

Anonymous Authors

The TAVGBench dataset we proposed comprises 1.7 million entries sourced from AudioSet [2], each containing 10 seconds of synchronized audio and video. However, the videos are not precisely split according to specific events, resulting in numerous transitions within the data. Additionally, the dataset includes videos with subtle movements. These characteristics may impact the performance of the TAVG model. In the future, we aim to develop more precise video split methods [1] and create higher-quality datasets that maintain consistent scene composition.

3 MORE EXAMPLES

We provide additional visualizations to illustrate our results. We show the results of more categories of video generation in the demo video. Including music scenes, landscapes, daily scenes, animation scenes, game scenes, *etc.* The details are shown in demo.mp4.

REFERENCES

[1] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *arXiv preprint arXiv:2402.19479* (2024).

[2] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.