Hear What You See: Video-to-Audio Generation with Diffusion Transformer and Semantic-Temporal Alignment-Ranked Direct Preference Optimization

Supplementary Material

1. More Related Work

1.1. Diffusion Transformer

Diffusion Transformer (DiT) models [2, 12, 13, 19] have emerged as a dominant paradigm in AI content creation due to their scalability and incredible performance in visual generation tasks. With flow matching [10] offering more stable and robust training than the score-based diffusion process [15], some flow-based diffusion transformers [4, 8, 11] are proposed to enhance the efficiency and quality of text-to-image generation. Notably, Stable Diffusion 3 (SD3) [4] employs multimodal DiT blocks to enable bidirectional information flow between image and text tokens. Unlike SD3, Flux [8] incorporates a sequence of single-modality DiT blocks after multimodal DiT blocks to further strengthen image generation, whose structure design has proven effective for other modalities such as audio [3, 6]. To improve the training stability and inference efficiency of flow-based DiT models, Lumina-T2X [5], Lumina-Next [20], and Lumina-Video [9] introduce more efficient attention mechanisms for integrating visual tokens with textual conditions. Built upon these advancements, our flow-based VisioSonic introduce a video-text-audio co-attention mechanism to efficiently interact aligned video-audio latents with multimodal conditions. In addition, our work removes the single-DiT stream and exclusively adopts pure transformer blocks, achieving a conciser model structure and faster inference speed than MMAudio [3].

2. Effects of Different Model Components

We conduct an ablation study to assess the contribution of each component, as shown in Table 1. Removing the conditioner leads to a significant degradation in audio quality, reflected by a substantial increase in FD and KL scores. In addition, removing the token aligner degrades semantic consistency, as indicated by the drop in IB-score, while omitting Syncformer weakens temporal synchronization, resulting in a higher DeSync score.

Table 1. Comparison results of different components.

	FD_{PaSST}↓	KL_{PaSST}↓	IB-score↑	DeSync↓
VisioSonic Base	**58.27**	**1.30**	**32.8**	**0.45**
- Conditioner	102.7	2.7	21.9	0.86
- Token Aligner	60.9	1.45	31.5	0.52
- Syncformer	61.1	1.32	32.7	0.79

Table 2. Comparison of other baselines w/ and w/o STAR-DPO.

	MMAudio	+STAR-DPO	Frieren	+STAR-DPO
IB ↑	32.27	**33.56**	22.78	**25.67**
DeSync ↓	0.44	**0.42**	0.851	**0.71**

3. Generalization of STAR-DPO

To evaluate the generalization capability of STAR-DPO, we apply it to two representative V2A baselines, MMAudio and Frieren, and observe consistent improvements in both semantic and temporal alignment, as shown in Table 2. Specifically, STAR-DPO improves IB-score and reduces DeSync for both models, demonstrating its effectiveness across different architectures and its ability to consistently enhance audio-visual alignment.

4. Robustness of Automated Labeling

To assess the robustness of the automated labeling, we conduct a human study on 200 randomly sampled videos from 50 classes in the VGGSound validation set. For each video, annotators select the best and worst audio among five candidates based on temporal alignment, semantic alignment, and overall quality. As shown in Table 3, the agreement between human annotations and the pretrained ranker is consistently high for both Top-1 (winner) and Bottom-1 (loser) selections across all criteria. These results suggest that the pretrained ranker aligns well with human judgment, supporting the reliability of the automated labeling.

Table 3. Agreement between automatic and human labeling.

	Semantic Agree.	Temporal Agree.	Overall
Top-1 (winner)	81%	82%	80%
Bottom-1 (loser)	85%	87%	85%

5. Effect of Reward Model Ratios

To construct audio-video preference pairs, we leverage existing ImageBind and SynchFormer as the reward models to judge generated audio samples. Specifically, the ImageBind uses internal audio and video encoders to extracts audio and video semantic latents, which are used for similarity calculation to obtain the audio-video semantic alignment score.

Table 4. Comparison results on various weight ratios between ImageBind and Synchformer rewards.

