StyleScape: Stylized and Depth-Consistent 3D Scene Generation

Emily Zhang
Stanford University
emily49@stanford.edu

Chetan Nair
Stanford University
cnair@stanford.edu

Abstract

In this paper, we propose a method for 3D-consistent scene generation to create diverse, visually appealing, and infinite walkthroughs of specific films. Our approach builds on the SceneScape pipeline and uses Dreambooth to train Stable Diffusion models for few-shot, scene-driven generation of novel views with specific characters, settings, and styles. We use ControlNet to condition these models for inpainting tasks, ensuring structurally and contextually coherent scenes as the model inpaints the masks recursively generated from the depth estimation model and the constructed 3D mesh. We conducted experiments on four stylistically distinct films, and our method outperforms the baseline on both quantitative aesthetic metrics and qualitative human evaluations, based on content and style resemblance with the target film.

1. Introduction

The field of 3D scene generation has seen significant advancements, particularly in its applications across graphics, virtual reality (VR), and animation. The ability to generate realistic 3D scenes from textual descriptions is a challenging and computationally intensive task and there is a lot of research being conducted in this field currently.

In this paper we explore the generation of consistent 3D scenes based on text prompts. In particular, we aim to implement and extend the method proposed in the SceneScape [1] paper which proposes a pipeline for text-driven, depth-consistent 3D scene generation.

The method described in the paper has demonstrated impressive results in generating semi-realistic, non-stylized 3D scenes. These scenes are created by using an inpainting model to generate 2D images from a text prompt, projecting them at different camera angles, then using inpainting to complete a consistent scene. However, for applications in entertainment and other creative industries, there is a need for generating more diverse and unique 3D scenes that adhere to various styles. To introduce this diversity, we finetune multiple diffusion models on stylized animated films.

To achieve this, we plan to leverage insights from the Google DreamBooth [6] paper to help finetune our inpainting diffusion model. The DreamBooth approach focuses on personalized text-to-image generation, enabling fine-tuning of diffusion models on specific styles and subjects with a limited number of images. This method will help us adapt our diffusion models to generate 3D scenes in a variety of unique styles effectively.

1.1. Contributions

Our contributions in this paper are:

- Experimentation with various text-to-image diffusion models and adapters for the task of few-shot scene generation
- A process of fine-tuning Stable Diffusion with Dreambooth that allows the model to synthesize backgrounds, characters, and styles of a scene into new compositions. This is a novel area of employing Dreambooth for scene-driven generation instead of subject-driven generation.
- Integration of ControlNet [9] to adapt Stable Diffusion for the inpainting task such that contextual and structural consistency is preserved.

2. Related Work

3. Diffusion and Inpainting

Instrumental to our work are diffusion models for image generation [5], which are a class of generative models that output images through a gradual denoising process. In the forward process, over a series of time steps, Gaussian noise is added to the data. This gradually moves the distribution towards pure noise. Structure in the data is removed. In the reverse process, the model denoises the data step by step, a process that the model learns. The training objective is to minimize the difference between the two processes so that the model can accurately denoise at each step.

There has also been extensive work done in training diffusion models specifically for inpainting tasks [4]. Inpaint-

ing is a conditional generation task whereby the model generates parts of an image that are missing after conditioning on the existing parts. The training process involves randomly masking parts of the input data which a diffusion model learns to generate seamlessly. This is ensured by a recontruction loss that is applied which penalizes generations that don't blend in well with the surrounding context and also a perceptual loss

3.1. Controlling Mechanisms: ControlNet and IP-Adapter

There exists a number of different control mechanism for guiding the diffusion generation process more precisely. This is needed for us to be able to effectively generate a consistent scene. ControlNet [9] and IP-Adapter [7] are two methods that fare relatively well.

ControlNet introduces this improved level of control by conditioning on specific input images like edge maps or segmentation maps. The architecture of the model is modified so that, during training, the model conditions both on the noise vector of diffusion but also on an additional control input. The control input is concatenated with the noise at different stages of the diffusion. This allows for more contextually consistent frame-by-frame generation of images that maintains the structure and content of prior frames, essential to successfully generating a coherent walkthrough.

