ICCV
#*****

SpecMaskFoley: Efficient Yet Effective Synchronized Video-to-audio Synthesis via Pretraining and ControlNet

Anonymous ICCV submission

Paper ID *****

Abstract

Foley synthesis is a task of wide interest that aims to synthesize high-quality audio which is both semantically and temporally aligned with video frames. To avoid the non-trivial task of training audio generative models from scratch, adapting pretrained audio generative models for video-synchronized foley synthesis presents an attractive direction. ControlNet, a method for adding fine-grained controls to pretrained generative models, has been applied to foley synthesis, but its use has been limited to handcrafted human-readable temporal conditions. In contrast, from-scratch models achieved success by leveraging high-dimensional deep features extracted using pretrained video encoders. We have observed a performance gap between ControlNet-based and from-scratch foley models. To narrow this gap, we propose SpecMaskFoley, a method that steers the pretrained SpecMaskGIT model toward video-synchronized foley synthesis via ControlNet. To unlock the potential of a single ControlNet branch, we resolve the discrepancy between the temporal video features and the time-frequency nature of the pretrained SpecMaskGIT via a frequency-aware temporal feature aligner, eliminating the need for complicated conditioning mechanisms widely used in prior arts. Evaluations on a common foley synthesis benchmark demonstrate that SpecMaskFoley could even outperform strong from-scratch baselines, substantially advancing the development of ControlNet-based foley synthesis models. Demo samples are uploaded as supplementary files.

1. Introduction

Text-to-audio (TTA) synthesis targets at synthesizing realistic sound events by natural language prompts [7, 9, 17, 26, 36]. In spite of the success of TTA systems in terms of sound quality as well as the semantic alignment between audio and text, the use of such systems in *foley* synthesis is very limited. Foley synthesis aims to synthesize audio that is not only semantically but also *temporally* aligned with video frames [5]. Automated foley synthesis has broad impact to creative industries, thus has gained arising attention in the research community.

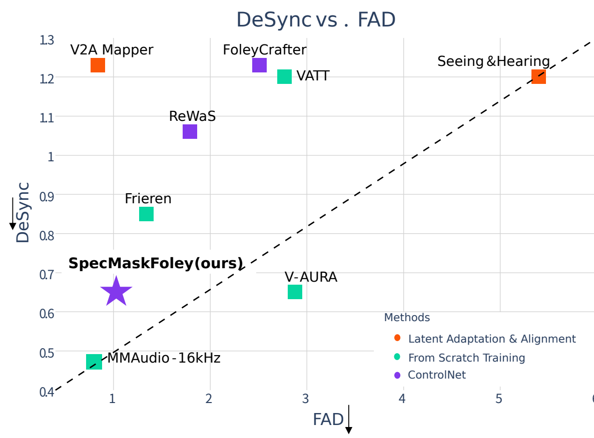

Figure 1. Audio synthesis quality (FAD [22]) and audio-video temporal alignment [5, 20] of different methods. The proposed SpecMaskFoley achieves competitive performance.

Temporal alignment for foley synthesis can be learned via the joint generative modeling of audio-visual pairs ([35, 43]), though the training can be expensive. Although there have been attempts to achieve temporal control of audio objects using text prompts [17], the mainstream is to explicitly condition audio generative models with temporal features that are synchronized with the video. To avoid the non-trivial task of training an audio generative model from scratch, steering *pretrained* audio generative models toward foley synthesis presents an attractive direction. Adapting or aligning pretrained audio and video latent spaces has been investigated ([39, 42]), but the temporal alignment between video and audio is poor [5].

ControlNet ([44, 47]), proposed to add *spatial* controls to pretrained image generative models, has been introduced to foley synthesis to inject handcrafted human-readable *temporal* conditions into pretrained TTA models ([21, 45]). Attempts have also been made to train foley synthesis models

