LoVA: Long-form Video-to-Audio Generation

Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, Ruihua Song[*]
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing 100872, China
{chengxin000, xihuaw, yihanwu, wangyuyue123, rsong}@ruc.edu.cn

Abstract

Video-to-audio (V2A) generation is important for video editing and post-processing, enabling the creation of semantics-aligned audio for silent video. However, most existing methods focus on generating short-form audio for short video segment (less than 10 seconds), while giving little attention to the scenario of long-form video inputs. For current UNet-based diffusion V2A models, an inevitable problem when handling long-form audio generation is the inconsistencies within the final concatenated audio. In this paper, we first highlight the importance of long-form V2A. Besides, we propose LoVA, a novel model for **Lo**ng-form **V**ideo-to-**A**udio generation. Based on Diffusion Transformer (DiT) architecture, LoVA proves to be more effective at generating long-form audio compared to existing autoregressive models and UNet-based diffusion models. Extensive experiments demonstrate that LoVA achieves comparable performance on 10-second V2A benchmark and outperforms all other baselines on a benchmark with long-form video input.

1 Introduction

Video-to-Audio (V2A) generation, which aims to create synchronized and realistic sound effects for silent videos, finds widespread use in video editing, sound effect creation, and autonomous content enhancement [1]. However, current V2A methods predominantly focus on generating fixed-length short audios, typically less than 10 seconds on VGGSound [3] or Audioset [8] benchmarks. These methods generate fixed-length audios through autoregressive approaches truncated to a maximum length [14, 22, 27], or fixed-length noise denoising by UNet-based diffusions [21, 30, 29]. Despite their success in generating fixed-length short audios, the challenge of creating audio for variable-length, long-form videos exceeding 10 seconds in real-world scenarios remains unexplored. Our work aims to address this short-to-long duration gap in the V2A domain.

When adapted to long-form V2A, current autoregressive and UNet-based diffusion models both exhibit limitations. As depicted in Figure 1(b): (1) *Autoregressive models* regard audio as a series of audio frames (i.e., tokens). Ideally, they can generate an infinite number of audio frames without truncation. However, this one-by-one generation leads to low efficiency for long audio sequences. It also yields lower audio quality compared to diffusion models due to frame discretization [29]. (2) *UNet-based diffusion models* struggle with long-range relation modeling, with generation performance being constrained by the length of the training data [13], a limitation confirmed by prior studies [4, 31, 10] and our experimental results (Section. 4.2). To better accommodate long-form V2A, these models split long videos into shorter clips, equivalent to their pretraining data length, generate audio for each clip, and then concatenate these to form the final long audio. However, such splitting process can result in inconsistencies, i.e., with distinct sounds from the same video. This is evident in Figure 1(a) with results from DiffFoley [21], TiVA [29], and FoleyCrafter [30], where short 8s/10s audio clips

[*]Corresponding authors

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

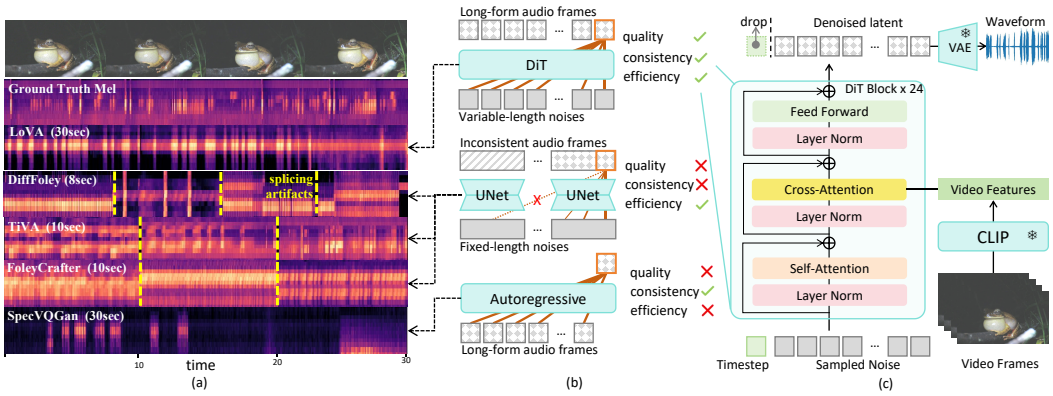

Figure 1: (a) Long-form V2A example. Current (8s/10s) UNet-based diffusion V2A models (DiffFoley, TiVA, FoleyCrafter) exhibit inconsistency when generating long-form (30s) audio, as indicated by clear mel-spectrogram boundaries and structural variances. In contrast, our LoVA produces consistent results akin to the ground truth. (b) Comparison of three distinct long-form V2A methods. From bottom to top: autoregressive methods, UNet-based diffusions, DiT-based diffusions (our LoVA), characterized by inefficient one-by-one generation manner, inconsistent fixed-length splits generation, and our parallel processing of arbitrary-length sequences respectively. (c) Overview of LoVA. Capable of accepting videos of any length, it samples and denoises on the corresponding length of the latent noise sequence and then decodes it to generate audio of any length.

