Stable Diffusion (back view)

After introducing the assumption that $\delta_C = p_{\text{neg}} \approx -\delta_D$, and thus $\varepsilon_\phi(Z_t; \mathcal{X}, t) - \varepsilon_\phi(Z_t; y) = p_{\text{neg}} = \delta_D + \delta_N - (\delta_D + \delta_N + \delta_C = p_{\text{neg}}) \approx \delta_D$. The NFSD loss is formulated as:

$$\nabla_\theta \mathcal{L}_{\text{NFSD}} = w(t) \left(\delta_{\text{D}} + s\delta_{\text{C}}\right) \frac{\partial \mathbf{x}}{\partial \theta}$$

This loss function enables the efficient optimization of the 3D Gaussians, leading to improved image and NeRF quality without the need for a large scaling factor $s$, as in SDS.

# 4 Experiments and Analysis

## 4.1 Implementation Details

### 4.1.1 Base Model and Training Approach:

Our project utilized Stable Diffusion version 1.5 as the foundational model for generating 3D character models. To enhance its capabilities, we applied LoRA (Low-Rank Adaptation) training, which resulted in the development of two distinct versions of the tuned model, each optimized for different aspects of character generation.

**Darth Vader: First Version - Initial Approach:**

- *Data Collection:* The initial dataset comprised 45 images of Darth Vader, incorporating sources from the internet, films, and images generated from 3D models.
- *Automated Labeling:* We employed a pretrained ConvNext model to automate the labeling process for these images.
- *Training Duration and Observations:* The training spanned across 10 epochs. We set the LoRA rank=32 and learning rate=1e-4 with AdamW optimizer. This version successfully captured key character features but exhibited limitations in accurately rendering the character's back. 3

**Darth Vader: Second Version - Enhanced Refinement:**

- *Dataset Optimization:* We expanded the dataset to 65 images, rigorously filtering out low-quality and blurry images from the initial collection. This revision also included a greater variety of images, particularly focusing on the character's side and back views.
- *Manual Labeling for 3D Preparation:* Each image was meticulously labeled with 'front', 'side', and 'back' view tags to aid in accurate 3D content generation.
- *Training Duration and Observations:* The training spanned across 20 epochs. We set the LoRA rank=64 and learning rate=1e-4 with AdamW optimizer. This revised model demonstrated superior adherence to specified viewing angles and significantly improved the accuracy of detail representation, especially on the character's back in 2D image tests.

**Iron Man**

- We also trained an iron man model in the same method as the second version of Darth Vader. The only difference is that for the character of Iron Man, we used only 36 images, with just 3 to 4 images taken from the side view. It indeed leads to some problems in our experiments. Yet, judging from the results, the performance of this model was still impressively satisfactory.

Figure 4: 3D generation process with training step: 500, 1500, 3000, 5500.

#### 4.1.2 Optimizing LoRA Weights and Model Evaluation

The optimization of LoRA weights proved crucial in balancing detail reproduction and command adherence. The optimal performance was achieved with a LoRA weight setting of 0.7. To assess the quality of our models, both versions were evaluated using FID (Fréchet Inception Distance) and CLIP scores. These assessments were conducted at various epochs to track the progress and quality of the models, providing valuable insights into their performance over time.

#### 4.1.3 ControlNet 3D Generation Guidance

An integral part of our 3D character generation process involved the innovative use of ControlNet, specifically tailored for guiding the generation of 3D characters in various poses. We began by generating 3d openpose from 2d images and storing a comprehensive library of 3D keypoints representing multiple human body postures. During the training phase, these keypoints were projected onto the current camera view to create OpenPose posture models. This projection was instrumental in accurately capturing the dynamics of human posture from various angles. We also render depth map from SMPL (a Skinned Multi-Person Linear Model) corresponding to the 3D Openpose, to enhance the correctness of our ControlNet.

To infuse the posture information into the generation process, we utilized a pretrained OpenPose ControlNet and a pretrained Depth ControlNet. The ControlNets were adept at injecting the current viewpoint's posture information into the U-Net architecture, a crucial step in our workflow. By employing this method, we were able to guide the generation of 3D characters in specific actions, tailoring each character to our desired pose and orientation.

A significant advantage of this approach was its effectiveness in preventing the issue of multi-heads or repeated features – a common challenge in 3D generative models. The targeted injection of posture data ensured that the generated characters remained coherent and true to the intended pose, contributing greatly to the realism and accuracy of our 3D models.

