SCENEDERELLA: Text-Driven 3D Scene Generation for Realistic Artboard

Jungmin Lee†
Dept. of Advanced Imaging
GSAIM, Chung-Ang University
Seoul, Republic of Korea
jngmnlee@vilab.cau.ac.kr

Haeun Noh†
Dept. of Artificial Intelligence
Chung-Ang University
Seoul, Republic of Korea
nhe8354@vilab.cau.ac.kr

Jaeyoon Lee
Dept. of Artificial Intelligence
Chung-Ang University
Seoul, Republic of Korea
leejaeyoon@vilab.cau.ac.kr

Jongwon Choi*
Dept. of Advanced Imaging
GSAIM, Chung-Ang University
Seoul, Republic of Korea
choijw@cau.ac.kr

* Corresponding author(s)
†These authors contributed equally to this work.

2025 IEEE International Conference on Consumer Electronics (ICCE) | 979-8-3315-2116-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCE63647.2025.10929907

***Abstract*—As global media and Video-on-Demand (VOD) markets expand, efficiently producing high-quality 3D scenes has become a key challenge. Our study named *SCENEDERELLA* generates 3D scenes useful for the film industry using simple text prompts. Since traditional diffusion models create fantastic images, they have limitations in generating scenes truly desired by real-world users. However, fine-tuning the clip encoder of the diffusion model enables the production of realistic 2D scenes for users. Once the initial 2D scenes align with the user's conceptual expectations, we construct 3D scenes using a depth estimator and 3D Gaussian Splatting. The study successfully demonstrates the generation of 3D movie scenes based on various movie scripts. Our method overcomes the limitations of traditional physical set production and offers a new approach that enables the rapid creation of diverse scenes. Additionally, it can be used as a pre-visualization video to attract investment in movies and dramas. Our approach is expected to introduce a new paradigm in content production, with film, gaming, and advertising. Additional results and interactive demos are available at our project website.**

***Index Terms*—3D Scene Generation, Text-to-3D, Film Pre-visualization**

I. INTRODUCTION

The media and entertainment industry evolves rapidly, with the growth of Video On Demand (VOD) systems based on Over-The-Top (OTT) platforms being particularly significant. Global OTT market revenue is expected to reach $316.4 billion in 2024, which is 9.89% higher than the market size in 2023 [1]. Under these circumstances, competition in the OTT market has intensified, especially around virtual studios, leading to an increased financial burden on content production, making it increasingly difficult for individuals to create content.

This situation highlights the need for new technologies that may reduce production costs and time. Traditional scene production tools require extensive manual labor and specialized modeling skills, making the process time-consuming and inefficient. Some recent approaches propose text-driven 3D scene generation, but they often generate scenes in a style different from the user's intent [2]. Even when a realistic scene is needed as the backdrop for a disaster movie, these methods sometimes produce the scene in an unrealistic animation style.

To address these limitations, we propose a novel method for generating 3D scenes that match the desired style using only text prompts. The approach utilizes a fine-tuned image diffusion model and a monocular depth estimator. To address the limitations of the diffusion model in generating realistic scenes, DreamBooth [3] was leveraged to fine-tune the text encoder based on real-world images. Users can iteratively generate images until the desired 2D scene is produced, and then expand the scene using the diffusion model's in-painting based on the generated 2D prior image. Finally, 3D Gaussian Splatting is applied to create the final 3D scene.

This study presents a new approach to film production by utilizing 3D scene generation technology tailored to the needs of film industry professionals.

- We provide an integrated 3D scene generation solution that produces realistic and unrealistic scenes by fine-tuning the text encoder of the diffusion model.
- Users can choose the scene generation mode based on the genre of the film they are producing to create scenes with any desired atmosphere.

II. RELATED WORKS

As the demand for tools to create 3D scenes in films has increased, interest in 3D scene generation technology has risen sharply. Traditional scene creation tools such as physical sets and Computer-Generated (CG) imagery, require extensive manual labor and specialized modeling skills [4], [5]. Editing is almost impossible and it takes high costs and time consumption. To address these limitations, the film industry has adopted generative models like 'text to video' and 'text to image' [6], [7]. However, these methods produce only 2D scenes, which lack the vividness and depth necessary for fully immersive experiences.