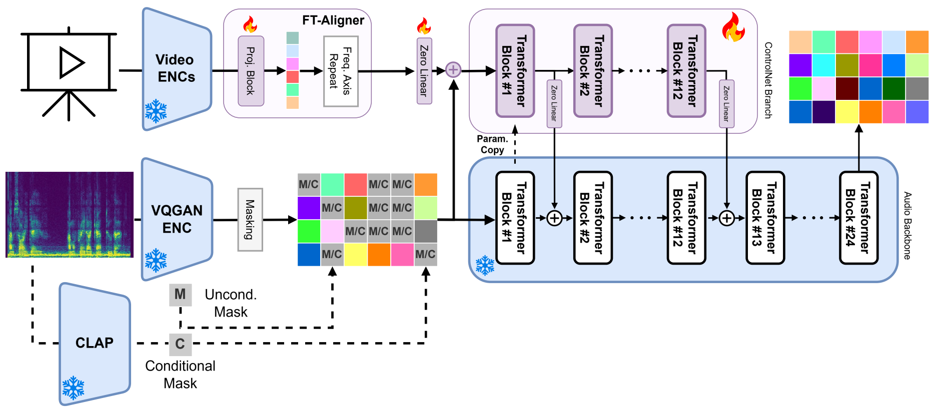

Figure 2. Overview of SpecMaskFoley. Ice icons: frozen modules. Fire icons: trainable modules. A CLAP embedding is treated as a conditional mask C following [7] to condition the audio backbone with audio prompts during training and text prompts during inference.

from-scratch with high-dimensional *deep features* extracted using pretrained video encoders [5, 31, 38, 40]. With advanced video encoders([18, 20, 32]), from-scratch models like Frieren [40] and MMAudio [5] brought foley synthesis to a new level.

As shown in Fig. 1, a notable performance gap can be observed between from-scratch models and ControlNet models, indicating that the effective use of pretrained audio generative models in the foley synthesis task is yet to be explored. To narrow this gap, we propose *SpecMaskFoley*, a method that steers the pretrained SpecMaskGIT model [7] toward video-synchronized foley synthesis via ControlNet. Our contributions lie in the following aspects. **(1) Competitive foley synthesis performance as a ControlNet-based method.** SpecMaskFoley outperforms not only prior ControlNet-based foley models [21, 45], but also strong from-scratch baselines, including Auto-Regressive model (AR) [38], MaskGIT model [31], and flow-matching model [40], in a widely used foley synthesis benchmark. **(2) Simplicity in neural network architecture.** Prior foley synthesis methods tend to combine multiple conditioning mechanisms for better video-audio alignment, *e.g.*, the combination of cross-attentions adaptors and ControlNet [45]. Nevertheless, SpecMaskFoley utilizes a single ControlNet branch with a Frequency-aware Temporal feature Aligner (FT-Aligner) to effectively adapt the temporal video features to our time-frequency ControlNet branch. The success of our simple design underscores the potential of ControlNet to process complex, high-dimensional features, paving the way to future extensions. Please find our review on related works in appendix.6.

2. SpecMaskFoley

As shown in Fig.2, SpecMaskFoley consists of a pretrained SpecMaskGIT [7] TTA model as the audio backbone, ControlNet branch, and an FT-Aligner to adapt temporal deep video features to our time-frequency ControlNet.

2.1. SpecMaskGIT

SpecMaskGIT features a distillation-free approach to efficient high-quality TTA synthesis [7]. The efficiency, effectiveness and flexibility of this method are based on: (1) A highly compressed latent space created by a 2-D SpecVQGAN [19]; (2) the light-weight ViT backbone trained with the Mased Language Modelling task ([13, 43, 48]) upon the latent space of SpecVQGAN; (3) the use of parallel iterative synthesis, *i.e.*, MaskGIT sampler, for fast inference ([1, 28]). We chose to use SpecMaskGIT as the backbone of SpecMaskFoley for the following reasons. First, SpecMaskGIT has been reported [7] to be competitive in a TTA benchmark – on par with AudioLDM [30], a TTA model widely used in foley synthesis research ([21, 39, 42, 45]) – while being more efficient in inference than prior arts. Second, the lightweight and efficient nature of SpecMaskGIT enables fast concept verification and iterative try-and-error exploration with academic level of resources. Finally, MaskGIT methods showed promising results in foley synthesis ([31, 43]).

2.2. Deep Video Features

Instead of using handcrafted features as with prior ControlNet methods ([21, 45]), we adopt the same deep video features as in [5, 38] to our ControlNet branch: A 25 Hz high-frame-rate video feature extracted using Synchformer [20] and an 8 Hz semantic feature extracted using CLIP [34]. Both features are 1-D sequences of high dimensionality. Although the 8 Hz CLIP features have been used in MMAudio [5], we hypothesize that CLIP features present more semantic rather than temporal synchronization information, thus can be averaged across all frames to form a global semantic condition. This global condition is directly added to Synchformer features, resulting in a 1-D temporal features of shape $[t, d]$, where t is the number of temporal frames of the deep feature, and d is the dimension.