ImageBind:Synchformer	Distribution matching					Audio quality	Semantic align.	Temporal align.
	$FD_{PaSST}\downarrow$	$FD_{PANNs}\downarrow$	$FD_{VGG}\downarrow$	$KL_{PANNs}\downarrow$	$KL_{PaSST}\downarrow$	IS↑	IB-score↑	DeSync↓
1:1	**55.48**	**4.36**	**0.99**	**1.44**	**1.29**	**18.41**	**33.1**	**0.41**
1:2	55.61	4.62	1.09	1.46	1.30	17.66	32.9	0.42
2:1	57.50	5.08	**0.99**	1.47	**1.29**	17.70	33.0	0.43

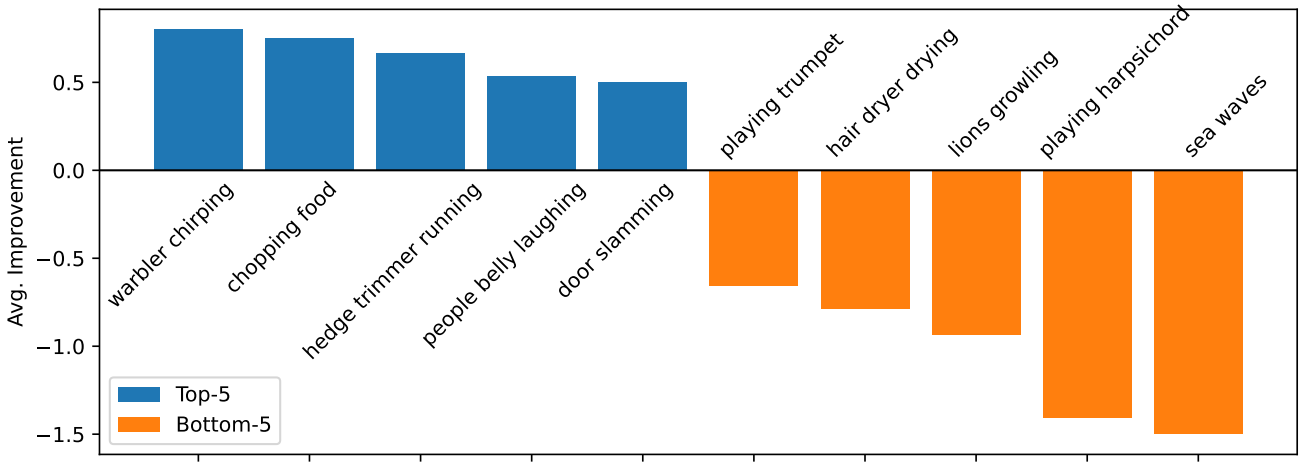

Figure 1. Improvement results of Top-5 and Bottom-5 categories.

Meanwhile, the Synchformer is used to extract high-frame-rate video and audio features and then compute the similarity along the temporal direction, yielding audio-video temporal alignment score. Afterwards, these two scores are weighed summed to form the final reward score for the best-of-N selection for inferred polices. To evaluate the impact of different weight ratios, we conduct a comparison experiment with various reward ratios like $1:1$, $1:2$, and $2:1$ as shown in the Table 4. We can observe that different weight ratios present a minor performance variation, and the ratio $1:1$ achieves the best score across all evaluation metrics. It shows that both ImageBind and Synchformer rewards contribute equally to synthesizing high-quality audio-video preference data.

6. Category-Level Analysis of STAR-DPO

We further analyze the effect of STAR-DPO across different video categories and failure modes. STAR-DPO improves temporal alignment in approximately ∼70% of the evaluated classes. As shown in Figure 1, the improvements are more pronounced in action-driven events with clear temporal cues (e.g., door slamming or clapping), whereas categories dominated by background sounds or off-screen sources remain more challenging.