IP-Adapter is another method of controlling diffusion that conditions the generated image on an image prompt. The model architecture is modified to incorporate the style, structure and / or content of the conditioning image. Additional layers are added so that features from the image prompt are maped onto the diffusion models's latent space. These mapping allow the model to condition the noise vector on the text and image prompt. Leveraging this provides us greater control guiding the frame generation process by feeding in prior frames.

3.2. DreamBooth

DreamBooth [6] is an approach that aims to finetune diffusion models on specific styles and subjects with a limited number of images. The training procedure involves introducing a unique identifier to the text prompts that are used to finetune the model. Then, the model is trained on a small set of images (usually less than ten) that capture a new style or subject. The unique identifier is used during the training process so that the model learns to associate it with the particular style or subject. The model employs a reconstruction loss for the regions that are masked and also a semantic loss that keeps the content in line with the prompt. Through this training, the model preserves most of it's priors learned from pretraining and also adapts to the new style or subject on which it's finetuned. We attempt to leverage this capability of the DreamBooth process to stylize our scenes.

3.3. SceneScape

Foundational to our work is the pipeline described in SceneScape [1] which leverages a pre-trained inpainting diffusion model alongside a pre-trained monocular depth estimation model to incrementally generate a walkthrough of a static scene represented as a triangular mesh from text prompts. Initially a text description is used to generate a 2D image. A pre-trained monocular depth estimation model is used to generate a depth map of this generated image. Using that, the image is then projected onto a 3D space. The current frame content is projected onto the next camera position, which produces a mask for the visible content, a depth mask and a masked frame. An inpainting model then fills in the occluded parts based on the masks produced from the projection. This generates subsequent frames in a consistent manner and these frames are all finally put together to generate the walkthrough.

While SceneScape performs extremely well, it primarily focuses on non-stylized scene generation. In this paper, we aim to extend this to more diverse and stylized output through our approach.

4. Methods

Given that the original SceneScape text-to-scene pipeline performed well on a variety of diverse scene walkthroughs, from underwater caves to grand libraries, we hypothesized that generating scenes conditioned on the *style* and *content* of specific films would be possible by just changing the 2D part of the pipeline, without modifying the 3D part.

In particular, we replace the original inpainting model, which is the default inpainting version finetuned from Stable Diffusion v2-base. We experimented with three different modifications:

1. At inference, use the IP-Adapter image prompt adapter to include an additional image condition for the film scenes on the default text-to-image inpainting model. This requires no training.

2. Finetune the default inpainting model using the Dreambooth method on the training images of a film. This was an implementation we wanted to try, as there is no published analysis on the effectiveness of this method.

3. Finetune a non-inpainting diffusion model using the Dreambooth method on the training images of a film. Then, apply ControlNet conditioned on the inpainting mask to use this model as an inpainting model during inference. Unlike (2), this has the benefit of learning entire images without masking, which is also the method described in the original Dreambooth paper.

To prevent catastrophic forgetting and because multi-token Dreambooth is not a developed area, for the last two models that require finetuning, we train a separate diffusion model on each of the four films.

Note that all three of our methods still use a text-to-image diffusion model, instead of a purely image-to-image method, so that we can leverage the strong semantic prior learned by Stable Diffusion, which makes few-shot image generation much easier. As the original SceneScape work demonstrated promising, generalizable 3D generation without expensive training on large 2D or 3D domain-specific datasets, we intend to present a modification upon their work that does not introduce inaccessible data or compute requirements.

5. Data

Given that the Dreambooth method we will use for fine-tuning latent stable diffusion models only requires a few input images depicting the same class, we manually collected three images each from four different films. These films, listed in Table 1, were carefully selected to represent a diverse range of visual styles. For example, the animation in Studio Ghibli's *Spirited Away* is distinctly different from the CGI for Hogwarts in *Harry Potter*. The three images of each film are shots of the same setting from different camera angles so that Dreambooth can find similarities among the training images. We believe that these few-shot data parameters are lenient and do not add much difficulty for using our pipeline in comparison to the original SceneScape's purely text-based pipeline, since these images were found by a quick Google search for the respective film. See our complete set of training images at `https://tinyurl.com/stylescape-231n`.