2.3. ControlNet and FT-Aligner

As shown in Fig.2, we implement a Transformer version of ControlNet similar to [4, 6, 16], which initializes the ControlNet branch with the pretrained model weights and then connects the ControlNet branch to the backbone via zero-initialized linear layers. Although this design was presented for continuous Diffusion Transformers [4, 6, 16], we found it works well in our discrete model. Note that the only training loss in SpecMaskFoley is cross entropy loss, the standard training target for discrete models ([1, 7, 28, 43]).

A major challenge to control the pretrained SpecMaskGIT model with *1-D* temporal features is the *2-D* time-frequency nature of SpecMaskGIT. SpecMaskGIT works in a 2-D latent space, where each token represents a 16-by-16 time-frequency patch in the original Mel-spectrogram. To achieve effective temporal alignment, it is essential to inject the *identical* temporal feature to all tokens belonging to the same time frame. To this end, we propose a Frequency-aware Temporal feature Aligner (FT-Aligner) to address this challenge. Assuming the audio token sequence sent into SpecMaskGIT has a shape of $[F, T, D]$, denoting the number of tokens along the frequency axis, number of tokens along the temporal axis, and dimensionality for each token. For 1-D temporal video features of shape $[t, d]$ described in Sec.2.2, we use a Projection Block containing a 1-D conv layer followed by an adaptive average pooling to downsample the 1-D features to $[1, T, D]$, then *repeat* the sequence along the frequency axis, resulting 2-D feature embeddings with the shape of $[F, T, D]$ that preserves original 1-D temporal information. We empirically found this FT-Aligner facilitates the convergence in training. Without this careful feature alignment, we would not be able to train the model.

2.4. Multi Classifier-Free Guidance

Classifier-Free Guidance (CFG [15]) has been used to improve TTA synthesis quality by balancing between diversity and audio-text alignment ([1, 2, 7]). Inspired by StemGen [33], we extend the original single-condition CFG formula to two conditions, effectively using both video and text conditional features when available:

$$\ell_{\text{foley}} = \ell_{\text{uncond}} + t[(\ell_{\text{text\&video}} - \ell_{\text{uncond}}) + (\ell_{\text{video}} - \ell_{\text{uncond}})], \quad (1)$$

where ℓ_{uncond} denotes the logits gained from the audio backbone without the CLAP conditioning, ℓ_{video} denotes logits obtained from conditioning SpecMaskFoley with deep video features via ControlNet but without CLAP feautures in the backbone, $\ell_{\text{text\&video}}$ denotes logits earned by conditioning SpecMaskFoley with both CLAP text features and deep video features simultaneously, and t denotes the CFG scale. We do not use logits gained by conditioning the audio backbone with CLAP text features, and we also randomly replace the CLAP conditioning with an unconditonal mask (shown in Fig.2) for 90% of the training steps, as we found text prompts alone are less important in Sec.**??**.

3. Experiments

3.1. Datasets

The pretraining of the TTA backbone in SpecMaskFoley was conducted on the sum of the unbalanced and balanced subset of AudioSet [10], a dataset that has been widely used in general audio representation learning ([25, 48]) and generative modeling ([17, 30]) due to its massive amount of audio clips as well as its diversity in sound sources and recording environments.

The ControlNet branch of SpecMaskFoley is trained on VGGSound [3], the only audio-visual dataset used in this study, which contains around 500 hours sounding video clips. On top of its synchronized audio-video pairs, video clips in VGGSound come with tags from a 310-class taxonomy. We follow the data split and preprocessing pipeline in MMAudio [5], in which the train set contains around 180K 10-second video clip. However, we do not truncate the videos to 8s, as our audio backbone was pretrained with 10-second audio clips.

Following common practice in ReWaS[21], VATT[31], FoleyCrafter[45], and MMAudio[5], we concatenate tags of the test set as the text input to SpecMaskFoley during evaluation.