exhibit clear mel-spectrogram boundaries and structural differences, thereby reducing the quality of the concatenated long audio. Thus, balancing efficiency, consistency, and quality in long-form V2A remains a significant challenge for existing methods.

To address the aforementioned challenge, we introduce LoVA, a **Lo**ng-form **V**ideo-to-**A**udio generation model adept well at handling long-duration problem. As depicted in Figure 1(b), the expected long-form V2A model should possess the capabilities of: (1) maintaining the variable-length audio as a sequence of lossless frames to ensure quality; (2) modeling the full sequence interactions among frames, rather than the localized interactions learned by convolutional UNets, to ensure consistency when extending to long sequences; (3) generating multiple frames in parallel for efficiency.

Diffusion Transformer (DiT)[23] treats latent data as token sequences in the diffusion process, aligning well with the sequential nature of audio. It also demonstrated promising results in generating high-quality images[23], videos [2], and audios [6, 7] in an efficient parallel sequence generation manner. Thus, we introduce DiT into the V2A domain and model the denoising process on noisy latent audio frames, termed as LoVA. For long-form V2A problems, LoVA simply extracts extended video features and prepares correspondingly longer sequences of noisy audio frames for denoised lengthy audio generation, akin to a consistent frog croak over a 30s video as depicted in Figure 1(a).

Furthermore, due to the absence research in the long-form V2A domain, we have established a long-form V2A evaluation based on a variable-length long video dataset UnAV100 [9], as an addition to the current standard short-form evaluation. We conducted extensive experiments on this long-form evaluation to validate the performance of autoregressive, UNet-based diffusion, and DiT-based diffusion (our LoVA) methods, as well as their duration-extending characteristics in long-form V2A. Overall, our main contributions are as follows:

- We first introduce the long-form generation problem in the V2A field and establish an evaluation framework for long-form V2A as a complement to existing V2A evaluations.
- We first employ DiT into V2A area and propose a new model, LoVA, which is better suited for generating long-form audio than existing methods.
- We conducted extensive experiments on standard short-form and newly established long-form evaluation, validating the SOTA results achieved by LoVA. We also draw some duration-extending characteristics for different V2A methods. Demo samples for different V2A methods are available at https://ceaglex.github.io/LoVA.github.io/.

2 Method

The long-form V2A task aims to generate an audio sequence a of equivalent duration from any given long video v. We introduce LoVA, a Latent Diffusion Transformer designed for this task. As depicted in Figure 1(c), LoVA preprocesses long-form video into features, applies denoising on a noise sequence of corresponding length, and eventually generates long-form audio through VAE decoding. We will sequentially elucidate the preliminary knowledge of Latent Diffusion Model (LDM) [25], the Architecture of LoVA and its training in the following subsections.

2.1 Preliminary: V2A LDMs

Given audio-video pairs (a, v), the typical V2A LDM compresses a into latent variables z (i.e.,z_0) using a VAE encoder , and encodes v into conditional video features c. A diffusion process then introduces Gaussian noise ϵ to the clean latent z_0 based on timestep t and predefined noise schedule $\bar{\alpha_1}, ..., \bar{\alpha_t}, ..., \bar{\alpha_T} : z_t = \sqrt{\bar{\alpha_t}} z_0 + \sqrt{1 - \bar{\alpha_t}}\epsilon, \epsilon \sim N(0, 1)$.

During the denoising process, LDM aims to recover z_0 from z_T by progressively estimating the added noise at each timestep t, given the condition c and input noisy data $z_t : \hat{\epsilon}_t = D(z_t, c, t)$.

The training objective is to minimize the L2 loss between the added noise ϵ and the predicted noise $\hat{\epsilon}_t$ at each step $t : \mathcal{L} = ||\hat{\epsilon}_t - \epsilon||^2$.

2.2 Architecture of LoVA

As shown in Figure 1(c), LoVA has three components: an audio VAE V, a video encoder CLIP and a DiT-based denoiser D.