Some recent methods employ text-to-image diffusion models to learn 3D representations [8]–[10]. However, most of these methods focus on object generation, and research on text-driven 3D scene generation is relatively scarce. The existing studies for text-based 3D scene generation are limited in producing 3D scenes in a style different from the user's intent,

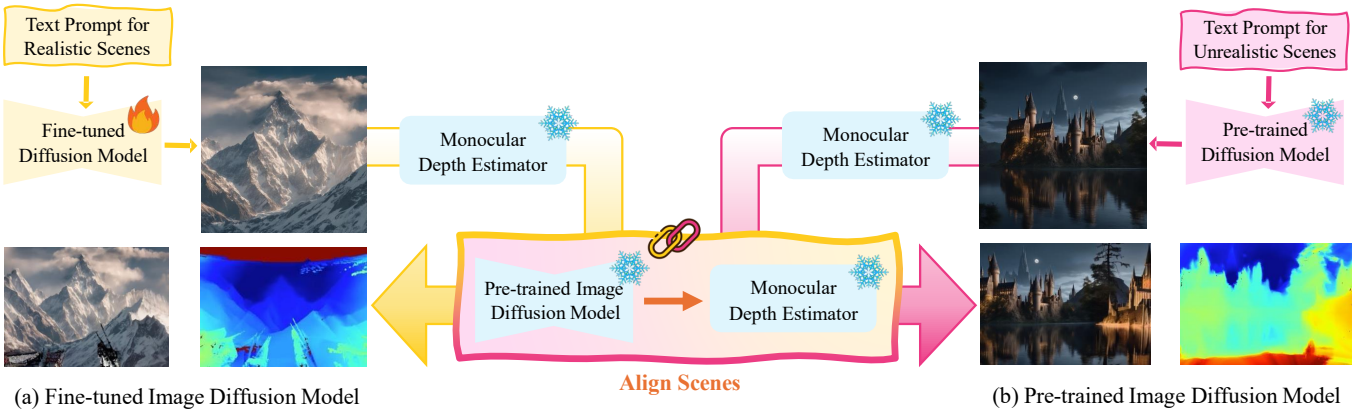

Fig. 1. The pipeline of SCENEDERELLA

as they generate initial images leveraging pre-trained diffusion models [2].

Diffusion models fail to align with the true distribution of real-world images when learning pixel distributions from a limited set of training images. This misalignment results in omitting important visual elements and the inability to distinguish style information, making it difficult to generate realistic images [11], [12]. Therefore, we propose a 3D scene generation specialized for the film industry that can meet user requirements in any scenario.

III. METHODS

Our SCENEDERELLA pipeline is illustrated in Fig. 1. The overall process follows the LucidDreamer [2] framework, but the diffusion model is replaced with a higher version to generate unrealistic scenes. The diffusion model has also been further fine-tuned to generate better realistic scenes.

First, the user selects either the unrealistic or realistic mode, depending on the desired scene style. Once a text prompt describing the scene is provided, the diffusion model generates a corresponding image. The pixels of the generated 2D image are then projected into 3D space. The point cloud is expanded through camera movement and in-painting, allowing for the generation of additional images and the construction of a larger 3D scene. After initializing the point cloud, 3D Gaussian Splatting [13] is applied to produce the final film scene.

A. Image Generation of Unrealistic Scene

Finding specific scenes online that match a desired scenario is highly challenging. Due to the difficulty of finding relevant text-image pairs for film production, we implement a system that generates the desired scenes using only text prompts. For fantasy films, like science fiction films, animation, fairy tales, and similar genres, our system generates scenes by inputting prompts that describe unrealistic backgrounds.