#### 4.1.4 Final 3D Content Generation Process

In the final phase of our project, we integrated the trained LoRA and SD models for comprehensive 3D content generation. This integration was crucial in realizing our project's objectives. A key feature of our approach was the dynamic adjustment of prompts in the generation process, similar to the technique used in DreamFusion. These prompts were adjusted according to the horizontal angle of view, ensuring that the generated content was aligned with the intended perspective.

Additionally, we implemented Noise-Free Score Distillation (NFSD) to enhance our models' performance. Unlike the standard SDS loss, which required a CFG scale over 50 and often led to oversaturation, NFSD maintained effective performance at a CFG scale of just 7.5. This advancement significantly improved the quality of our models.

Each model underwent a rigorous training process, consisting of 5500 iterations on an Nvidia 3070 GPU. The first 3500 iterations included a densification of Gaussian every 350 iterations.The generation processes are shown in 4.

To further enhance the level of detail in the models, we employed a stepwise rendering process. The initial 30% of steps utilized a resolution of 128x128, followed by 30% at 256x256, and the final 40% at

a higher resolution of 512x512. This stepwise increase in resolution allowed for a gradual build-up of details, resulting in models with a higher degree of finesse and realism.

## 4.2 Prior Analysis

In this part, we will try different combination for our controlnet, including baseline(no controlnet), openpose controlnet, depth controlnet and openpose&depth multicontrolnet, we will conduct experiment on those combinations and analyze their performance in generating 3D models.
Note: For Iron Man, our prompt remained "Mark 42, Iron Man, full body, masterpiece." For Darth Vader, our prompt was kept as "1boy, full body, Darth Vader, 3D asset, 4K, ultra quality, realistic." These are obtained from our 2D Stable Diffusion tests.

### 4.2.1 Baseline (Stable Diffusion + LoRA)

In our initial approach, we trained a baseline model utilizing Stable Diffusion and LoRA, devoid of any active control net intervention. The outcome, as illustrated in Figure 5, distinctly reveals that in the absence of a control net, the model struggles to produce a 3D model with an accurate pose. While the model is capable of generating a 3D model with the correct shape, the resulting figures appear blurry and indistinct.

Notably, when tasked with generating Iron Man, the model produced a 3D model that exceeded the image boundaries even when we stressed the prompt "full body", an outcome that is considered unacceptable for our purposes.

### 4.2.2 OpenPose ControlNet Only

For our second model, we utilized Stable Diffusion and LoRA, incorporating OpenPose as a guiding prior for the generation process. The methodology for integrating the OpenPose prior through the ControlNet has been detailed in the preceding section. Observations from Figure 6 demonstrate that the implementation of the ControlNet is indeed crucial; the model successfully generates a 3D model with an accurate pose. However, despite the correct pose, the model appears noticeably bulkier than anticipated. This discrepancy may stem from an imbalanced distribution of side-view images of the 3D model within our dataset.

Additionally, we noted that during the training phase, the Gaussian splatting balls tended to increase in size and solidity as the training progressed. This phenomenon is likely a contributing factor to the "bulkiness" observed in the generated 3D model.

To address this issue, we introduced a depth prior into the model. It is our expectation that this addition will help constrain the model, enabling it to generate a 3D model that avoids the "bulkiness" problem.

### 4.2.3 Depth ControlNet Only

Now we tried to use ControlNet on the model with only depth prior, instead of using the OpenPose image. From the result shown in Figure 7(a), it is evident that using depth alone to train the model leads to some undesirable outcomes. Notably, some parts of the hands and legs appear to be transparent, indicating a significant loss of detail in these areas. This results in a model that looks rather unsatisfactory. The depth prior seems unable to fully capture the complexity of the human body, leading to these inaccuracies.

### 4.2.4 OpenPose + Depth ControlNets

When we combined both OpenPose and Depth Prior, the results were quite promising. As seen in Figure 7(b), the model was able to generate a 3D model with the correct pose and shape. While there was a slight blur compared to the model with only the OpenPose prior, it was minimal and could potentially be reduced with further training and refinement.

Interestingly, the combined use of OpenPose and Depth Prior seemed to balance each other's limitations. The model was not excessively bulky, as was the case with the OpenPose prior alone, nor did it produce incorrect poses, as was observed with the Depth Prior alone. This suggests that the combination of these two priors effectively harnessed their strengths while mitigating their weaknesses.