3.2. Implementation Details

For the TTA backbone, we use the official checkpoint of SpecMaskGIT [7], which was pretrained on the AudioSet. More details of pretraining can be found in [7]. The standard Mel-spectrogram transform from vocoders [23] is used, which transforms 10-second audio clip at the sampling rate 22.05kHz into 848 frames with 80 Mel bins. The Mel-spectrogram is further tokenized using SpecVQGAN with a 16-by-16 downsampling factor, resulting in a 2-D token map of $[F = 5, T = 53]$, while each token in the map is represented by a 256-dimension embedding from a 10bit codebook. Each 10-second video clip is processed by Synchformer [20] into 240 feature frames, and by CLIP [34] into 80 feature frames respectively. These CLIP feature frames are then globally averaged (Sec.2.2).

The audio backbone in SpecMaskFoley employs 24 Transformer blocks, in which the attention dimension is $D = 768$ with 8 heads and the feedforward dimension is 3072, resulting in around 170M parameters. We copy the first 12 Transformer blocks, *i.e.*, half of the audio backbone, to initialize the ControlNet branch in SpecMaskFoley. The total number of *trainable* parameters is around 126M. To align deep video features with the shape of the 2-D audio token map, the proposed FT-Aligner first downsamples the aforementioned deep features to $[F = 1, T = 53,$

Table 1. Benchmarking on VGGSound test set. AR.: Auto-regressive. Mask.: MaskGIT. Diff: Diffusion and flow matching. Bold: best score. Underline: the second and third best scores. Inference time is computed on a H100 GPU with batch size 1 for one 10-second clip.

Method	Params↓	Type	Pretrained Backbone	Video features	Audio Synthesis FD↓	Audio Synthesis FAD↓	Audio Synthesis KL↓	AV Semantic IB Similarity↑	AV Sync. DeSync (s)↓	Infer. Time (s)↓
Latent Adaptation & Alignment										
Seeing & Hearing [42]	415M	Diff.	AudioLDM [30]	ImageBind [11]	219	5.40	2.3	**34.0**	1.20	14.6
V2A Mapper [39]	230M	Diff.	AudioLDM [30]	CLIP [34] & CLAP [41]	84.6	0.84	2.56	22.6	1.23	-
ControlNet + Handcrafted Features										
ReWaS [21]	620M	Diff.	AudioLDM [30]	Energy Curve	141	1.79	2.82	14.8	1.06	16.0
FoleyCrafter [45]	1.22B	Diff.	AudioLDM [30]	Onset timestamps	140	2.51	2.23	25.7	1.23	1.7
From Scratch Training										
VATT [31]	-	Mask	-	eva-CLIP [37]	132	2.77	**1.41**	25.0	1.20	-
V-AURA [38]	695M	AR	-	Synchformer [20]	218	2.88	2.07	27.6	0.65	16.6
Frieren [40]	**160M**	Diff.	-	CAVP [32] & MAViL [18]	106	1.34	2.86	22.8	0.85	-
MMAudio-16kHz [5]	**160M**	Diff.	-	Synchformer [20] & CLIP [34]	**70.2**	**0.79**	1.59	29.1	**0.48**	1.23
ControlNet + Deep Features										
SpecMaskFoley (ours)	300M	Mask	SpecMaskGIT [7]	Synchformer [20] & CLIP [34]	109	1.03	1.76	26.4	0.65	**0.47**

$D = 768$] then repeats the features 5 times along the frequency axis.

We trained the ControlNet branch for 140K steps on a *single* A6000 GPU. Other training details can be found in appendix.7

3.3. Metrics

We use the av-benchmark [1] to evaluate the quality of foley synthesis from the following aspects. Please find details in appendix.8.

4. Results: VGGSound Benchmarking

Benchmarking results on VGGSound are presented in Tab. 1, in which the top-3 scores are highlighted.

Several observations can be made. First, MMAudio [5] ranks in the top-3 among all metrics, which sets a high standard for foley syntheis and indicates room of improvement for SpecMaskFoley. Second, the proposed SpecMaskFoley outperforms those latent adaptation methods in DeSync while maintaining competitive audio synthesis quality, indicating the effectiveness of using ControlNet to enhance the video-audio synchronization. Next, SpecMaskFoley outperforms previous ControlNet-based methods in audio synthesis quality, audio-video semantic matching, audio-video synchronization, and inference speed, pushing the boundary of ControlNet-based methods for foley synthesis. We believe the advantage of SpecMaskFoley comes from the well-trained SpecMaskGIT backbone, the use of deep video features and the effective use of these features with our ControlNet and FT-Aligner.