The initial 2D images generated based on these prompts are created by Stable Diffusion XL [14] which replace the diffusion model previously used in LucidDreamer [2]. Additionally, the in-painting model used to upscale the images to a higher resolution also employs the Stable Diffusion XL [14]. These modifications are intended to improve the overall visual fidelity of the output, enabling the generation of scenes similar to movie backgrounds.

B. Image Generation of Realistic Scene

In realistic scene generation, backgrounds required for producing content such as disaster films, war films, dramas, and documentaries are generated. The key difference from the unrealistic scene generation presented in Section A is that the CLIP encoder [15] of the diffusion model is fine-tuned with real-world images. This is because the diffusion model cannot achieve perfect photorealism.

Diffusion models struggle to distinguish and represent subtle nuances such as style, texture, and lighting, which often leads to the creation of more fantastical scenes. To address this issue, we leveraged DreamBooth [3] to fine-tune the text-to-image generation model to ensure high visual fidelity in realistic contexts. While fine-tuning the generative models with a small number of real-world subject images, the semantic knowledge of the original model is preserved.

C. 3D Scene Generation

Based on the generated initial 2D images, pixels are projected into 3D space to create an initial point cloud. The camera moves continuously while missing parts are in-painted to expand the point cloud. As the camera moves, newly generated images are converted into 3D point clouds and integrated to form a larger scene. The ZoeDepth [16], a monocular depth estimation model, calculates the relative depth of each image. The scale factor is optimized to minimize depth differences between the newly generated images and the existing point cloud. This optimization plays a critical role in resolving inconsistencies between point clouds while maintaining the overall image consistency.

The Gaussian Splatting model is trained using the generated point cloud. The initial point cloud serves as the center point for the Gaussian splats. The projected images guide the optimization of the positions and volumes of each point,

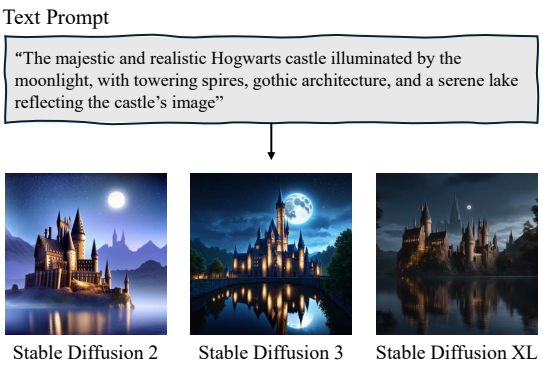

Fig. 2. Comparison of the generated image from different stable-diffusion models.

Fig. 3. Generated image from Stable Diffusion XL with various text prompts.

enhancing the details of the scene. Reprojected images further refine the point cloud, improving detail accuracy. The Gaussian Splatting approach generates the final output, which produces high-quality 3D movie scenes with multi-view consistency.

IV. EXPERIMENTS

A. Experimental Settings

SCENEDERELLA requires a dataset only for fine-tuning Stable Diffusion to generate realistic scenes, and no additional training datasets are needed for other cases. We identified that scenes involving natural disasters or accidents incur significant production costs in set design, and thus collected real images related to these themes from the internet. For each class, we carefully selected 10 to 20 images that best depict the respective scenes.

Implementation Details. We used the Stable Diffusion XL [14] for initial scene creation as the text-to-image generation module. Through our experiments, we determined that Stable Diffusion XL [14] is suitable for generating movie scenes, but the model was also configured to allow users to choose Stable Diffusion 2 [17] or 3 [18] if needed. After the initial 2D scene generation was completed, we used