Film (Scene)	Studio	Year
Spirited Away (Train)	Studio Ghibli	2002
Harry Potter 1 (The Great Hall)	Warner Bros.	2001
When Marnie Was There (Anna's room)	Studio Ghibli	2014
Frozen (Arendelle Castle)	Disney	2014

Table 1: Film scenes in our dataset with their corresponding production studios and years

6. Experiments

For inference on all of the diffusion models, we use the parameters in table 2. For training and inference, we use the Hugging Face diffusers library [2].

6.1. Experimentation with IP-Adapter

We first attempted our inference-time method of conditioning default Stable Diffusion Inpainting on the image scenes using IP-Adapter. We used the same prompt as the default SceneScape baseline (for Frozen, this is *"POV, walkthrough, Arendelle castle from Frozen, masterpiece, indoor scene, best quality"*) and conditioned on the same scene image at every time step. We tested different values of the IP-adapter scale to vary the amount of image vs. text conditioning that is applied, and even tried only applying IP-Adapter to the down-part block 2 and up-part block 0 layers of the model, which control layout and style. However, all of these resulted in distorted inpainting results (as shown in Figure 1), which is likely due to IP-Adapter treating the scene as a subject, and trying to incorporate it all into the masked regions, which are warped regions from depth estimation and thus also geometrically complex.

Parameter Name	Value
Number of Inference Steps	50
Classifier-free Guidance	7.5
Size	600px
Negative Inpainting Prompt	*text, writings, signs, text, white border, photograph border, artifacts, blur, smooth texture, foggy, fog, bad quality, distortions, unrealistic, distorted image, watermark, signature, fish-eye look, windows, people, crowd, outdoor, landscape, view, chandelier*

Table 2: Diffusion Parameters for Inference

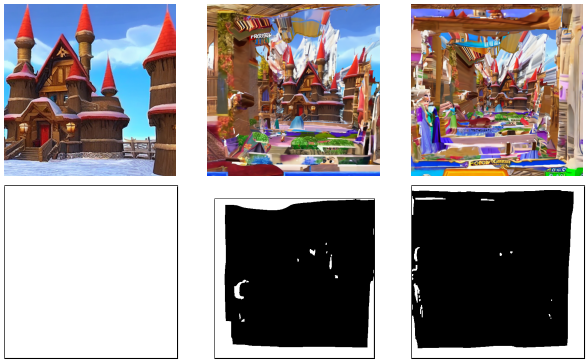


Figure 1: Sample frames from the video generated by IP-Adapter conditioned on a *Frozen* scene with scale 0.6

6.2. Experimentation with Dreambooth on Inpainting Model

Now, we move on to the more involved fine-tuning methods. By using masked versions of the training images to

Figure 2: Training images, scenes from *Frozen*

(a) Samples generated by Stable Diffusion Inpainting v2 model finetuned with Dreambooth on images shown in Figure 2

(b) Samples generated by Stable Diffusion v1.5 base model finetuned with Dreambooth on images shown in Figure 2

finetune the Stable Diffusion Inpainting model, we hoped to retain inpainting performance while synthesizing content and style from the film scenes. For finetuning, we use the parameters and prompt in Table 3. For all Dreamboothed models, we train the text encoder along with the UNet to improve results, especially on faces and characters. Our model checkpoint after finetuning on the *Frozen* data is available at `https://huggingface.co/emily49/frozen-stable-diffusion-inpaint`.

Parameter Name	**Value**
Resolution	512px[1]
Learning rate	$5e^{-6}$
Batch size	1
Number of steps	500
Instance prompt	*"A photo of sks **[Film name]**"*[2]
Inference prompt	*"POV, walkthrough, **[Scene name]** from sks **[Film name]**, masterpiece, indoor scene, best quality"*

Table 3: Parameters for Dreambooth fine-tuning and inference

While this model excelled at inpainting, with little to no visual discrepancies or gaps, we found two qualitative issues with the resulting generated images and scenes.

The first is that the fine-tuned inpainting model did not learn many of the details and overall style of the scene. For example, see Figure 2 and 3a, in which the former is the Dreambooth training images and the latter is three samples from inference on our trained checkpoint with an empty mask[3]. We can see that the generated images are not similar to the training images, and rely far more on the semantic prior of *"castle"* than on the new token *"sks frozen"*. This is likely because much of the references are masked during training, in order to ensure the model does not forget the inpainting capabilities.