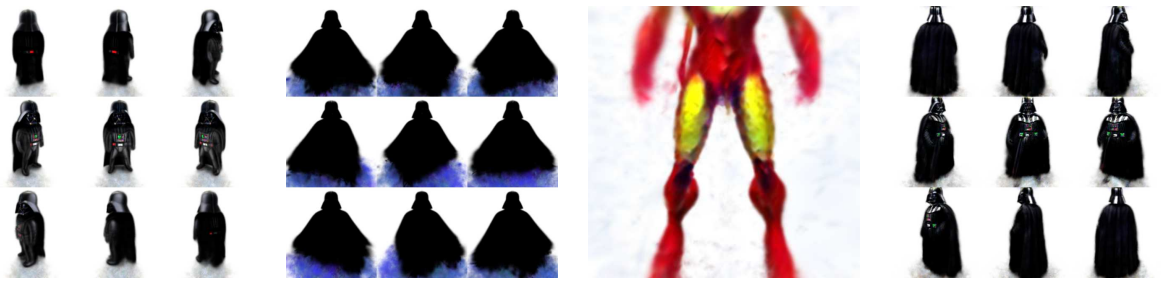

Figure 5: Baseline Model - The first three images on the left depict results without the use of LoRA, while the fourth image demonstrates the outcome when LoRA is enabled.

Figure 6: OpenPose ControlNet

We attempted to directly incorporate depth images as a part of the loss function, but this approach disrupted the model's clothing, particularly the cloaks. Using a depth Controlnet with lower control strength did not cause this disruption and was able to assist in the correct generation of character hands to some extent. However, we found that it was necessary to periodically reset the transparency of the Gaussian spheres to better adhere to the depth directives. This requirement led to the need for longer iteration steps for the model to materialize effectively.

In conclusion, the combination of OpenPose and Depth Prior showed great potential for generating high-quality 3D models. With further refinement and exploration of new strategies, we believe that we can achieve even better results in future work.

#### 4.2.5 Summary

In our recent experiment, we focused on enhancing the quality of 3D models generated by our model, which integrates Stable Diffusion and LoRA techniques. We specifically investigated the impact of various priors, including OpenPose and depth prior, on the model's generative process.

Our initial observations with the baseline model, which operated without the application of any control net, revealed some limitations. This model predominantly produced 3D models with inaccuracies in pose, leading to blurry and indistinct outputs. The integration of the OpenPose prior marked a significant improvement, particularly in the accuracy of poses. However, this enhancement came with an unexpected consequence—the models tended to exhibit an increased bulkiness.

To address this issue, we introduced a depth prior. When utilized independently, the depth prior did not yield the desired outcomes; the model particularly struggled with accurately generating leg poses. However, a notable improvement was observed when both the OpenPose and depth priors were applied in tandem. This combination enabled the generation of 3D models with both accurate poses and appropriate shapes.

Despite this success, a slight drawback was noted—the models appeared somewhat blurry and less defined compared to those generated using only the OpenPose prior. We hypothesize that with further training, these issues could be mitigated. However, due to the constraints imposed by our current hardware setup, specifically a single RTX 3070 GPU, our experimentation was limited in this regard.

((a)) Depth ControlNet ((b)) Depth and OpenPose MultiControlNet

Figure 7: Comparison of Depth ControlNet and Depth and OpenPose MultiControlNet

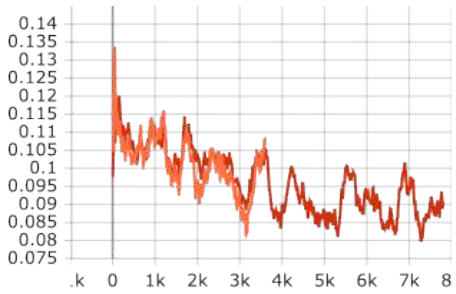


Figure 8: Training Loss by Step (Darth Vader: orange for version 1 and red for version 2)

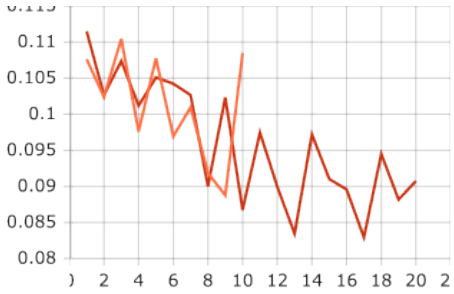


Figure 9: Training Loss by Epoch (Darth Vader: orange for version 1 and red for version 2)