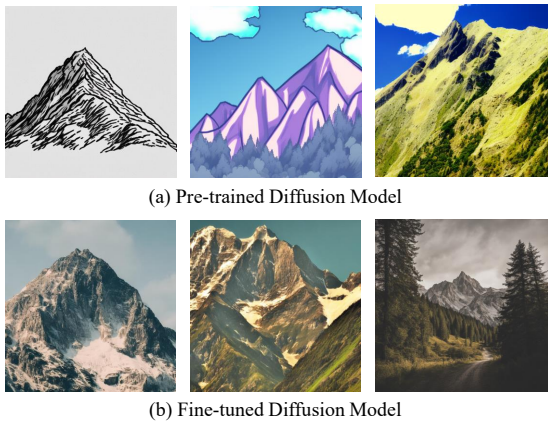

Fig. 4. Comparison of pre-trained and fine-tuned diffusion models for mountain scene generation

both Stable Diffusion XL [14] and 1.5 [19] to in-paint the masked images, applying the same text prompt. Since the XL model [14] generates scenes relatively slowly, we primarily used Stable Diffusion 1.5 [19] during the in-painting stage. All experiments, including fine-tuning, were conducted on an RTX 3090 24GB GPU.

B. Selection of Text-to-Image Model

We compared various Stable Diffusion models as 2D priors and found that Stable Diffusion XL [14] generated images that best matched the cinematic scenes. In Fig. 3, a prompt describing the movie 'Harry Potter' was used, and it can be observed that as the diffusion models advanced, the details of the castle and the rendering of light progressively improved.

In Stable Diffusion 2 [17], the structure of the castle and the background were expressed too simply, additionally in Stable Diffusion 3 [18], the moon and water surface appeared artificial. However, in Stable Diffusion XL [14], the rendering of light and shadow became more refined, resulting in better harmony between the background and the castle structure. As the diffusion models improved, the images became more detailed and expressive, providing a feeling closer to cinematic scenes.

Since 3D scene generation heavily relies on the 2D prior, the performance of the diffusion model is crucial. Therefore, we selected Stable Diffusion XL [14] as the model for expanding scenes into 3D through in-painting. The results of experimenting with various prompts are presented in Fig. 3, demonstrating that the 2D prior image is well-suited for XL [14] in generating the initial scene. These experiments validate the effectiveness of the 2D prior in guiding the model to produce accurate and contextually relevant first scenes.

C. Fine-tuning for Realistic Image Generation

We fine-tuned the model using DreamBooth [3] with real-world images to accurately represent not only unrealistic scenes but also realistic ones. For each class, 10 real-world images were used to fine-tune the diffusion model, enabling

(a) Pre-trained Diffusion Model

(b) Fine-tuned Diffusion Model

Fig. 5. Comparison of pre-trained and fine-tuned diffusion models for volcano as a subcategory of the mountain class

(a) Car

(b) Car Accident

Fig. 6. Visualization of the fine-tuned class and its subcategory

the generation of realistic backgrounds. We conducted experiments to evaluate how realistically the model can generate scenes involving the 'car' and 'mountain' classes. These tests aimed to assess the model's ability to create visually convincing and contextually appropriate representations of these specific elements.

As shown in Fig. 4(a), the results of the pre-trained diffusion model depicted mountain scenes in a style commonly seen in the animation or cartoon genre. These images emphasized vibrant colors and simplified details, which created a sense of dissonance when compared to the realistic representation required for cinematic scenes. The level of detail was limited, and the handling of light and shadow was also depicted unrealistically.

In contrast, the results of the fine-tuned diffusion model shown in Fig. 4(b) generated more realistic scenes based on real-world mountains. While the three resulting images had slight differences in lighting and color, the natural lines and shadows depicted the details of the mountains more realistically. In particular, the texture and lighting were rendered with greater precision, providing a level of realism suitable for use in film scenes. These results demonstrated that fine-tuning overcame the limitations of the pre-trained model and enabled the generation of realistic scenes closer to real-world images.

Fig. 5 shows the fine-tuning results for the volcano subcategory within the mountain class. As seen in (a), the volcanic scenes generated by the pre-trained diffusion model are characterized by a simplified representation of the mountain and exaggerated lava. In particular, the movement of the lava and the depiction of the volcanic eruption are not realistic, and the rendering of light and shadow is also limited. These images are somewhat distant from the natural representation expected in realistic cinematic scenes.

