# LD-LAudio-V1: Video-to-Long-Form-Audio Generation Extension with Dual Lightweight Adapters

Anonymous ICCV submission

Paper ID 12

## Abstract

*Generating high-quality and temporally synchronized audio from video content is essential for video editing and post-production tasks, enabling the creation of semantically aligned audio for silent videos. However, most existing approaches focus on short-form audio generation for video segments under 10 seconds or rely on noisy datasets for long-form video-to-audio synthesis. To address these limitations, we introduce LD-LAudio-V1, an extension of state-of-the-art video-to-audio models and it incorporates dual lightweight adapters to enable long-form audio generation. In addition, we release a clean and human-annotated video-to-audio dataset that contains pure sound effects without noise or artifacts. Our method significantly reduces splicing artifacts and temporal inconsistencies while maintaining computational efficiency. Compared to direct fine-tuning with short training videos, LD-LAudio-V1 achieves significant improvements across multiple metrics:* $FD_{passt}$ *450.00* $\rightarrow$ *327.29 (+27.27%),* $FD_{panns}$ *34.88* $\rightarrow$ *22.68 (+34.98%),* $FD_{vgg}$ *3.75* $\rightarrow$ *1.28 (+65.87%),* $KL_{panns}$ *2.49* $\rightarrow$ *2.07 (+16.87%),* $KL_{passt}$ *1.78* $\rightarrow$ *1.53 (+14.04%),* $IS_{panns}$ *4.17* $\rightarrow$ *4.30 (+3.12%),* $IB_{score}$ *0.25* $\rightarrow$ *0.28 (+12.00%),* $Energy\Delta 10ms$ *0.3013* $\rightarrow$ *0.1349 (+55.23%),* $Energy\Delta 10ms$*(vs.GT) 0.0531* $\rightarrow$ *0.0288 (+45.76%), and* $Sem.Rel.$ *2.73* $\rightarrow$ *3.28 (+20.15%). Our dataset aims to facilitate further research in long-form video-to-audio generation and is available at* https://github.com/deepreasonings/long-form-video2audio.

## 1. Introduction

Video-to-audio (V2A) synthesis, commonly known as Foley sound generation, represents a fundamental challenge in multimedia content creation that aims to generate semantically and temporally aligned audio for silent videos [20]. This task demands not only understanding visual semantics and their relationship with audio to produce contextually appropriate sounds, but also ensuring precise temporal synchronization, as humans are highly sensitive to audio-visual misalignments as subtle as 25 milliseconds [10, 11]. While recent advances in V2A synthesis have demonstrated capabilities in generating high-quality audio for short video segments [3, 14, 19, 20], existing approaches perform short-time audio generation, typically within 10 seconds or less. Current V2A models have limitations when extending to longer temporal contexts. Commonly used approaches can be categorized into two main categories: autoregressive generation methods that produce audio in a sequential token-by-token manner [12, 18, 20], and diffusion-based models with fixed-length denoising processes [14, 19]. However, both categories contain challenges when adapted for long-time synthesis, particularly in maintaining accurate alignment between semantic and temporal domains when guided by visual information.


Figure 1. Long-form video-to-audio (V2A) generation with dual lightweight adapters.

Recent research has attempted to address long-form audio generation challenges through various approaches, including diffusion transformer-based (DiT) architectures [4] and multi-agent systems [21]. Yet, these models have not been specifically designed to reduce splicing artifacts and temporal inconsistencies, nor do they demonstrate capabil-