Last but not least, SpecMaskFoley remains competitive even compared with strong from-scratch baselines. methods. SpecMaskFoley substantially outperforms VATT [31], a from-scratch MaskGIT method, in most metrics in Tab. 1. Compared with V-AURA [38], SpecMaskFoley presents superior audio synthesis quality and inference speed, while maintaining the same level of audio-video synchronization. Overall, SpecMaskFoley performed similarly to Frieren [40], a from-scratch flow-matching method, but our method is more advantageous in metrics such as FAD, KL, IB similarity, and DeSync. Note that although SpecMaskFoley uses more parameters than Frieren and MMAudio, the *trainable* parameter amount is only 126M, enabling *single* GPU training, hence is friendly to low-resource researchers.

According to the above observation, we believe that SpecMaskFoley has substantially narrowed the gap between from-scratch and ControlNet-based methods in the field of foley synthesis.

Ablation studies can be found in appendix.9

5. Conclusion

We addressed the challenging task of foley synthesis in this work. To avoid the non-trivial task of training audio generative models from scratch, it is attractive to use pretrained TTA models. Many previous ControlNet-based foley synthesis methods have limited themselves to handcrafted human-readable temporal conditions. On the other hand, from-scratch models achieved success by leveraging high-dimensional deep features extracted from pretrained video encoders. To narrow the performance gap between ControlNet-based and from-scratch foley models, we proposed SpecMaskFoley. Our method steers the pretrained SpecMaskGIT model toward video-synchronized foley synthesis by directly processing deep video features with ControlNet. A frequency-aware temporal feature aligner was introduced to resolve the discrepancy between the temporal video features and the time-frequency nature of the pretrained SpecMaskGIT, simplifying the conditioning mechanism in SpecMaskFoley. Evaluations on a common foley synthesis benchmark demonstrate that SpecMaskFoley could even outperform strong from-scratch baselines. Future work includes transplanting our methodology to 1-D audio MaskGIT models, and extending the scope of audio ControlNet toward spatial audio.

[1] https://github.com/hkchengrex/av-benchmark

References

[1] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proc. CVPR 2022*, pages 11315–11325, 2022. 2, 3, 1

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *Proc. ICASSP 2020*, pages 721–725. IEEE, 2020. 3

[4] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 3

[5] 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, 3, 4

[6] Ruben Ciranni, Giorgio Mariani, Michele Mancusi, Emilian Postolache, Giorgio Fabbro, Emanuele Rodolà, and Luca Cosmo. Cocola: Coherence-oriented contrastive learning of musical audio representations. *arXiv preprint arXiv:2404.16969*, 2024. 3

[7] Marco Comunità, Zhi Zhong, Akira Takahashi, Shiqi Yang, Mengjie Zhao, Koichi Saito, Yukara Ikemiya, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. Specmaskgit: Masked generative modeling of audio spectrograms for efficient audio synthesis and beyond. *arXiv preprint arXiv:2406.17672*, 2024. 1, 2, 3, 4

[8] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 1

[9] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *Proc. ICASSP 2025*, pages 1–5. IEEE, 2025. 1

[10] 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 *Proc. ICASSP 2017*, pages 776–780. IEEE, 2017. 3

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proc. CVPR 2023*, pages 15180–15190, 2023. 4, 1

[12] Riccardo Fosco Gramaccioni, Christian Marinoni, Emilian Postolache, Marco Comunità, Luca Cosmo, Joshua D Reiss, and Danilo Comminiello. Stable-v2a: Synthesis of synchronized sound effects with temporal and semantic controls. *arXiv preprint arXiv:2412.15023*, 2024. 1

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR 2022*, pages 16000–16009, 2022. 2

[14] 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 *Proc. ICASSP 2017*, pages 131–135. IEEE, 2017. 1

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[16] Siyuan Hou, Shansong Liu, Ruibin Yuan, Wei Xue, Ying Shan, Mangsuo Zhao, and Chao Zhang. Editing music with melody and text: Using controlnet for diffusion transformer. In *Proc. ICASSP 2025*, pages 1–5. IEEE, 2025. 3