(1) Audio VAE: LoVA employs a 1D-Conv-based VAE [7, 13] to compress the audio waveform $a \in [n, T]$, where n and T are the audio channels and time length. The resultant latent data is $z_0 = V(a) \in [n, T', h]$, with T' and h denoting the compressed time length and latent space size.

(2) Video Encoder: Numerous previous works [27, 29] have demonstrated the effectiveness of CLIP [24] in V2A task. For a video composed of a sequence of frames $v : [f_1, \ldots, f_i, \ldots, f_N]$, LoVA also uses the CLIP visual encoder to extract features from each video frame and concatenate them to form the video condition $c : c = \mathrm{Concat}([\mathrm{CLIP}(f_1), \ldots, \mathrm{CLIP}(f_i), \ldots, \mathrm{CLIP}(f_N)]) \in [N, h_C]$, where N is the frame number and h_C is the CLIP hidden size.

To further assist the model in learning the relationship between the audio sequence and video frames when extending to long-form V2A generation, we add a learnable positional embedding layer PE_c to the video condition $c : c = c + PE_c([1, ..., i, ..., N])$. Finally, the video frame features c are fed into the DiT denoiser as a visual condition to guide the denoising process.

(3) DiT Denoiser: Diffusion Transformer (DiT) [23] is a novel diffusion structure that integrates the denoising diffusion models [12] with the Transformer architecture [28]. In LoVA's DiT denoiser, before being fed into each DiT block, the noised audio latent sequence z_t is first processed with an additional positional embedding layer $PE_z : z_t = z_t + PE_z([1, ..., i, ..., T'])$.

Timestep t is embeded and appended at the beginning of the input sequence. Conditional input c is processed by cross-attention layers in DiT block. Finally, conditioned on timestep t and video features c, DiT denoiser takes z_t as input tokens to estimate noise at each timestep.

2.3 Training and Inference of LoVA

In the optimization phase of LoVA, the Audio VAE and Video Encoder are maintained frozen, as per [7, 24]. The DiT Denoiser, including all blocks, PE_c, PE_z, and time embeddings, undergoes training. The training is governed by the L2 Loss. During inference, LoVA can accommodate videos of arbitrary lengths, handling variable-length video conditions and noisy latent sequences through the extension of PE_c and PE_z. Finally, variable-length audio is obtained through VAE decoding.

3 Experimental Settings

3.1 Implementation Details

We implement a two-phase training approach: *pre-training* with short-form data and then *fine-tuning* with long-form data, referred to as **LoVA (w/o tuning)** and **LoVA (w/ tuning)** respectively. LoVA (w/o tuning) utilizes AudioSet-balanced [8] and VGGSound [3] datasets, encompassing 20,280 and 180,379 10-second videos respectively. LoVA (w/ tuning) employs the UnAV100 dataset [9], made up of 6,489 videos ranging from 10 to 60 seconds. More training details can be found at A.1. To assess LoVA's performance against baselines in short-form V2A generation, we use the VGGSound [3] test set of 15,273 10-second videos. For long-form V2A generation evaluation, we utilize the UnAV100 [9] test set, comprising 2,167 cases with an average duration of 42.1s.

3.2 Baselines

We implement the public code of five baselines to replicate the results.: SpecVQGAN [14], IM2WAV [27], DiffFoley [21], TiVA [29], and FoleyCrafter [30], in which the first two are auto-regressive models while the other three are diffusion-based models. To ensure a fair comparison, we adapt all of them for long-form V2A generation. Modification details can be found at A.2.

3.3 Metrics

Following previous works [14, 19, 20, 21, 27], we apply Fréchet Audio Distance (FAD) [16], Inception Score (IS) [26], and mean KL-divergence (MKL) to evaluate the quality of generated audio. To ensure a fair comparison and eliminate the effect of different sampling rate, we downsample the generated audio from LoVA's to 16kHz and then resample them to the required sampling rate of classifiers (16kHz for VGGish [11], 32kHz for PaSST [18] and PANN [17]). Since these audio classifiers are trained on 10-second audio data, they cannot be directly applied to the evaluation of long-form audio. Thus for the evaluation of long-form V2A, we segment the generated audio into 10-second clips with 5-second overlapping windows. For FAD, we average features from all audio clips to get the final feature of long-form audio. For IS and MKL, following previous works [5], we get the mean results of classification logits and then apply a softmax. Besides, we introduce the number of inferences per audio (Num.Infer.) as a indicator of potential

4 Experimental Results

Table 1: Comparison of LoVA with baselines on VGGSound and UnAV100 benchmark. We employ multiple classifiers to evaluate audio quality ("VGG" denotes VGGish, "PaSST" denotes PaSST, and "PANN" denotes PANN). The best score on each metric is highlighted with bold type and the second best score is in underline.

