GenDeck: Towards a HoloDeck with Text-to-3D Model Generation

Manuel Weid* Navid Khezrian† Aparna Pindali Mana‡ Forouzan Farzinnejad§ Jens Grubert¶

Coburg University of Applied Sciences and Arts

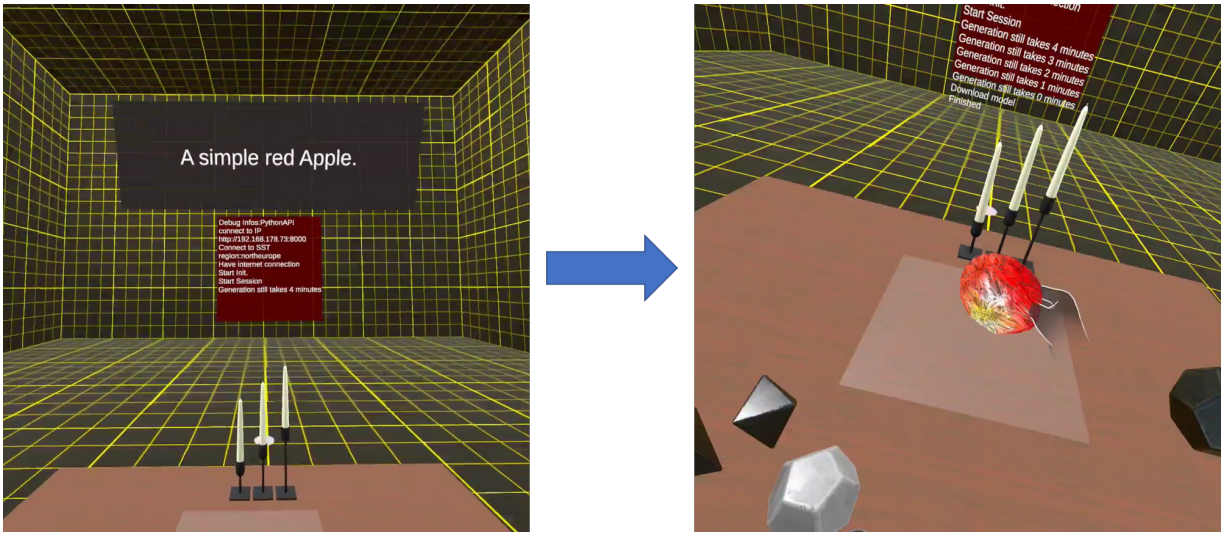

Figure 1: We present the proof-of-concept system GenDeck - an application to experience text-to-3D model generation inside an immersive Virtual Reality environment. Left: Initially, the user can create a textual description of the object they want to have created using speech input. Right: After the processing is finished, the final model is downloaded and ready for further use.

Abstract

Generative Artificial Intelligence has the potential to substantially transform the way 3D content for Extended Reality applications is produced. Specifically, the development of text-to-3D and image-to-3D generators with increasing visual fidelity and decreasing computational costs is thriving quickly. Within this work, we present GenDeck, a proof-of-concept application to experience text-to-3D model generation inside an immersive Virtual Reality environment.

Index Terms: Human Computer Interaction—Knowledge Work—Extende Reality;

1 Introduction

Artificial Intelligence-Generated Content (AIGC) is impacting the creative industry. AIGC already allows for high-fidelity text [2], image [11], and video generation [12] alongside further modalities (see recent survey [4]). It has been projected to have a substantial impact on future XR (and Metaverse) experiences [3]. Text-to-3D generation [5] is particularly relevant for the generation of content for Extended Reality (XR), but initial methods (e.g., [6, 10]) tended to have heavy computational demands requiring hours to generate individual objects. Recently, substantial improvements have been made both in terms of computational needs [8, 13] and visual fidelity [9, 14, 15] with further advancements being published regularly.

*e-mail: manuel.weid@stud.hs-coburg.de
†e-mail: navid.khezrian@hs-coburg.de
‡e-mail: Aparna.Pindali-Mana@stud.hs-coburg.de
§e-mail: forouzan.farzinnejad@hs-coburg.de
¶e-mail: jens.grubert@hs-coburg.de

Tools for immersive content generation inside VR also became popular in recent years (such as Gravity Sketch[1] but still require a substantial amount of expertise for the generation of 3D models. Driven by the idea of the HoloDeck, we propose a virtual environment, where users can generate 3D objects simply by describing their appearance using voice input (e.g., "generate a red apple). The idea of immersive content generation, and subsequent editing, could contribute to improving the speed and quality of 3D modeling for and inside of virtual environments.

To contribute to this goal, we introduce a modular framework combining a virtual environment created with Unity, along API calls to internet services for 3D content generation. Our software will be made available under https://gitlab.com/mixedrealitylab/gendeck.

2 The Demo Experiences

The user wears a VR-headset (Meta Quest 3) and is inside a room with a desk in front of them. They can activate speech recognition with a button press. Now the user can describe the object they want to have modelled using speech. The recognized text appears in front of them and status updates are shown about the progress of the overall process. Once the generation process is finished, the generated 3D model appears on a table in front of the user. Afterward, the user can grab and manipulate the object (scaling, rotation, translation). The user can also choose from different text-to-3D generators to experience various time/quality trade-offs. Please note, that, depending on the underlying machine learning model and desired 3D model quality, the generation process can vary between seconds and several minutes.