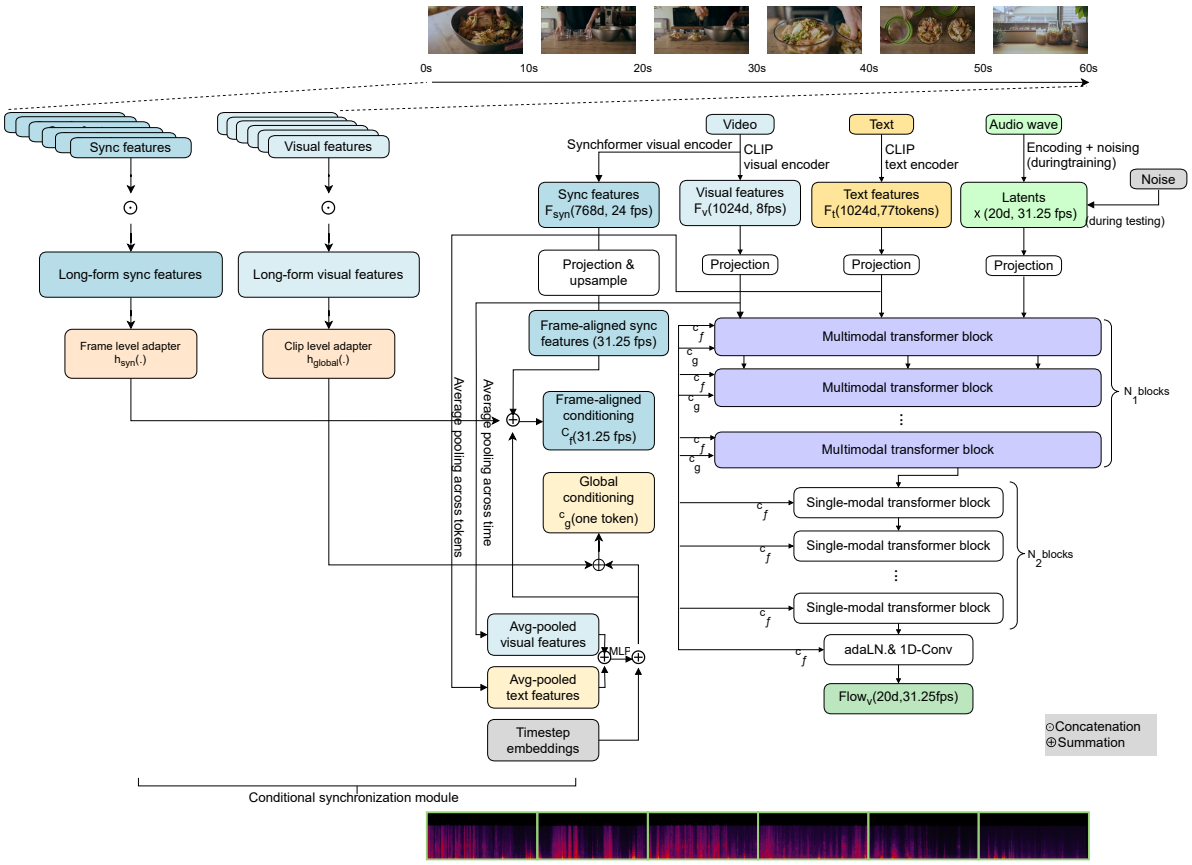


Figure 2. The framework details of LD-LAudio-V1 for long-form V2A generation.

ity for high-quality generation across both short-form and long-form scenarios. Furthermore, high-quality long-form V2A datasets are lacking, with some being closed-source or containing voice and music noise [7].

To address these limitations in long-form V2A generation, we first develop a high-quality V2A dataset containing pure audio effects that are human-annotated and free from voice and music noise. Second, we propose LD-LAudio-V1, a data-driven approach that extends short-form V2A synthesis models through dual lightweight adapters specifically designed to reduce temporal inconsistencies and splicing artifacts in long-form audio generation. As shown in Figure 1, existing models segment videos, while ours uses dual lightweight adapters (frame- and clip-level).

## 2. Related Work

### 2.1. Video-to-Audio

V2A approaches can be categorized into autoregressive and diffusion-based categories. Autoregressive methods generate audio tokens sequentially [8, 15, 17], which are then decoded into audio signals. Latent diffusion and flow matching techniques have substantially enhanced Foley production quality and efficiency [14, 19]. Recent works of Multi-Foley [2] combines mask denoising with reference audio for multi-modal control. MMAudio [3] utilizes a multi-modal transformer with flow matching and synchronization modules for enhanced temporal alignment. These studies inadequately address modality differences between audio and video and lack reasoning guidance. The newer work uses Factorized Contrastive Learning [13] to enhance cross-domain alignment and propose a Chain-of-Thought (CoT)-like V2A approach, which facilitates both general V2A (VGGSound [1]) and professional V2A (piano performance) through step-by-step guidance.