[17] 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. 1, 3, 2

[18] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Proc. NeurIPS 2023*, 36:20371–20393, 2023. 2, 4, 1

[19] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021. 2

[20] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *Proc. ICASSP 2024*, pages 5325–5329. IEEE, 2024. 1, 2, 3, 4

[21] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024. 1, 2, 3, 4

[22] 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. 1

[23] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *NeurIPS*, 33:17022–17033, 2020. 3

[24] 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 TASLP*, 28:2880–2894, 2020. 1

[25] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021. 3, 1

[26] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 1

[27] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Proc. NeurIPS 2023*, 36:27980–27993, 2023. 1

[28] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proc. CVPR 2023*, pages 2142–2152, 2023. 2, 3, 1

[29] WeiHsiang Liao, Yuhta Takida, Yukara Ikemiya, Zhi Zhong, Chieh-Hsin Lai, Giorgio Fabbro, Kazuki Shimada, Keisuke Toyama, Kinwai Cheuk, Marco A Martínez-Ramírez, et al. Music foundation model as generic booster for music downstream tasks. *arXiv preprint arXiv:2411.01135*, 2024. 1

[30] 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. 2, 3, 4

[31] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see–video to audio generation through text. *arXiv preprint arXiv:2411.05679*, 2024. 2, 3, 4, 1

[32] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Proc. NeurIPS 2023*, 36:48855–48876, 2023. 2, 4, 1

[33] Julian D Parker, Janne Spijkervet, Katerina Kosta, Furkan Yesiler, Boris Kuznetsov, Ju-Chiang Wang, Matt Avent, Jitong Chen, and Duc Le. Stemgen: A music generation model that listens. In *Proc. ICASSP 2024*, pages 1116–1120. IEEE, 2024. 3

[34] 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 *Proc. ICML 2021*, pages 8748–8763. PmLR, 2021. 2, 3, 4, 1

[35] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proc. CVPR 2023*, pages 10219–10228, 2023. 1

[36] Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. Soundctm: Uniting score-based and consistency models for text-to-sound generation. *arXiv preprint arXiv:2405.18503*, 2024. 1

[37] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 4, 1

[38] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *Proc. ICASSP 2025*, pages 1–5. IEEE, 2025. 2, 4, 1

[39] 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. In *Proc. AAAI 2024*, pages 15492–15501, 2024. 1, 2, 4

[40] 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. *Proc. NeurIPS 2024*, 37:128118–128138, 2024. 2, 4, 1

[41] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP 2023*, pages 1–5, 2023. 4, 1

[42] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proc. CVPR 2024*, pages 7151–7161, 2024. 1, 2, 4

[43] Shiqi Yang, Zhi Zhong, Mengjie Zhao, Shusuke Takahashi, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Visual echoes: A simple unified transformer for audio-visual generation. *arXiv preprint arXiv:2405.14598*, 2024. 1, 2, 3

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. CVPR 2023*, pages 3836–3847, 2023. 1

[45] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foley-crafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. 1, 2, 3, 4

[46] Yaoyun Zhang, Xuenan Xu, and Mengyue Wu. Smooth-foley: Creating continuous sound for video-to-audio generation under semantic guidance. In *Proc. ICASSP 2025*, pages 1–5. IEEE, 2025. 1

[47] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Proc. NeurIPS 2023*, 36:11127–11150, 2023. 1

[48] Zhi Zhong, Hao Shi, Masato Hirano, Kazuki Shimada, Kazuya Tateishi, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. Extending audio masked autoencoders toward audio restoration. In *IEEE WASPAA 2023*, pages 1–5, 2023. 2, 3, 1

SpecMaskFoley: Efficient Yet Effective Synchronized Video-to-audio Synthesis via Pretraining and ControlNet

Supplementary Material

6. Appendix: Related Works

While significant progress has been made in TTA synthesis [7, 9, 17, 26, 36], it is still difficult to achieve fine-grained temporal control using natural language prompts [17]. The joint generative modeling of synchronized visual and audio data leads to an aligned latent space, such as in MMDiffusion [35] and VisualEchoes [43]. However, jointly modeling multiple modalities from scratch is non-trivial. The adaptation or alignment of pretrained audio and video latent spaces has been explored to mitigate the burden of from-scratch training ([39, 42]), but resulted in poor temporal alignment between video and audio [5].