7. Subjective Evaluation

We conduct the human evaluation for the subjective performance of our model with five comparison models, such as Seeing&Hearing [17], FoleyCrafter [18], VATT [1], Frieren [16] and MMAudio [3]. We sample 10 videos from the VGGSound testing set and feed them into each model to generate corresponding audio, resulting in 60 video-aligned audio samples. For each audio-video pair, we ask volunteers to rate it based on audio quality, audio-video semantic consistency, and audio-video synchronization. For each evaluation dimension, we will give specific instructions for volunteers and define the scale from 1 to 5 to represent strongly disagree, disagree, neutral, agree, and strongly agree, respectively. The given instructions are presented as follows:

(a) **Audio quality:** Please rate audio quality based on whether it is noisy, unclear, or muffled. Note that, please ignore the visual information and only focus on the audio.

(b) **Audio-video semantic alignment:** Please judge if the audio effects are likely to occur in the scenario depicted by the video based on semantic coherence.

(c) **Audio-video temporal alignment:** Please judge if the audio effects are delayed/advanced/synchronized compared to the video, or if audio events happen at the wrong time

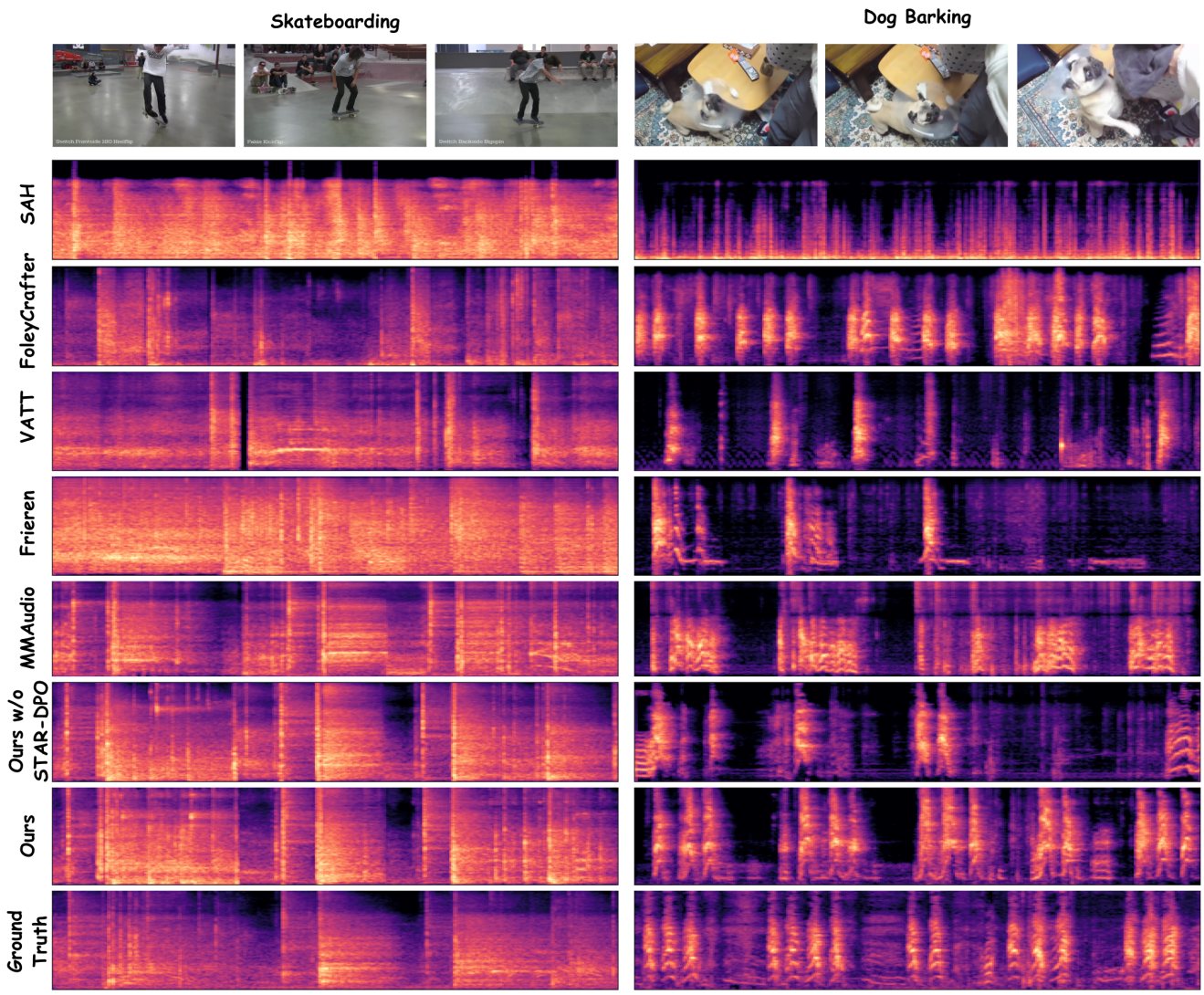

Figure 2. More visualizations of generated audio spectrograms on **VGGSound test subset**.

8. Limitation and Future Work

Although our VisioSonic can synthesize off-screen sounds by modifying language captions, it remains challenging to generate high-fidelity and nuanced off-screen audio. To address this limitation, it is essential to curate a mixed dataset containing both on-screen and off-screen audio-video pairs for training. Furthermore, our proposed STAR-DPO leverages pre-trained ImageBind and Synchformer as reward models to construct positive and negative audio-video pairs. While this approach is computationally efficient and requires no manual annotations, it remains suboptimal for selecting audio-video pairs that align well with human preferences. In future work, we plan to design a more effective audio-video reward model and train it from scratch using human-annotated preference data. Lastly, our current model primarily focuses on synthesizing video-based sounds, without fully considering other audio types such as speech and music. We aim to extend VisioSonic to jointly support the generation of any audio sources, including sound effects, speech, and music.

9. More Visualizations

We present more visualizations of the generated audio spectrograms on the VGGSound test subset, as shown in Figure 2 and Figure 3. We also show qualitative visualizations on out-of-domain datasets, including MovieGen Audio Bench [14] in Figure 4, Sora videos [2] in Figure 5, and Hunyuan videos [7] in Figure 6.

