

Echoes Over Time: Unlocking Length Generalization in Video-to-Audio Generation Models

Supplementary Material

In this supplementary material, we provide the details of experimental settings, our proposed method, and additional results.

1. Datasets and Settings

Training datasets. As stated in the main paper, our primary video-to-audio dataset is VGGSound, which serves as the core resource for training and evaluation. To further enhance the capability of our model, we incorporate additional training using text-to-audio datasets, specifically WavCaps [9] and Clotho [3]. These supplementary datasets provide rich textual descriptions paired with audio, enabling the model to learn from diverse textual cues. It is important to emphasize that, when leveraging these datasets, we only utilize the textual information to complement audio generation, without introducing any extra visual context. This approach ensures that the improvements gained from these resources stem solely from text-based learning rather than multimodal inputs (*i.e.*, visual cues).

Evaluation datasets. For the UnAV100 benchmark, we utilize the official test set provided by UnAV100 in its original form, without introducing any modifications. During the evaluation phase, captions are deliberately withheld for all instances within this set. This design ensures that the task remains strictly focused on video-to-audio generation, eliminating any dependency on textual inputs and thereby preserving the video-to-audio evaluation setting. For the LongVale datasets, the original evaluation sets predominantly consist of short video clips, many of which have audio segments shorter than one minute. To address this limitation and create a more balanced evaluation scenario, we selectively sample additional videos from the training split of LongVale [4] and eliminate short videos from the original test set. These selected videos are incorporated into the evaluation set to increase diversity and length. As a result of this augmentation, the final evaluation set comprises around 1K videos, each averaging approximately 45 seconds in duration. This adjustment ensures a more representative and robust evaluation for tasks involving video-to-audio generation.

2. The Details of MMHNet

Flow matching. We use flow matching in our proposed approach. To be specific, we use 25 training steps and apply the same for inference. We train with learning rates of $1e-4$ with the AdamW optimizer for 200K iterations.

Temporal synchronization features. The temporal synchronization feature is encoded using Synchformer model [6]. A 1D convolutional layer (kernel size = 7, padding = 3) is employed to project the input into a hidden representation, followed by a SELU activation function [8]. Subsequently, a ConvMLP layer with a kernel size of 3 and padding of 1 is applied.

Visual and text semantic features. Semantic visual and textual features are encoded using the CLIP model [10] to capture cross-modal representations. The subsequent projection layer incorporates a ConvMLP block with a kernel size of 3 and padding of 1, enabling local spatial interactions while preserving the original sequence length.

Audio features. A 1D convolutional layer (kernel size = 7, padding = 3) is utilized to project the input into a hidden representation, followed by a SELU activation function [8]. Subsequently, a ConvMLP layer with a kernel size of 7 and padding of 3 is applied.

Audio Variational Auto Encoder (VAE). As described in the main paper, audio latents are obtained by first applying a short-time Fourier transform (STFT) to the input audio and extracting the magnitude component as mel spectrograms. We use 44 kHz audio with a latent frame rate of 43.07. The mel bins, FFT size, hop size, and window size are set to 128, 2048, 512, and 2048, respectively. These spectrograms are then encoded into latent representations using a pretrained VAE. During inference, the generated latents are decoded back into spectrograms via the VAE and subsequently converted into audio waveforms using a pretrained Vocoder, such as BigVGAN-V2 [7]. For the VAE architecture, we adopt the 1D convolutional design from Make-An-Audio 2 [5], employing a downsampling factor of 2.

Non-Causal Mamba-2. For the Non-Causal Mamba component, we adopt VSSD [11] as the primary building block. This module largely follows the architectural principles of Mamba-2 [2], but with a key distinction: the computation is performed in a non-sequential manner. By removing the strict sequential processing constraint, the model can process multiple tokens simultaneously and capture a global view of the entire token sequence.

3. Additional Experiments

We conduct a further ablation study to observe the performance gain of each specific module in our proposed model. Note that we conduct this ablation study using the small version of MMHNet.

Ablation on routing strategies. We ablate on having the

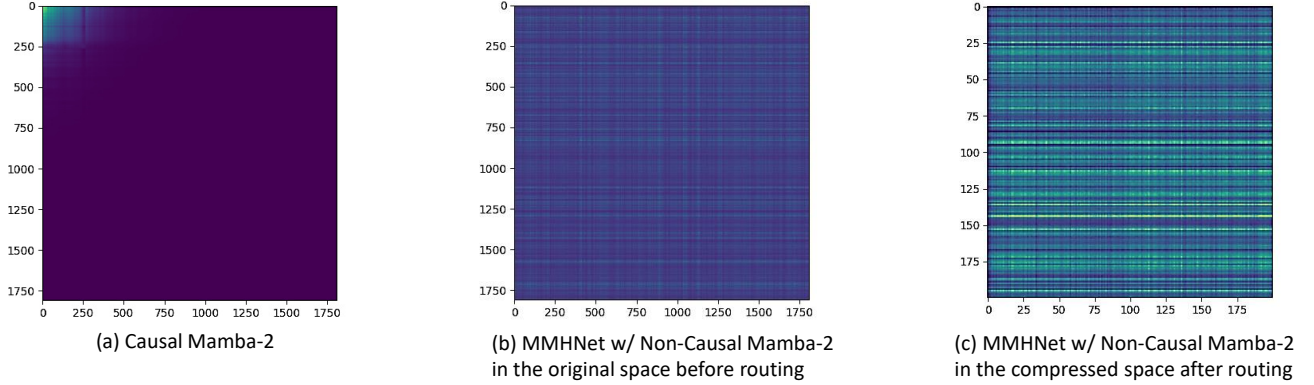


Figure A1. Visualization of heatmaps for activation matrices in Causal Mamba-2 and Non-Causal Mamba-2 within MMHNet: (a) Causal Mamba-2, used as a Transformer replacement, shows activation scores in the transition matrix that gradually decay during extended audio generation (up to 5 minutes). (b) Non-Causal Mamba-2 maintains visible activation scores in the transition matrix prior to routing. (c) After routing, the transition matrix becomes more pronounced in the compressed representation space.