ControlNet ([44, 47]), originally designed to add spatial controls to pretrained image generative models, has been introduced to foley synthesis to explicitly inject handcrafted human-readable temporal conditions into pretrained audio generative models. Typical handcrafted features include the energy curves ([12, 21]), onset timestamps [45], and binarized CLIP stamps [46]. FoleyCrafter [45] further combines ControlNet with parallel cross-attention adaptors to make use of CLIP [34] visual features.

Parallel to the investigation of ControlNet-based models, attempts have been made to train foley synthesis models from scratch. From-scratch models often take the advantage of various high-dimensional deep features extracted using pretrained visual encoders. For example, VATT [31] trains a MaskGIT ([1, 7]) model controlled by eva-CLIP [37] visual features. V-AURA trains an AR Transformer ([26, 29]) with the deep features extracted by Synchformer [20]. Both VATT and V-AURA are trained upon a discrete latent space created by 1-D VQ-GANs ([8, 27]). Frieren [40] trains a flow-matching model upon the latent space created by a 1D-VAE-GAN ([17]) conditioned by deep features from both contrastive learning ([32]) and masked Transformer models ([18]). MMAudio [5] brought the foley synthesis task to a new level. Similar to Frieren, MMAudio is a flow-matching model built upon a 1D-VAE-GAN, while taking the advantage of the Synchformer features ([20, 38]) and CLIP visual features [34].

A method that more effectively leverages pretrained audio generative models is likely to narrow the notable performance gap observed between from-scratch and ControlNet-based models.

7. Appendix: Implementation Details

We trained the ControlNet branch for 140K steps on a *single* A6000 GPU. Following common practice ([25, 48]), we employ a linear warmup for the first 2.8 K steps then a cosine annealing of the learning rate (LR) for the remaining training. The batch size is set to 64, the base LR is set to 1e-3. The LR equates to the base LR times the batch size divided by 256 ([7, 28]).

Unless denoted, we use the multi-CFG described in Sec.2.4 with the CFG scale linearly increasing from 0 to 3 across the 12 inference steps with the Gumbel temperature ([7]) set to 9.0 during evaluation.

8. Appendix: Metrics

We use the av-benchmark [2] to evaluate the quality of foley synthesis from the following aspects: **Audio synthesis quality**. Following common practice ([5, 7, 17, 26, 36]), we compute the Frechet Distance (FD) and Kullback–Leibler (KL) distance on the basis of PaSST [25], a 32 kHz audio classifier. The 16 kHz VGGish classifier [14] is also used for FD (denoted as "FAD"). Note that we exclude PANN [24] from the metric computation, as it has been reported as not being robust in some scenarios [3]. **Audio-video semantic matching.** The semantic similarity between the input video and the generated foley audio is evaluated by the cosine similarity between ImageBind [11] video and audio features, denoted as "IB similarity". **Audio-video synchronization**. The synchronization between input video and generated audio is measured by DeSync ([5, 20]) in seconds.

9. Appendix: Ablation Studies

Effectiveness of the ControlNet branch. SpecMaskFoley is competitive as shown in Tab.1, but it is still unclear that to what extent the ControlNet branch has contributed to this success. Therefore, we prompt SpecMaskGIT, the audio backbone without ControlNet, to see the behavior change before and after adding the ControlNet branch.

As shown in Tab.2, SpecMaskGIT largely deviates from the desired foley synthesis when prompted with the concatenated audio tags of the test set, which might have been caused by the limited capability of the text encoder in CLAP [41]. Results prompted by the CLAP audio features of the

[2] https://github.com/hkchengrex/av-benchmark
[3] https://github.com/haoheliu/audioldm_eval

Table 2. Comparison of pretrained backbone, inference steps, and CFG settings. Bold: best score. Underline: second and third best score.