		VGGSound					UnAV100					
Method	Sampling Rate (kHz)	FAD↓ (VGG)	KL↓ (PANN)	KL↓ (PaSST)	IS↑ (PANN)	IS↑ (PaSST)	FAD↓ (VGG)	KL↓ (PANN)	KL↓ (PaSST)	IS↑ (PANN)	IS↑ (PaSST)	Num. Infer.↓
AutoRegressive												
SpecVQGAN	22.05	6.26	3.16	3.12	4.00	3.77	9.21	2.28	2.17	2.84	2.52	**1.00**
IM2WAV	16	5.77	2.28	2.24	5.77	5.19	6.99	1.10	1.05	4.32	4.28	4.64
Diffusion												
DiffFoley	16	6.10	2.76	2.88	8.12	9.56	7.74	1.22	1.28	4.42	5.02	5.70
FoleyCrafter	16	2.34	2.29	2.28	8.53	9.83	2.82	1.06	**1.05**	6.91	7.47	4.64
TiVA	16	**1.05**	2.13	**2.00**	9.31	8.02	6.36	1.48	1.54	3.11	2.77	4.64
LoVA (w/o tuning)	**44.1**	1.70	**2.06**	2.10	**9.69**	**9.87**	**2.44**	**1.05**	1.06	**7.69**	**7.96**	**1.00**
LoVA (w/ tuning)	**44.1**	1.70	**2.06**	2.10	**9.73**	**9.91**	**2.42**	**1.05**	1.06	**7.71**	**7.96**	**1.00**

4.1 Comparison with SOTA models

To evaluate the performance of LoVA on long-form V2A generation, we compare LoVA with baselines on the UnAV100 dataset. As shown in Table 1, being trained on the same 10-second data without any fine-tuning, LoVA outperforms all other baselines on 4 out of 5 metrics, with the fewest Num.Infer. It proves the effectiveness of DiT model in long-form V2A tasks. Utilizing DiT, LoVA generates high sampling rate audio that is 6 times longer than current UNet-based LDMs. Besides, being fine-tuned on the long-form dataset, LoVA (w/ tuning) achieves the best performance regarding most metrics. We also do extensive subjective experiment to evaluate different models' performance on long-form V2A dataset, the results aligns with the autonomous evaluation and can be found at A.3. This remarkable performance underscores LoVA's superiority in handling long-form V2A generation.

To evaluate the performance of LoVA for the short-form V2A generation, we conduct experiments on the VGGSound test set. As shown in Table 1, LoVA achieves comparable or even better performance to existing state-of-the-art V2A models. Specifically, LoVA achieves the best performance in PANN KL, PANN IS and PaSST IS scores, and the second-best score in FAD and PaSST KL scores. These results indicate that LoVA can generate realistic audio and accurately capture semantic information from the video. In summary, LoVA shows the best results in both short-form and long-form V2A generation across most evaluation metrics.

4.2 Comparison between Different Diffusion Denoiser

We conduct ablation studies to validate the effectiveness of DiT modules when performing long-form V2A generation. Being similar with LoVA, some previous UNet-based diffusion models, like FoleyCrafter [30], can also generate long-form audio by modifying the shape of the latent space. On the UnAV100 benchmark, we cut the video into clips with different durations and generate audio for each clip individually.Then we obtain the long-form audio by concatenating all generated audio clips. FoleyCrafter is adapted to different clip durations by resizing the latent space.

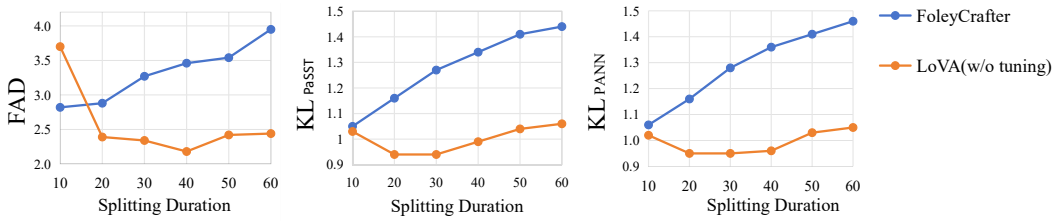

Figure 2: Comparison of long-form audio generation ability between UNet and DiT structure. The experimennt is carried on UnAV100 test dataset. Different splitting duration means different video segment durations, as well as different generated sequence lengths per inference.

