DLCV (DS265) Project Presentation

- By

Akanksha Sharma

(25709)

Video to Audio

Introduction

- Challenging task
- Requires alignment in terms of semantics, temporal, causal etc.
- https://youtu.be/PUKGyEve7XQ

Related Work

- AudioLDM[5]
- MMDiffusion[6]
- MeLFusion[3]
- ReWaS[7]
- MMAudio[1]
- FoleyCrafter[2]

Implementation of stat-of-the-art

- MMAudio[1] (transformer based model)

- FoleyCrafter[2] (diffusion based model)

Approach

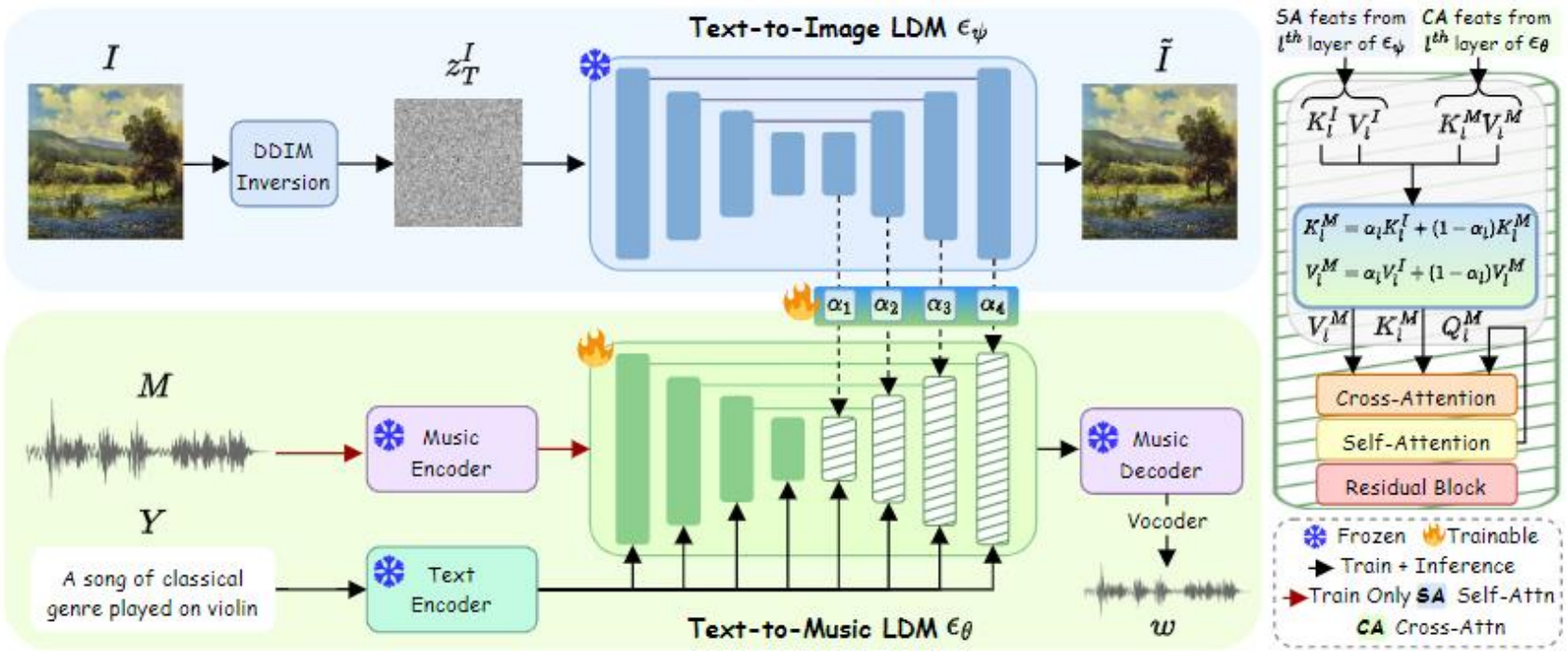

Figure 1: MeLFusion[3] Architecture

- WAVE: Warping DDIM Inversion Features for Zero-shot Text-to-Video Editing[4]

Dataset

- AudioSet
- Released in 2017
- Contains YouTube links for audio-videos, class of audio, time stamp

References

1. Cheng, Ho Kei, et al. "Taming multimodal joint training for high-quality video-to-audio synthesis." *arXiv preprint arXiv:2412.15322* (2024).
2. Zhang, Yiming, et al. "Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds." *arXiv preprint arXiv:2407.01494* (2024).
3. Chowdhury, Sanjoy, et al. "Melfusion: Synthesizing music from image and language cues using diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
4. Feng, Yutang, et al. "Wave: Warping ddim inversion features for zero-shot text-to-video editing." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
5. Liu, Haohe, et al. "Audioldm: Text-to-audio generation with latent diffusion models." arXiv preprint
6. 48 arXiv:2301.12503 (2023).Ruan, Ludan, et al. "Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
7. Jeong, Yujin, et al. "Read, watch and scream! sound generation from text and video." arXiv preprint arXiv:2407.05551 (2024).