The car images shown in Fig. 6 also exhibited greater realism after fine-tuning. The natural textures and colors effectively conveyed a sense of realism and blended well with cinematic scenes. In particular, Fig. 6(b) shows the results of subdividing the car class into the car accident subcategory. This demonstrated that fine-tuning was possible with minimal data when specific scenes were required. Such class subdivision was useful for filming high-cost scenes like volcanic eruptions or car accidents. These results showed that fine-tuning overcame the limitations of pre-trained models, enabling the generation of more realistic scenes even for complex scenarios.

V. DISCUSSION

We have developed a text-based 3D scene generation system, which allows the generation of 3D scenes applicable to the film industry. By improving the diffusion model, which is a 2D image generation model, we have enabled the effective creation of 2D film scenes across various genres based on user requirements. Subsequently, the generated 2D scenes were used as the basis for completing the 3D scene using 3D Gaussian Splatting.

Additionally, a demo program was developed to enable professionals in the film industry to easily use the system, as shown in Fig.7. By integrating Gradio [20], a user-friendly interface allows for the quick generation of images and 3D scenes through text input. Users can describe the scene with text prompts and refine it using negative prompts. By clicking 'Generate Initial Scene', the scene is instantly visualized, and users can iteratively adjust it as needed. Once satisfied, they can choose the camera trajectory, select the rendering perspective, and press 'RUN' to generate the 3D scene. This intuitive interface simplifies the process, making it accessible even to those unfamiliar with complex tools, reducing both time and cost in film production.

The final 3D scene output can be downloaded and used in various modeling programs. In Fig. 8, we visualized a scene from the movie 'Avatar' using the Supersplat [21]. This demonstrates the successful generation of 3D scenes based on a movie script, overcoming the limitations of physical set production. These results can be applied not only in film and drama production but also in pre-visualization videos to attract investment in content. SCENEDERELLA is significant

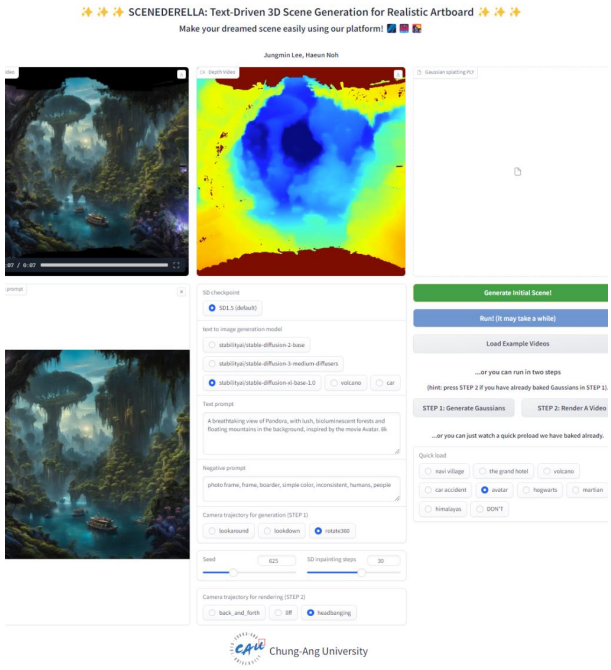

Fig. 7. UI of SCENEDERELLA demo

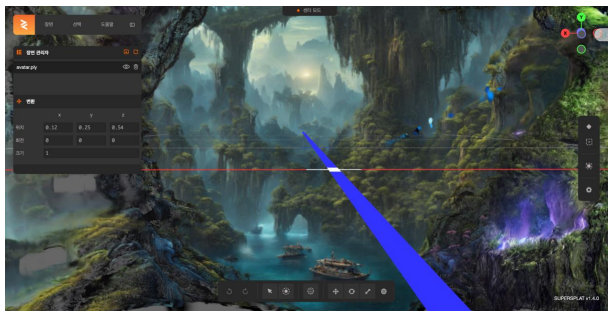

Fig. 8. Visualization of a scene from the movie 'Avatar'

in that it has laid the technical foundation for introducing a new paradigm in content creation for films, games, and advertisements.