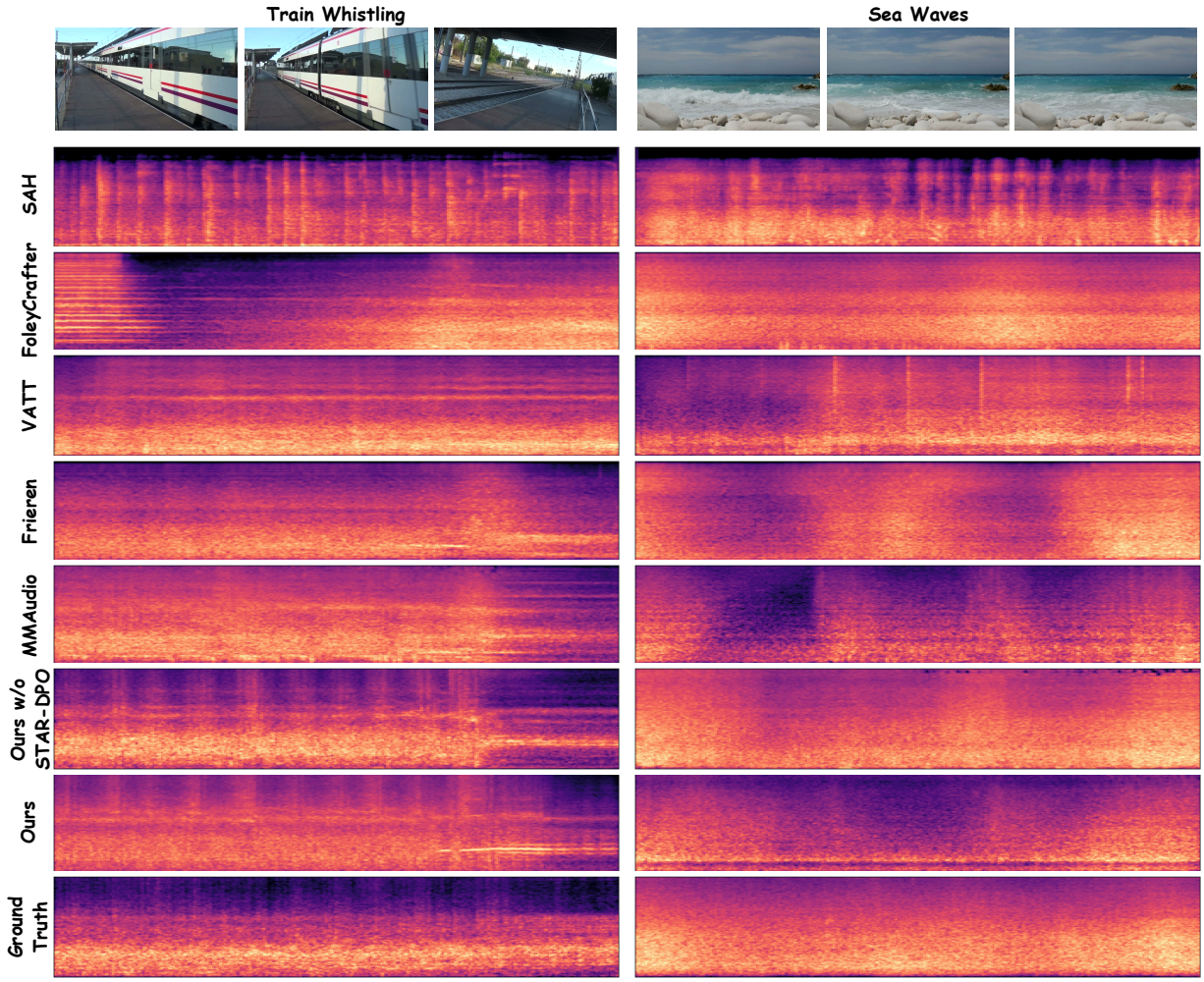

Figure 3. More visualizations of generated audio spectrograms on **VGGSound test subset**.

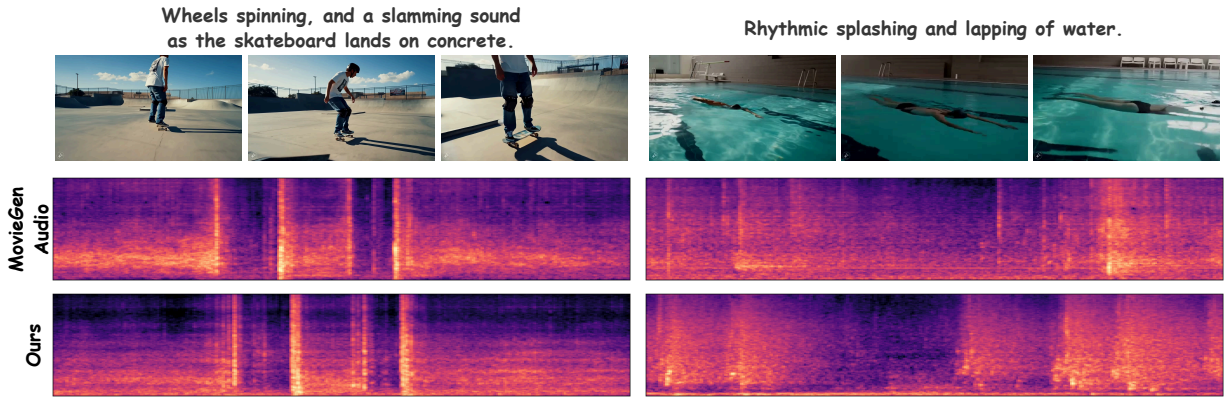

Figure 4. Visualization of generated audio spectrograms on **MovieGen Audio Bench**.