As shown in Fig 2, as the clips duration increases, the FAD and KL metrics degenerate. The best scores are achieved when the splitting duration is set to 10 seconds, which aligns with the training data on which FoleyCrafter were trained. However, for DiT-based LoVA (w/o tuning), metrics do not show obvious degenerative phenomenon as the duration increases. It should be stressed that FoleyCrafter and LoVA are all trained on 10-second data only, but performs differently on long-form audio generation. This difference highlights a critical limitation of the UNet structure when it extends to long-form audio, while proves DiT's effectiveness to handle long audio sequence.

5 Conclusion

In this paper, we identify the significant gap between current V2A models and real-world V2A applications, particularly in generating long-form audio. To address this, we introduce a new task termed long-form video-to-audio generation. Besides, we introduce LoVA, a DiT-based V2A generation model, which is tailored for long-form V2A generation tasks. Experimental results indicate that LoVA shows SOTA performance than previous models on both the 10-second VGGSound and long-form UnAV100 benchmarks, excelling in audio quality, sampling rate, and supported duration. In future work, we plan to (1) further explore temporal synchronization between video and audio, and (2) investigate methods for controlling generated audio, including text and duration control.

References

[1] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014.

[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. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

[3] 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.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[5] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

[6] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*, 2024.

[7] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[9] 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.

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[13] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023.

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

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

[16] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

[17] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[18] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

[19] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[20] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[21] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation. *arXiv preprint arXiv:2309.10537*, 2023.

[23] 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.

[24] 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.

[25] 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*, pages 10684–10695, 2022.

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[27] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[28] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[29] Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In *ACM Multimedia 2024*, 2024.

[30] 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.

[31] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

A Appendix

A.1 Implementation Details

Throughout both phases, the weights of Audio VAE and Video Encoder remain frozen [7, 24]. For the pre-training phase, the DiT Denoiser, PE_c, PE_z, and time embedding undergo training. During the fine-tuning phase, updates are only applied to PE_z, PE_c, and the final DiT block. .....Beside, We sample our audio at 44.1kHz and video frames at 8 FPS. In the inference stage, we set the guidance scale to 5.0, and employ the DPM++ 3M SDE sampler [15] to execute denoising over 150 steps.

A.2 Baselines

For the autoregressive baseline SpecVQGAN, we use the long-form video as input, adjust the generated sequence length, and obtain aligned long-form audio. For the three diffusion-based baselines, we divide the original video into short-form fixed-length clips(8 seconds or 10 seconds consistent with their training settings), generate corresponding audio separately, and then concatenate the generated audio segments. It should be mentioned that for IM2WAV we use the same divide-generate-concatenate procedure due to its slow inference speed.

A.3 Subjective Evaluation

We randomly select 40 videos from UnAV100 test set for human evaluation. Evaluators are asked to give a 5-level Likert scale on 4 aspects: Overall quality (Overall), Sound Quality (SoundQua), Semantic Relevance (SemRel) and Consistency. A higher score denote better performance.

Table 2: Subjective evaluation results for different models. Each entry comprises the mean score from 40 evaluators, followed by the 95% confidence interval. Bold typeface indicates the highest score for each metric, while underlined values represent the second-highest scores.

Method	Overall ↑	SoundQua ↑	SemRel ↑	Consistency ↑
DiffFoley	2.79 ± 0.12	2.89 ± 0.14	3.02 ± 0.12	2.94 ± 0.14
FoleyCrafter	2.93 ± 0.13	3.15 ± 0.15	3.06 ± 0.14	3.13 ± 0.15
IM2WAV	$\underline{2.93 \pm 0.10}$	2.81 ± 0.14	$\underline{3.12 \pm 0.10}$	$\underline{3.21 \pm 0.13}$
TiVA	2.76 ± 0.13	$\mathbf{3.50 \pm 0.13}$	2.78 ± 0.11	2.79 ± 0.13
LoVA (w/o tuning)	$\mathbf{3.45 \pm 0.10}$	$\underline{3.42 \pm 0.13}$	$\mathbf{3.55 \pm 0.12}$	$\mathbf{3.81 \pm 0.14}$
LoVA (w/ tuning)	$\mathbf{3.51 \pm 0.11}$	$\underline{3.42 \pm 0.13}$	$\mathbf{3.56 \pm 0.12}$	$\mathbf{3.82 \pm 0.13}$

As shown in table 2, LoVA model achieves best result on human evaluation. It outperforms other models in Semantic Relevance, Consistency, and Overall Quality, while ranking second in Sound Quality. These findings largely corroborate the automatic evaluation results, suggesting that LoVA and its Dit structure are well-suited for long-form V2A tasks.