### 2.2. Long-Form Video-to-Audio

For long-form V2A synthesis, segmentation approaches like MMAudio divide extended videos into shorter clips for independent processing. These approaches contain tempo-

ral inconsistencies and loss of global context [17]. Diffusion transformers such as LoVA [4] demonstrate minute-level synthesis without explicit segmentation [2, 7], and they demonstrate comparable performance in the comparison with the state-of-the-art models under heavy parameters trained with non-quantitation evaluation of inconsistency. A multi-agent based long-form V2A is developed with professional dubbing workflows through collaborative role specialization [21]. Temporal fusion module emerges as a computationally efficient approach for long-form V2A generation [5]. Similarly, Omni-based transformers are developed for end-to-end and fast V2A [6]. However, these long-form video-to-audio models do not demonstrate effectiveness in reducing inconsistency when generating long-form pure sound effects without noise. We curate a long-form V2A dataset and propose an extended model for generating multi-clip long-form V2A.

# 3. Methods

The long-form V2A task aims to generate an audio sequence $a$ of equivalent duration from a long video $v$. We propose LD-LAudio-V1, which extends the state-of-the-art models with dual lightweight adapters to handle multi-clip coherence. Our framework is shown in Figure 2. It processes long-form input multi-clip sequences to extract global features and fuses them with short-form features to increase coherence in multi-clip audio generation.

## 3.1. Feature Representation

We represent all features as one-dimensional tokens without using absolute position encoding, which allows generalization to different durations at test time. Visual features are extracted at 8 fps as 1024-dimensional features, and text features consist of 77 tokens as 1024-dimensional features, both extracted from CLIP [16]. Audio latents exist in variational autoencoder (VAE) latent space at 31.25 fps as 20-dimensional latents by default. Synchronization features are extracted with the Synchformer tool [9] at 24 fps as 768-dimensional features. All features follow the same temporal ordering at different frame rates and are projected to a hidden dimension after initial processing layers.

## 3.2. Frame-Level Synchronization Module

For a sequence of video frames, we first extract per-frame visual features using a pre-trained vision encoder. Synchformer at 24 fps as 768-dimensional features. This synchronization module processes frame-level visual features to generate fine-grained temporal conditioning signals that are closely aligned with the temporal dynamics of audio. The Synchformer extracts features $F_{syn}$ and then performs projection and up-sampling to frame-aligned sync features $F_{syn}^{frame}$ with 31.25 fps sampling rate.

## 3.3. Clip-Level Contextualization Module

We integrate a clip-level contextualization module to produce a global semantic representation of the video. Features from the clip visual encoder $F_v$ and clip text encoder $F_t$ are first projected, then averaged and concatenated as $F_g^{con}$. The $F_g^{con}$ is fused with timestamp embeddings to produce global conditioning features $c_g$ that capture the semantic context of the video content.

## 3.4. Dual Lightweight Adapters for Coherence

We extend short-form V2A to long-form audio generation through a multi-clip coherence extension module. This module applies the dual lightweight adapters, denoted as $h_{syn}$, $h_{global}$, to capture long-form consistency of audio from videos and then fuse them with the short-form videos.

- **Light-weight Dual Adapters.** The long-form input multi-clip sequence from a video: $\{V_{clip}^{(i)}\}_{i=1}^{L}$ and is processed as a union long-form video from the same clip visual encoder and text encoder to extract the global visual features $F_v^{global}$, global text features $F_t^{global}$, and global synchronization features $F_{syn}^{global}$. These global features are processed similarly to $F_v$, $F_t$, and $F_{syn}$ to obtain global-level (multi-clip) conditions $c_g^{global}$ and local-level conditions $c_f^{global}$ after projection and average pooling (See Figure 2).
- **Fusion.** For inference of audio for each video clip $V_i$, the final conditions are computed by combining global and local features through the lightweight adapters: $c_g^{final} = c_g + h_{global}(F_v^{global})$, $c_f^{final} = c_f + h_{syn}(F_{syn}^{global})$.

## 3.5. A Unified Multi-Modal Synthesis Transformer

The same multi-modal transformer architecture used for short-form generation is applied to generate audio using the combined final conditions: $c_g^{final}$ and $c_f^{final}$.

$$x^{(l+1)} = \text{DiT}^{(l)}\left(x^{(l)},\ c_g^{final},\ c_f^{final}\right), \quad \text{for } l = 1, 2, \ldots, L \tag{1}$$

where $x^l$ is the input of the $l_{th}$ transformer layers.