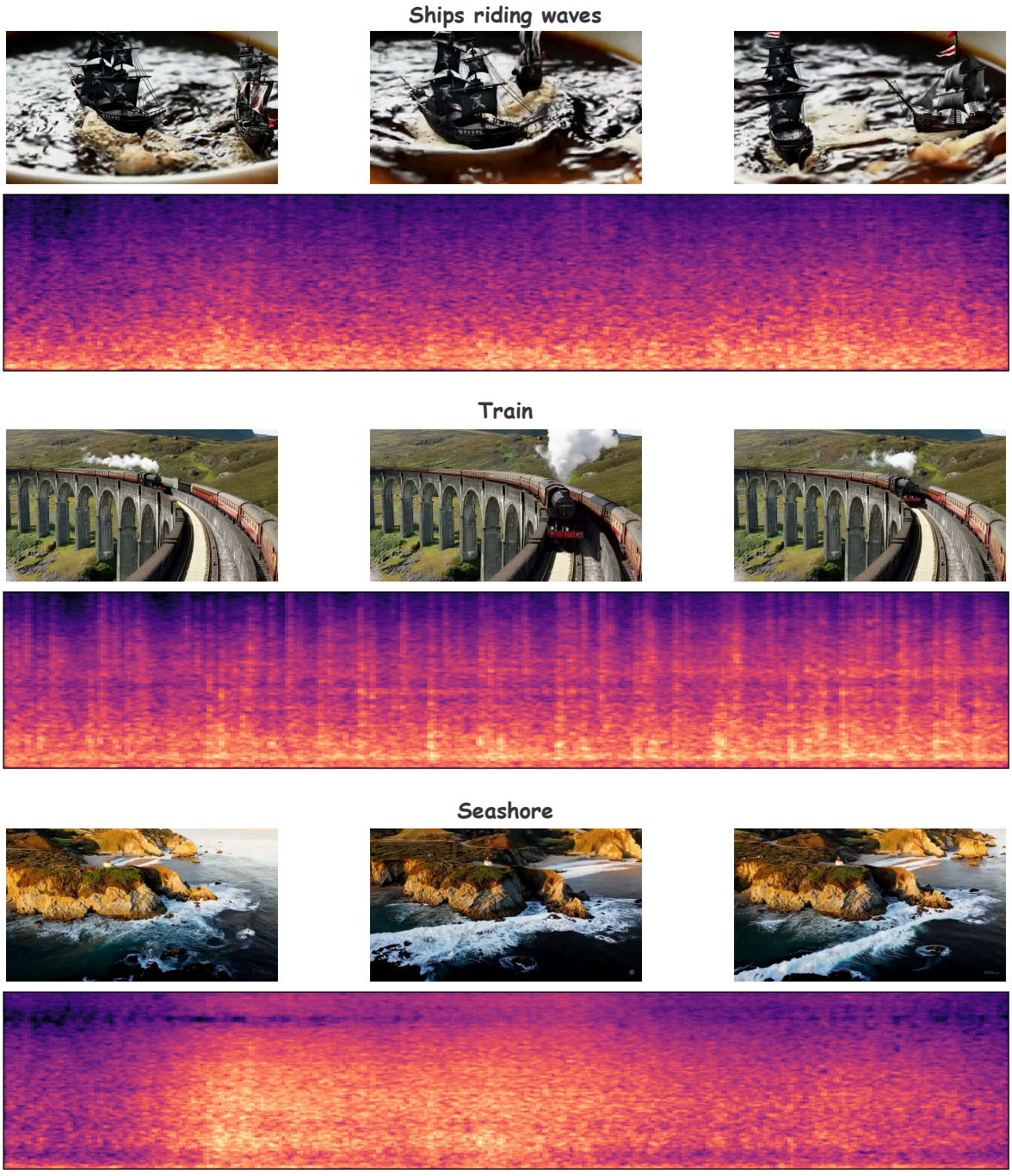

Figure 5. Visualization of generated audio spectrograms on **Sora** videos.