UnAV100						
Variant	FD _{PANNs} ↓	FD _{PASST} ↓	ISC _{PANNs} ↑	ISC _{PASST} ↑	IB-Score ↑	DeSync ↓
No Routing	6.31	264.43	6.58	6.72	35.00	0.621
Temp. Routing	6.57	214.01	7.06	7.09	33.82	0.474
Temp. + MM Routing	5.87	217.00	7.62	8.21	36.82	0.439

Table A1. Ablation study on routing strategies using the UnAV100 dataset.

UnAV100						
	FD _{PANN} ↓	FD _{PASST} ↓	ISC _{PANN} ↑	ISC _{PASST} ↑	IB-Score ↑	DeSync ↓
W/o Pos. Emb.	3.24	220.37	7.76	8.30	36.75	0.425
W/ Pos. Emb.	3.35	217.00	7.62	8.21	36.82	0.439

Table A2. We compare our proposed approach with and without positional embeddings applied on input conditions.

structure of temporal and MM routing in our proposed network structure as shown in Table A1. We observe that the model with a temporal routing mechanism could improve DeSync scores, which are related to temporal synchronization between audio and visual modalities.

Ablation on additional position embeddings for the temporal sync. condition. Beyond the current framework, we also conducted an experiment to assess the impact of positional embeddings. Specifically, we examined whether removing them would degrade performance and whether our design choice could be justified. As shown in Table A2, the use of positional embeddings has minimal impact on overall performance.

Analysis on Causal-Mamba and Non-Causal Mamba attention maps. Figure A1 illustrates the activation maps of the transition matrix of Mamba-2 across all tokens, taken from the first single-modal layer. From these visualizations, we observe that the activation scores in Causal Mamba-2 exhibit a noticeable decay as more tokens are processed.

Specifically, the activations are concentrated within the initial segment of the sequence, primarily spanning the first 250–300 tokens, which corresponds to approximately 10 seconds of audio. This pattern suggests that the model’s attention is biased toward early tokens, with diminishing influence on later tokens.

Running time. We evaluated the time required to convert long videos into audio across multiple samples. Our proposed method achieves speed improvement in the wall clock time compared to MMAudio [1], despite sharing a similar MMDiT-like architecture. For example, our approach with a large version can generate 500 seconds of audio in approximately 60 seconds, whereas MMAudio takes about 120 seconds for the same task, up to 2× improvement. All measurements were conducted on an H100 GPU with 80GB memory.

Similarity metrics. We also evaluated alternative similarity metrics, but, as shown in Tab. A3, they consistently underperform cosine similarity across most evaluation measures. This behavior is expected. Our routing mechanism relies on CLIP-based condition encoders, and CLIP is explicitly trained with a cosine-similarity objective. Using a mismatched distance metric would fundamentally misalign with the geometry of the CLIP embedding space and degrade token selection. Consequently, cosine similarity is the only principled and effective choice for our routing mechanism.

Distance Metric	FD _{VGG} ↓	FD _{PANNs} ↓	ISC _{PANNs} ↑	IB-Score ↑	DeSync ↓
Euclidean	3.53	6.76	8.80	35.48	0.433
Dot Product	3.69	7.03	8.57	35.39	0.424
Cosine Similarity	1.80	5.29	8.10	36.27	0.410

Table A3. Comparison with different distance metrics on UnAV100.

Variant	UnAV100					
	FD _{PANNs} ↓	FD _{PASST} ↓	ISC _{PANNs} ↑	ISC _{PASST} ↑	IB-Score ↑	DeSync ↓
CFG=2.0	6.5	241.36	7.70	7.16	33.08	0.586
CFG=3.0	7.33	257.02	6.88	6.30	28.91	0.571
CFG=4.0	5.29	209.06	8.10	7.35	36.27	0.410
CFG=5.0	5.80	183.86	8.40	7.30	36.77	0.412
CFG=6.0	6.75	224.04	8.26	7.26	35.14	0.435

Table A4. Analysis on different CFG scores.

Performance across different hyperparameters. We analyze different classifier-free guidance (CFG) values to identify the optimal setting for achieving the best results, as shown in Table A4. Based on this evaluation, we use a CFG value of 4.0 as the hyperparameter across all experiments.

References

- [1] 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. 2
- [2] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 1
- [3] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 1
- [4] Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18959–18969, 2025. 1
- [5] 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, 2023. 1
- [6] 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. 1
- [7] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022. 1
- [8] Boyu Li, Haobin Jiang, Ziluo Ding, Xinrun Xu, Haoran Li, Dongbin Zhao, and Zongqing Lu. Selu: Self-learning embodied mllms in unknown environments. *arXiv preprint arXiv:2410.03303*, 2024. 1
- [9] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multi-modal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–15, 2024. 1
- [10] 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. 1
- [11] Yuheng Shi, Mingjing Dong, Mingjia Li, and Chang Xu. Vssd: Vision mamba with non-causal state space duality. *arXiv preprint arXiv:2407.18559*, 2024. 1