However, in the 3D scene generation process, using the same text prompt repeatedly during in-painting causes repetitive objects to appear. While this method is effective in maintaining continuity in scene generation, it can lead to issues, such as the repeated appearance of the hotel building in the movie "The Grand Budapest Hotel," as shown in Fig. 9. To address this issue, future research will explore adding prompts specifically for the in-painting process or adjusting the attention map to develop a more stable in-painting method. Overcoming this limitation would enable the generation of more practical and diverse 3D scenes.

VI. CONCLUSION

We propose an integrated solution for text-based 3D scene generation. This approach allows for the effective creation of

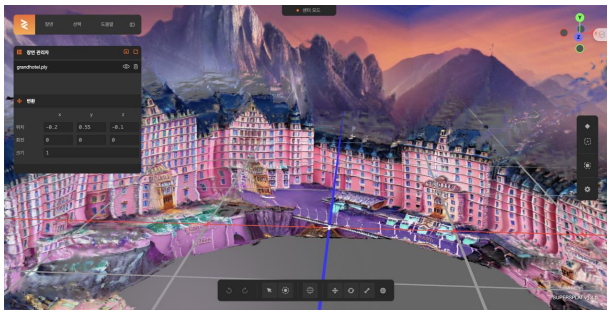

Fig. 9. Scene generation issue from the movie 'The Grand Budapest Hotel'

realistic or unrealistic 3D scenes in film production. To address the issues of unrealistic scene generation, we fine-tuned the text encoder using real-world images with DreamBooth [3]. This process enables the generation of 2D scenes in the desired genre, which are then expanded using the in-painting technique. Finally, 3D Gaussian Splatting is applied to produce the complete 3D movie scene.

By proposing an integrated solution for text-based 3D scene generation, we highlight the potential for innovation in scene production within the film industry. This method allows for 3D visualization of a film's concept and scenario, making it useful for pre-visualization to attract investment or as a background for actual film shooting. In particular, it can significantly contribute to early-stage visual conceptualization and cost savings by addressing the high costs and time consumption associated with traditional physical set production.

Moreover, this study applies not only to the film industry but also to various content production fields such as gaming and advertising, and it is expected to play an important role in future competition among OTT and VOD platforms. However, future research needs to address issues related to resolution and the repetition of elements during the in-painting process. With these improvements, 3D visualization technology holds the potential to introduce a new paradigm across the film industry and other related fields.

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang University)) and the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2023(2024)-RS-2024-00418847) supervised by the IITP.

REFERENCES

[1] Statista. (2024) Ott video - worldwide. [Online]. Available: https://www.statista.com/outlook/amo/media/tv-video/ott-video/worldwide

[2] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee, "Luciddreamer: Domain-free generation of 3d gaussian splatting scenes," *arXiv preprint arXiv:2311.13384*, 2023.

[3] 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," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.
[4] G. Esperdy, "From instruction to consumption: Architecture and design in hollywood movies of the 1930s." *Journal of American Culture*, vol. 30, no. 2, 2007.
[5] G. Luciano, *Essential computer graphics techniques for modeling, animating, and rendering biomolecules and cells: a guide for the scientist and artist*. Crc Press, 2019.
[6] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.
[7] 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 computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
[8] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
[9] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 300–309.
[10] Y. Chen, Y. Pan, H. Yang, T. Yao, and T. Mei, "Vp3d: Unleashing 2d visual prompt for text-to-3d generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4896–4905.
[11] S. Lin and X. Yang, "Diffusion model with perceptual loss," *arXiv preprint arXiv:2401.00110*, 2023.
[12] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, "Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.
[13] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
[14] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.
[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
[16] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
[17] S. AI, "Stable diffusion," https://github.com/Stability-AI/stablediffusion, 2022.
[18] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.
[19] Runway, CompVis, and S. AI, "Stable diffusion v1.5," https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5, 2022.
[20] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ml models in the wild," *arXiv preprint arXiv:1906.02569*, 2019.
[21] P. LTD. (2024) Supersplat v 1.4.0. [Online]. Available: https://playcanvas.com/supersplat/editor/