The second issue we discovered is that in the generated 3D walkthrough, the model experiences quality degradation from error accumulation rapidly, as can be seen in Figure 4. We suspect that this is likely because Dreambooth is traditionally used only for subject-driven generation, not scene-driven generation. In fact, we have found no prior research that uses Dreambooth to generate variations of scenes instead of subjects. Specifically for inpainting, the model is trying to incorporate all facets of class *"sks frozen"* into the small masked areas, which is theoretically and experimentally the incorrect behavior.

Figure 4: Frame 0 and Frame 12 of a video generated from finetuned Stable Diffusion Inpaint

[1] Training images are randomly cropped to 512×512.

[2] *sks* is the unique token identifier we choose to use for Dreambooth

[3] This means the entire image is generated; this is also how we generated the first frame in the video pipeline.

6.3. Dreambooth Stable Diffusion + ControlNet

Finally, we move onto the last and most successful method we developed. Since Dreambooth experimentally works ineffectively on the inpainting model, we finetuned a non-inpainting base Stable Diffusion model using the same parameters in Table 3. Like before, we trained both the UNet and the text encoder. This yielded excellent results on the image generation task, as seen in Figure 3b. The characters, architecture, color scheme, and style were all effectively learned and synthesized to generate new views of the film setting. We have successfully used Dreambooth for scene-driven image generation instead of subject-driven generation.

In order to then perform the inpainting task using this non-inpainting model, we apply the ControlNet Inpainting model [8], which is already pretrained on the image inpainting task, on top of the diffusion model. We can then pass the conditional mask to the ControlNet model for inpainting. However, ControlNet is known to behave poorly with edges, which was prominent in our 3D pipeline, which can have no empty gaps. We fixed this by passing in a mask that was slightly dilated by a 5x5 filter.

Note that pretrained ControlNet only works with the architecture of Stable Diffusion v1.5, which varies from Stable Diffusion v2 and forward. Because of this, our final method finetunes the older v1.5 Stable Diffusion base models, but it still exceeds the alternate methods and the SceneScape baseline, as we will report in the next section. The finetuned model checkpoints for all four films can be found at `https://huggingface.co/emily49`.

7. Results

We generated 30-frame[4] long videos on an NVIDIA L4 GPU with the same fast-moving camera translation and rotation as SceneScape to maintain a robust baseline for comparison. Figure 5 shows sampled frames through time from our generated videos; we recommend looking at the videos at `https://tinyurl.com/stylescape-231n`.

Since stylized scene walkthrough generation is a qualitative and visual task, we use human evaluation as our primary metric. For each of the four films, we generated a video with the existing baseline SceneScape pipeline using the same prompt that we gave to our model, excluding the *sks* token. With the baseline video (abbreviated SS) and our video, we employed the Two-alternative Forced Choice protocol, asking 34 survey respondents the following two questions about each film:

1. **Choose the video that more closely resembles the specific film and setting in content and style.** We ask that you please do a *quick Google photos search of "Interior of [Setting] from [Film]"* to see what each film setting looks like before answering. For example, for the first question, google "Interior of Arendelle Castle from Frozen".

2. **Choose the video that has better visual quality, e.g., sharper, less artifacts such as holes, stretches, or distorted geometry.** For this question, you do not need to consider the specific film or setting, just the geometric consistency and quality.

We included the second question because it was the question that respondents were asked in the SceneScape paper to compare against existing methods. While the primary objective of our model is to generate scenes that have the look and feel of certain films, we still want to retain most of the geometric consistency and quality such that the scene is convincingly 3D.

