Enabling Subtle Visual Modifications with AR+AI

Thesis director: Christian Sandor, ARAI Team, LISN
Co-Supervisor: Quentin Bammey, Centre Borelli, ENS Paris-Saclay
PhD Student: Hovhannes Margaryan, ARAI Team, LISN

Generative Artificial Intelligence (AI) has made incredible progress over the last few years. A related development is Augmented Reality (AR), with e.g. the recent release of Apple's Vision Pro headset. It is reasonable to assume that in the near future, humans will see the world around them through AR headsets that apply AI modifications to what they see. In this project, we will investigate how face-to-face interactions would be impacted when we use AR+AI (ARAI) to adjust environmental elements and conversational cues, including facial expressions, eye movements, and gestures. Thus, the research question of the thesis is: **is it possible to develop a distributed infrastructure for generating high quality subtle visual modifications in real time?**

Visual AI has progressed enormously and at a rapid pace, particularly after the advent of diffusion models (DM) [6, 2, 7, 5, 1]. Some exemplary works include text-to-video (OpenAI's SORA, announced on 15 Feb 2024), animation of photos based on example motions [8], and 3D gaussian splatting [3] for efficient rendering of neural radiance fields [4]. However, we are not aware of any work that addresses real-time modifications of a user's view of their environment with DMs. The means we envision (distributed architectures involving thin clients and heavy cloud computers) have not been explored either.

Generating fast, high-quality images and videos using DMs is very challenging. While achieving high quality or high speed individually has been demonstrated (e.g. FLUX.1-schnell[1], FLUX.1-dev[2]), a solution combining both has yet to emerge. For reference, a recent text-to-video model, Mochi 1[3] (released 23 Oct, 2024), requires 4 × H100 GPUs to generate a 5.4-second video at 30 frames per second (FPS).

The envisioned structure of the thesis includes the following stages:

- **First prototype of high-quality real time generation**: In the initial stage, our objective is to enhance the inference speed of text-to-image generation using Dr. Sandor's prototype while preserving image quality and diversity. Specifically, we aim to increase the frame rate from 20 FPS to 60 FPS. Additionally, we will establish a robust evaluation pipeline to assess and compare the quality of generated images across various experiments, benchmarking them against state-of-the-art methods.

- **Prototype refinement and extension to 3D and video**: In this stage, we will refine the results from the first stage and extend them to support additional modalities, such as multi-view generation, 3D, and video. We will also focus on temporal consistency, a critical factor for achieving coherence in 3D and video generation.

- **Text based Editing and Personalization**: In the last stage, we will explore the feasibility of real-time, text-guided editing and personalization across various modalities, including 2D images, 3D, and video. This investigation aims to enable AR system users to dynamically modify their environment in real time, allowing for immediate and personalized interaction.

[1] https://huggingface.co/black-forest-labs/FLUX.1-schnell
[2] https://huggingface.co/black-forest-labs/FLUX.1-dev
[3] https://www.genmo.ai/blog

References

[1] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), July 2023.

[4] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), July 2022.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[6] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[8] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model, 2023.