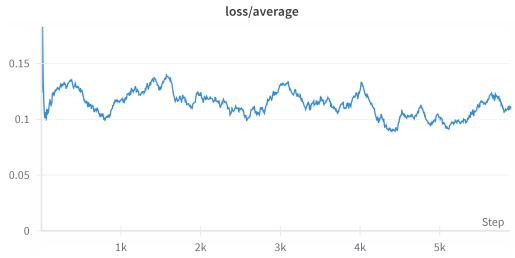


Figure 10: Training Loss by Step (Iron Man)

## 4.3 Training and Validation

Training loss are shown in 8 9 10. Notably, the training process of the Lora model is not monotonically decreasing, but overall, it shows a downward trend. There is also a certain correlation between the model's loss, CLIP score, and FID. Models with lower loss are better at generating scenes semantically related to the characters, demonstrating higher compatibility and image quality.

Figure 11 displays the changes in NFSD loss during the generation of 3D models. Since all the training losses looks almost the same, we are not visualizing all the losses. During the generation process, the loss initially increases and then decreases. We believe that the loss in the first half is mainly dominated by the weight $w_t$, where the training primarily focuses on the shape generation of the 3D model, and the residual term of the diffusion model is not significant. In the second half, the residual term of the diffusion model becomes dominant. During this phase, the 3D model refines details in the direction guided by the prompt, and the loss gradually decreases.

## 4.4 Parameter Analysis

In image synthesis, the Fréchet Inception Distance (FID) and the CLIP scores are two critical metrics. For our models at different stages, we generated images using prompts identical to those in the training dataset and assessed the FID and CLIP scores to gauge their quality and semantic alignment. The

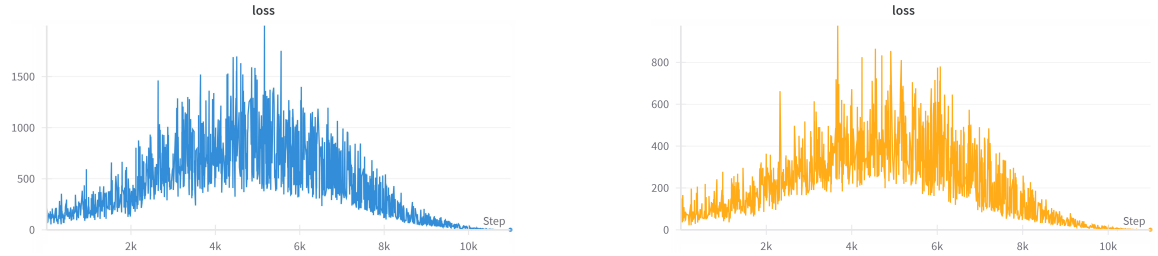


Figure 11: Examples of NFSD loss

| Training Epoch | FID Score (↓) | CLIP Score (↑) |
|---|---|---|
| w/o LoRA | 164.81 | 32.59 |
| v1-8 | 140.00 | 32.27 |
| v1-final | 145.22 | 32.32 |
| v2-16 | 139.74 | 32.10 |
| v2-18 | 146.24 | 32.01 |
| v2-final | **139.37** | **32.40** |

Table 1: Evolution of FID and CLIP Scores during Training

significant decrease in the FID score, coupled with the nearly consistent CLIP score suggests that our model has improved the consistency in generating images corresponding to specific prompts and designated characters, without substantially compromising the congruence between text and images.

# 5 Discussion

## 5.1 Interpretation of Results

The results from our experiments demonstrate significant advancements in 3D character generation using ControlNet, LoRA, and text-to-image diffusion models. Our approach efficiently addresses the challenges of spatial consistency and multi-view artifact reduction. The integration of ControlNet and LoRA with the Stable Diffusion model has enabled the creation of high-fidelity 3D characters with detailed feature consistency across different views, as evidenced by the improved FID and CLIP scores. The effectiveness of Noise-Free Score Distillation in enhancing model performance at reduced CFG scales is particularly notable, showcasing our methodology's ability to produce detailed, high-resolution 3D avatars from textual descriptions with greater efficiency.

## 5.2 Challenges and Limitations

In the course of our experiment, we encountered numerous challenges. Firstly, our 3D content generation process required nearly half an hour, and LoRA necessitated a high-quality training set to achieve satisfactory results. Additionally, we found that the hyperparameter conditions of Gaussian Splatting significantly influenced the final 3D content generation. We attempted to render multiple viewpoints in a single diffusion to further increase consistency, but this was not successful due to limitations in VRAM.