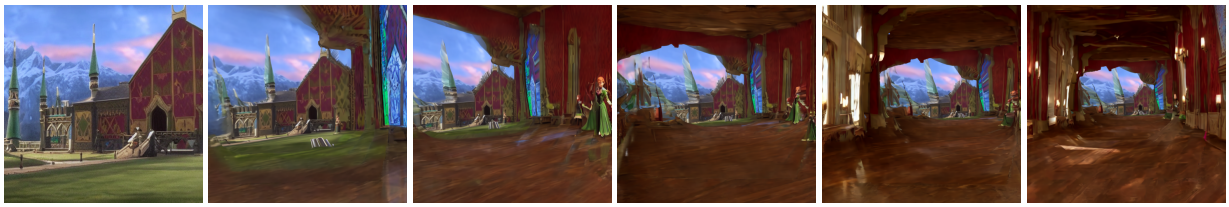
The results are shown in Table 4. We see that for all four films, our method considerably outperforms the baseline in terms of likeness to the corresponding film setting. For visual and geometric quality, both methods perform similarly. There are two films where the baseline outperforms our method. This is likely because the diffusion + ControlNet model was not trained end-to-end on the inpainting task and thus is more likely to create discrepancies (not to mention SD v1.5 is a lower version pretrained on less data). The two films where our model performs better on visual quality are actually more stylized films (*Spirited Away* and *When Marnie Was There*) since our diffusion model is better at generating stylized and depth-consistent images than the default inpainting model.

For a quantitative metric, we also computed a CLIP Aesthetics score [3], an aesthetic predictor on top of CLIP embeddings that was also used for SceneScape. We calculated this metric for a video by taking the average of the CLIP scores on the first, middle, and last frames. We see similar CLIP scores for the baseline model and our model on the same film, with our model marginally improving upon the score in aggregate.

	Frozen	*SA*	*HP*	*Marnie*	Avg.
Film-SS	22.2%	11.1%	5.60%	13.9%	13.2%
Film-Ours	**77.8%**	**88.9%**	**94.4%**	**86.1%**	**86.8%**
Quality-SS	**63.9%**	27.8%	**61.1%**	47.2%	**50.0%**
Quality-Ours	36.1%	**72.2%**	38.9%	**52.8%**	**50.0%**
CLIP-SS	**5.97**	**6.26**	6.20	6.13	6.14
CLIP-Ours	5.69	6.16	**6.91**	**6.22**	**6.25**

Table 4: Metrics from Human Evaluation and CLIP Aesthetic Score on baseline SceneScape method and our method across all films

[4]While the original SceneScape paper generated 50-frame videos, we were constrained by memory and compute, and 30 frames is substantial enough to evaluate aesthetic, stylistic, and geometric quality.



(a) *"POV, walkthrough, Arendelle Castle from sks frozen, masterpiece, indoor scene, best quality"*

(b) *"POV, walkthrough, train from sks spiritedaway, masterpiece, indoor scene, best quality"*

(c) *"POV, walkthrough, the Great Hall from sks harrypotter, masterpiece, indoor scene, best quality"*

(d) *"POV, walkthrough, Anna's room from sks marnie, masterpiece, indoor scene, best quality"*

Figure 5: Sample frames from video walkthroughs of each film using our method and the corresponding inference prompts

8. Conclusion

In this paper, we aimed to develop a method of generating 3D-consistent scene walkthroughs of specific films that retain style and content in a few-shot manner. We successfully finetuned a diffusion model on multiple views of the same film setting such that it synthesizes diverse, context-rich, and visually appealing images from the same setting. This is a novel use of the Dreambooth training method for scene-driven generation. We then integrated finetuned diffusion into the 3D pipeline pioneered by SceneScape using ControlNet for mask conditioning. The scene walkthroughs that we generated substantially outperformed SceneScape in their likeness to their corresponding films. However, both the baseline and our method struggle with geometric consistency and quality for stylized scenes, which suggest a domain-specific method or an alternative depth estimation model would be required for animated scenes. This is an area of future research for generative 3D models before they can be used to fully reconstruct convincing imaginary worlds.

References

[1] R. Fridman, A. Abecasis, Y. Kasten, and T. Dekel. Scenescape: Text-driven consistent scene generation. *arXiv preprint arXiv:2302.01133*, 2023.

[2] Hugging Face. Diffusers. https://huggingface.co/docs/diffusers.

[3] LAION AI. Clip aesthetic score predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022.

[4] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 10684–10695, June 2022.

[6] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[7] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[8] L. Zhang. Controlnet - v1.1 - inpaint version. https://huggingface.co/lllyasviel/control_v11p_sd15_inpaint.

[9] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.