[1] https://www.gravitysketch.com/ last access January 9th, 2024

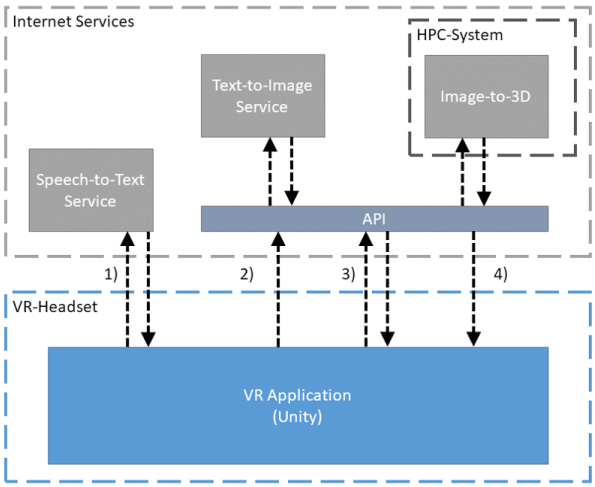

Figure 2: The VR application runs on a VR headset and utilizes various internet services. 1) The recorded voice of the participant is sent to a speech-to-text service. 2) Optionally, and depending on the chosen generative model, a text-to-image generator is called. 3) the text (or image) is then transferred to a 3D generator and 4) finally, the model is downloaded to the VR application. In our implementation, we use a dedicated service deployed on a high-performance computing (HPC) system and a custom API for the text-to-image and image-to-3D services.

3 System Overview

The system overview is shown in Fig. 2. The VR application is written in Unity. The recorded voice of the participant is sent to a speech-to-text service (e.g., Microsoft Azure speech-to-text) and processed. The retrieved text is then optionally forwarded to a text-to-image generator (in our case StableDiffusion [11]) and an image is generated. This image (or text) is then forwarded to a text/image-to-3D generator. In our case, we are using One-2-3-45 [8] and Wonder3D [9] but plan to integrate VolumeDiffusion [14] in the next step. While text-to-image generators are readily available as internet services, finding reliable text-to-3D generators that are accessible through an online API can still be challenging. Hence, as initial pre-processing step before running the application for the first time, we first download the respective text/image-to-3D repository, pack it into a Dockerfile[2] and convert into an Apptainer container[3]. Finally this is uploaded to a High-Performance Compute Server[4], which only supports the execution of Apptainer but not Docker files. For the text-to-image and image-to-3D generator, we use a custom API based on FastAPI[5]. Specifically, for the text/image-to-3D generator on the HPC system, we first log into the system via SSH, upload the text or image via SCP, request a compute resource (in our case an NVIDIA A100 80 GB) via SLURM[6] and start the compute job. As soon as it is finished, the 3D model is then downloaded to the VR application and is ready for further use.

4 Conclusion and Future Work

Within this work, we presented GenDeck, a proof-of-concept system to experience text-to-3D model generation inside an immersive VR environment. We foresee that both the quality of generated 3D models will steadily increase and compute time will rapidly decrease soon. Our system allows the modular use of various text-to-3D and image-to-3D generators in a flexible pipeline. Besides generating 3D models and animations in text-to-4D systems [1, 7]) an interesting future endeavor for us is to explore iteratively editing and refining initially generated 3D objects beyond traditional editing approaches. Specifically, we foresee multi-modal interaction techniques (such as speech + gesture) to update the parametrization of the underlying generative process (e.g., "add a leaf here").

[2]https://www.docker.com/ last access, January 10th, 2024

[3]https://apptainer.org/ last access, January 10th, 2024

[4]https://hpc.fau.de/systems-services/documentation-instructions/clusters/alex-cluster/ last access, January 10th, 2024

[5]https://fastapi.tiangolo.com/ last access, January 10th, 2024

[6]https://www.schedmd.com/ last access, January 10th, 2024

References

[1] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023.

[2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[3] V. Chamola, G. Bansal, T. K. Das, V. Hassija, N. S. S. Reddy, J. Wang, S. Zeadally, A. Hussain, F. R. Yu, M. Guizani, et al. Beyond reality: The pivotal role of generative ai in the metaverse. *arXiv preprint arXiv:2308.06272*, 2023.

[4] L. G. Foo and J. LIU. Ai-generated content (aigc) for various data modalities: A survey. *arXiv preprint arXiv:2308.14177*, 2023.

[5] C. Li, C. Zhang, A. Waghwase, L.-H. Lee, F. Rameau, Y. Yang, S.-H. Bae, and C. S. Hong. Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131*, 2023.

[6] 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*, pages 300–309, 2023.

[7] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023.

[8] M. Liu, C. Xu, H. Jin, L. Chen, Z. Xu, H. Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.

[9] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

[10] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[11] 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*, pages 10684–10695, 2022.

[12] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[13] C. Sun, J. Han, W. Deng, X. Wang, Z. Qin, and S. Gould. 3d-gpt: Procedural 3d modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023.

[14] Z. Tang, S. Gu, C. Wang, T. Zhang, J. Bao, D. Chen, and B. Guo. Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. *arXiv preprint arXiv:2312.11459*, 2023.

[15] J. Wu, X. Gao, X. Liu, Z. Shen, C. Zhao, H. Feng, J. Liu, and E. Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3202–3211, 2024.