# 4. Experiments

## 4.1. Datasets

We curate the first version of a long-form clean sound effects dataset, denoted as LPSE-1, to support the study of long-form audio generation from videos. Different from the previous audio-visual event datasets (AVE) dataset, our LPSE-1 consists of more than $6K$ videos with over $20K$ audio-visual events covering 120 different event categories. Each clip is more than 60 seconds, containing real-life audio-visual scenes. Unlike other AVE datasets that contain noise such as voice-over [1] or other audio types such as

| Method | $FD_{\text{passt}}\downarrow$ | $FD_{\text{panns}}\downarrow$ | $FD_{\text{vgg}}\downarrow$ | $KL_{\text{panns}}\downarrow$ | $KL_{\text{passt}}\downarrow$ | $IS_{\text{panns}}\uparrow$ | $IB_{\text{score}}\uparrow$ | DeSync↓ | Energy$\Delta$10ms↓ | Energy$\Delta$10ms(vs.GT)↓ | Sem.Rel.↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GT | \ | \ | \ | \ | \ | \ | \ | \ | 0.1103 | 0 | \ |
| MMAudio-L-44.1kHz Zeroshot | 455.49 | 33.04 | 2.28 | 2.87 | 2.32 | **4.65** | 0.27 | 1.44 | 0.3629 | 0.0524 | **3.77** |
| MMAudio-L-44.1kHz Finetuned | 450.00 | 34.88 | 3.75 | 2.49 | 1.78 | 4.17 | 0.25 | **1.38** | 0.3013 | 0.0531 | 2.73 |
| MMAudio-L-44.1kHz Long-form | 327.29(−27.27%) | 22.68(−34.98%) | 1.28(−65.87%) | 2.07(−16.87%) | 1.53(−14.04%) | 4.30(+3.12%) | 0.28(+12.00%) | 1.51(+9.42%) | 0.1349(−55.23%) | 0.0288(−45.76%) | 3.28(+20.15%) |

Table 1. Results of long-form V2A generation with dual adapters compared to baselines.

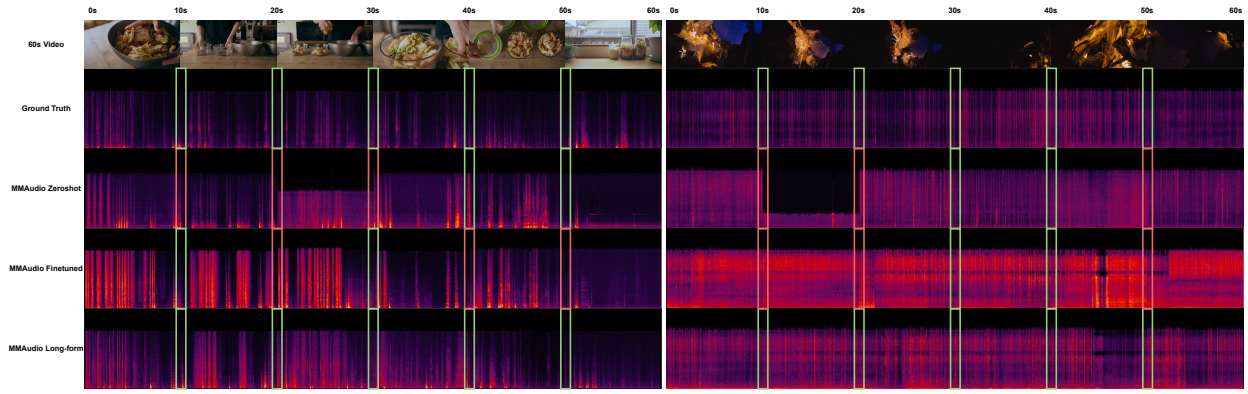


Figure 3. Examples of long-form V2A generation with different experimental settings.

music or speech, our long-form dataset contains pure sound effects. Each sample is manually human verified to ensure it contains only sound effects without other types of audio or irrelevant events from visual scenes.

## 4.2. Evaluation Metrics

### 4.2.1. Quality Metrics

Metrics for evaluating the quality of V2A are in four aspects, such as distribution matching, audio quality, semantic alignment, and temporal alignment [20]. Specific metrics inlucde: $FD_{passt}\downarrow$, $FD_{panns}\downarrow$, $FD_{vgg}\downarrow$, $KL_{panns}\downarrow$, $KL_{passt}\downarrow$, $IS_{panns}\uparrow$, $IB_{score}\uparrow$, $DeSync\downarrow$ [3].

### 4.2.2. Multi-clip Consistency Metrics

Additionally, we apply consistency metrics specifically for long-form V2A. These metrics include the average energy change within 10 ms before and after each segmentation point between two short-form video clips ($Energy\Delta 10ms\downarrow$), differences between the average energy change of the generated audio and ground truth ($Energy\Delta 10ms(vs.GT)\downarrow$), and Semantic Relevance ($Sim.Rel.\uparrow$) [4].