Method	Audio Synthesis FD↓	FAD↓	KL↓	AV Semantic IB Similarity↑	AV Sync. DeSync (s)↓	Infer. Time (s)↓
SpecMaskGIT-*16*-step						
w/ text prompt	223	4.97	2.77	10.2	1.28	0.32
w/ audio prompt	**99.8**	**0.91**	**0.95**	22.7	1.22	0.32
SpecMaskGIT-*12*-step						
w/ text prompt	226	5.07	2.75	9.9	1.26	**0.27**
w/ audio prompt	108	1.01	0.96	22.7	1.22	**0.27**
SpecMaskFoley-*12*-step (ours)	109	1.03	1.76	**26.4**	**0.65**	0.47
CFG w/o $\ell_{\text{text\&video}}$	125	1.15	1.96	24.2	0.79	0.32
CFG w/o ℓ_{video}	125	1.33	1.82	22.8	0.75	0.32

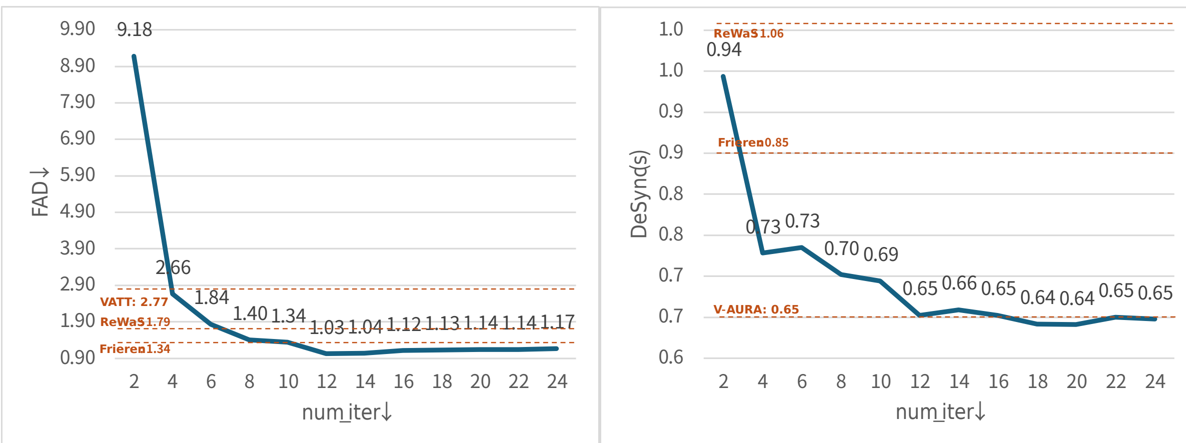

Figure 3. Left: FAD vs. Number of iterations. Right: DeSync vs. Number of iterations. Scores saturate after 12 iterations.

audio data in the test set present the upper bound of SpecMaskGIT in terms of audio synthesis, resulting in a competitive FAD score (0.91), as well as the best KL score (0.95) and fast inference speed (0.32s for 16 steps, 0.27s for 12 steps). Nevertheless, the audio-video semantic matching and temporal alignment remain weak as there is no video feature injected to SpecMaskGIT. As shown in Tab.2, our ControlNet branch effectively improved the audio-video semantic matching and temporal synchronization of the backbone, without significantly degrading the audio synthesis quality.

Impact of Multi-CFG. It is revealed in Tab.2 that Multi-CFG improves all metrics compared with standard CFG, with only a slightly increased, while still affordable inference cost.

Impact of inference steps is illustrated in Fig.3. While in the original SpecMaskGIT paper [7], the optimal number of inference steps was 16, with SpecMaskFoley, we found FAD and DeSync scores saturate after 12 steps. It is worth noting that, using as few as *4* steps, SpecMaskFoley outperforms VATT and FoleyCrafter in terms of FAD and DeSync; Using *6* steps, the FAD score becomes close to that of ReWaS. Fig.3 reveals SpecMaskFoley's few-step synthesis capability *without* any distillation.

Discussion: 2-D vs. 1-D VAE. As illustrated in Fig.2 and discussed in Sec.3.3, due to the use of a 2-D VAE in SpecMaskFoley, there are only 53 temporal frames in the latent space presenting a 10-second clip. However, the 1-D VAE used in Frieren and MMAudio presents a 10-second clip with more than 300 frames [17], preserving higher temporal resolution. According to the DeSync scores in Tab.1, this low temporal resolution has not been a bottleneck for SpecMaskFoley in the VGGSound benchmark, but may limit the potential of 2-D methods in the future. We leave the task of transplanting SpecMaskGIT and SpecMaskFoley to 1-D as our future work.