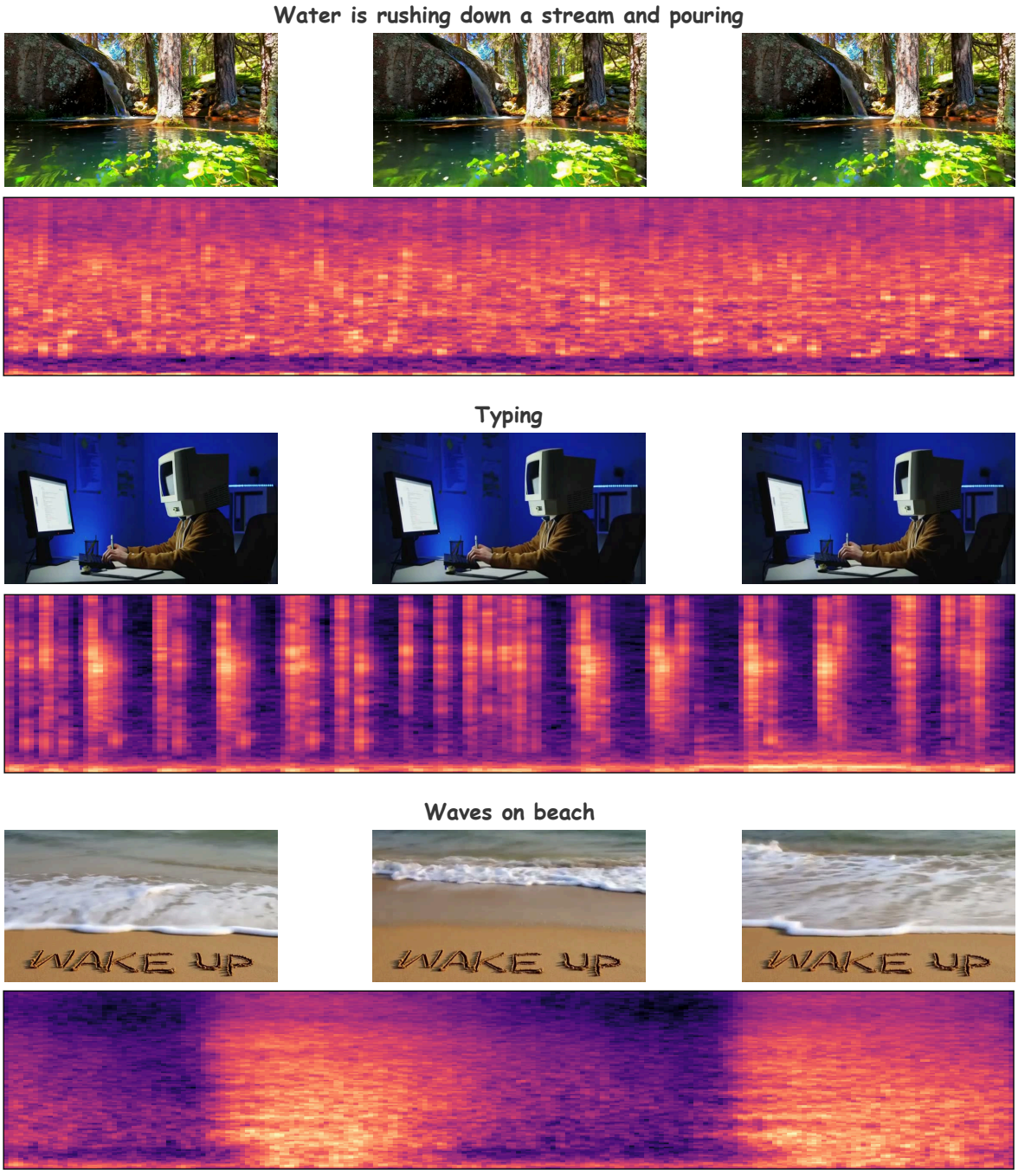

Figure 6. Visualization of generated audio spectrograms on **Hunyuan** videos.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021. 2

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 3

[3] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911, 2025. 1, 2

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1

[5] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Rongjie Huang, Shijie Geng, Renrui Zhang, et al. Lumina-t2x: Scalable flow-based large diffusion transformer for flexible resolution generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[6] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024. 1

[7] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[8] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[9] Dongyang Liu, Shicheng Li, Yutong Liu, Zhen Li, Kai Wang, Xinyue Li, Qi Qin, Yufei Liu, Yi Xin, Zhongyu Li, et al. Lumina-video: Efficient and flexible video generation with multi-scale next-dit. *arXiv preprint arXiv:2502.06782*, 2025. 1

[10] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1

[11] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 1

[12] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 1

[13] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[14] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 3

[15] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[16] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in neural information processing systems*, 37:128118–128138, 2024. 2

[17] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161, 2024. 2

[18] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026. 2

[19] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 1

[20] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024. 1