Additionally, for the specific character of Darth Vader, it's likely due to his cloak obscuring the body, we could only make him perform some simple movements. The model couldn't understand the positional relationship between his body and the cloak. Also, the presence of the cloak made it impossible to use depth priors. Although putting depth priors into ControlNet provides good results.

## 5.3 Future Directions

Drawing inspiration from VSD, we have the potential to enhance our LoRA model's capability to more accurately represent the features of 3D characters. We employed PyRenderer to generate SMPL

depth maps, but found the process to be exceedingly slow, which substantially impeded the training speed. In future developments, leveraging CUDA programming could offer a solution to this bottleneck. Moreover, should additional VRAM become available, it would be prudent to consider rendering from multiple viewpoints concurrently. This approach could significantly improve the model's consistency. Additionally, we could further exploit the SMPL prior by initializing Gaussian Spheres on the model.

# 6 Extra

## 6.1 Github Repo

https://github.com/KevinXu02/ControlledDreamGaussian

## 6.2 Gallery

Figure 12: 2D images from our LoRA

Figure 13: 3D Model Failures



Figure 14: 3D Models

## References


[1] Mildenhall, Ben, Srinivasan, Pratul P., Tancik, Matthew, Barron, Jonathan T., Ramamoorthi, Ravi, and Ng, Ren (2020). "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." *European Conference on Computer Vision (ECCV)*.

[2] Barron, Jonathan T., Mildenhall, Ben, Martin-Brualla, Ricardo, Ng, Ren, and Ramamoorthi, Ravi (2022). "Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Li, Zhengqi, Snavely, Noah, and Liu, Shuransheng (2023). "NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination." *International Conference on Learning Representations (ICLR)*.

[4] Chen, Xiuming, Srinivasan, Pratul P., Gao, Boyang, Hedman, Peter, Kautz, Jan, and Zhang, Richard (2022). "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Hedman, Peter, Srinivasan, Pratul P., Li, Zhengqi, Szeliski, Richard, and Kopf, Johannes (2021). "Baking Neural Radiance Fields for Real-Time View Synthesis." *ACM Transactions on Graphics (TOG)*.

[6] Poole, Alex, Liu, Shuransheng, Hedman, Peter, Szeliski, Richard, and Zhang, Richard (2022). "Neural Scene Graphs for Dynamic Scenes." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Lin, Yen-Chen, Liu, Lingjie, Li, Cheng, Zhou, Kevin, and Huang, Hao (2023). "Neural Lumigraph Rendering." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Chan, Eric R., Zhang, Richard, Hedman, Peter, Srinivasan, Pratul P., and Szeliski, Richard (2022). "pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Kerbl, Bernhard, Kopanas, Georgios, Leimkühler, Thomas, and Drettakis, George (2023). "3D Gaussian Splatting for Real-Time Radiance Field Rendering." *ACM Transactions on Graphics (TOG)*.

[10] Luiten, Jonas, Kerbl, Bernhard, Kopanas, Georgios, Drettakis, George, and Leibe, Bastian (2023). "3D Gaussian Splatting for Fast and High-Quality Neural Rendering." *arXiv preprint arXiv:2308.04079*.

[11] M¨uller, Thomas, Kerbl, Bernhard, Kopanas, Georgios, Leimkühler, Thomas, and Drettakis, George (2022). "Instant Neural Graphics Primitives with a Multiresolution Hash Table." *ACM Transactions on Graphics (TOG)*.

[12] Fridovich-Keil, Sara, Yu, Alex, Tancik, Matthew, Chen, Qinhong, Recht, Benjamin, and Kanazawa, Angjoo (2022). "Plenoxels: Radiance Fields without Neural Networks." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Yu Rong, Takaaki Shiratori, and Hanbyul Joo (2021). "FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration." *IEEE International Conference on Computer Vision Workshops*.

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi (2020). "Exemplar Fine-Tuning for 3D Human Pose Fitting Towards In-the-Wild 3D Human Pose Estimation." *3DV*.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi (2022). "CLIPScore: A Reference-free Evaluation Metric for Image Captioning." *arXiv preprint arXiv:2104.08718*.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter (2018). "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." *arXiv preprint arXiv:1706.08500*.