## 4.3. Initial Benchmark Results

We compare our long-form model against the zero-shot MMAudio-L-44.1kHz model and a model finetuned on short training videos. The results are presented in Table 1, with overhead costs in Table 2.

Our long-form model demonstrates significant improvements across multiple metrics compared to the short training videos finetuned model. Specifically, $FD_{passt}$ 450.00 to 327.29(+27.27%), $FD_{panns}$ 34.88 to 22.68(+34.98%), $FD_{vgg}$ 3.75 to 1.28(+65.87%), $KL_{panns}$ 2.49 to 2.07(+16.87%), $KL_{passt}$ 1.78 to 1.53(+14.04%), $IS_{panns}$ 4.17 to 4.30(+3.12%), $IB_{score}$ 0.25 to 0.28(+12.00%), $Energy\Delta 10ms$ 0.3013 to 0.1349(+55.23%), $Energy\Delta 10ms(vs.GT)$ 0.0531 to 0.0288(+45.76%), and $Sem.Rel$ 2.73 to 3.28(+20.15%). In Figure 3, we present several generated examples from different experimental settings.

| Method | Params | Inference time of clip(60s) |
|---|---|---|
| MMAudio-L-44.1kHz | 1.03B | 61.27s |
| MMAudio-L-44.1kHz Finetuned | 1.03B | 61.27s |
| MMAudio-L-44.1kHz Long-form with dual adapters | 1.07B | 62.75s |

Table 2. Computational costs of parameters and inference time.

## 5. Conclusion

We propose LPSE-1, a clean long-form V2A dataset of $6k$ video clips with each duration of $60s$ and $24k$ audio-visual events. Human annotators manually verify that the videos contain pure sound effects without noise or irregular music. We also propose LD-LAudio-V1, which extends the state-of-the-art short-form (10 seconds) V2A models to long-form (60 seconds) audio generation using dual lightweight adapters. Our approach reduces splicing artifacts and temporal inconsistencies in long-form V2A. Our approach adds only 4% more parameters while enabling effective long-form V2A generation from existing short-form models.

## References

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset.

In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 3

[2] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. *arXiv preprint arXiv:2411.17698*, 2024. 2, 3

[3] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*, 2024. 1, 2, 4

[4] Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, and Ruihua Song. Lova: Long-form video-to-audio generation. *arXiv preprint arXiv:2409.15157*, 2024. Long-form Video-to-Audio Generation. 1, 3, 4

[5] Alex Ergasti, Giuseppe Gabriele Tarollo, Filippo Botti, Tomaso Fontanini, Claudio Ferrari, Massimo Bertozzi, and Andrea Prati. Rflav: Rolling flow matching for infinite audio video generation. *arXiv preprint arXiv:2503.08307*, 2025. 3

[6] Zhengcong Fei, Hao Jiang, Di Qiu, Baoxuan Gu, Youqiang Zhang, Jiahua Wang, Jialin Bai, Debang Li, Mingyuan Fan, Guibin Chen, and Yahui Zhou. Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers. *arXiv preprint arXiv:2506.00830*, 2025. 3

[7] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951, 2023. 2, 3

[8] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021. 2

[9] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE, 2024. 3

[10] A. Sophia Koepke, Olivia Wiles, Yael Moses, and Andrew Zisserman. Sight to sound: An end-to-end approach for visual piano transcription. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1838–1842, 2020. 1

[11] Jangwon Lee, Bardia Doosti, Yupeng Gu, David Cartledge, David J. Crandall, and Christopher Raphael. Observing pianist accuracy and form with computer vision. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1505–1513, 2019. 1

[12] Bingliang Li, Fengyu Yang, Yuxin Mao, Qingwen Ye, Hongkai Chen, and Yiran Zhong. Tri-ergon: Fine-grained video-to-audio generation with multi-modal conditions and lufs control. *arXiv preprint arXiv:2412.20378*, 2024. 1

[13] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[14] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023. 1, 2

[15] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[17] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. *arXiv preprint arXiv:2409.13689*, 2024. 2, 3

[18] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models, 2023. 1

[19] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*, 2024. 1, 2

[20] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. 1, 4

[21] Yehang Zhang, Xinli Xu, Xiaojie Xu, Li Liu, and Yingcong Chen. Long-video audio synthesis with multi-agent collaboration, 2025. 1, 3