T-FOLEY: A CONTROLLABLE WAVEFORM-DOMAIN DIFFUSION MODEL FOR TEMPORAL-EVENT-GUIDED FOLEY SOUND SYNTHESIS

♯*Yoonjin Chung**, ♯*Junwon Lee**, ♯♭*Juhan Nam*

♯Graduate School of Artificial Intelligence, KAIST, Republic of Korea
♭Graduate School of Culture Technology, KAIST, Republic of Korea

ABSTRACT

Foley sound, audio content inserted synchronously with videos, plays a critical role in the user experience of multimedia content. Recently, there has been active research in Foley sound synthesis, leveraging the advancements in deep generative models. However, such works mainly focus on replicating a single sound class or a textual sound description, neglecting temporal information, which is crucial in the practical applications of Foley sound. We present **T-Foley**, a **T**emporal-event-guided waveform generation model for **Foley** sound synthesis. T-Foley generates high-quality audio using two conditions: the sound class and temporal event feature. For temporal conditioning, we devise a temporal event feature and a novel conditioning technique named Block-FiLM. T-Foley achieves superior performance in both objective and subjective evaluation metrics and generates Foley sound well-synchronized with the temporal events. Additionally, we showcase T-Foley's practical applications, particularly in scenarios involving vocal mimicry for temporal event control. We show the demo on our companion website.[1]

Index Terms— Foley Sound Synthesis, Controllable Sound Generation, General Audio Synthesis, Waveform Domain Diffusion

1. INTRODUCTION

Foley sound refers to human-created sound effects, such as footsteps or gunshots, to accentuate visual media. The significance of Foley sound lies in its ability to enhance the overall immersive experience for various forms of media [1]. Foley sounds are usually created by Foley artists who record and produce required sounds manually, synchronized with the visual elements.

The advent of neural audio generation has presented an opportunity to automate and streamline the Foley sound creation process, reducing the time and effort required for sound production. To synthesize proper sounds from specific categories, early studies usually focused on single sound sources such as foosteps [2], laughter [3], and drum [4, 5]. Subsequent research then further improved a model to be capable of generating multiple sound categories utilizing auto-regressive models [6, 7, 8], Generative Adversarial Networks(GANs) [9, 10], or diffusion models [11]. Recently, it has been possible to generate holistic scene sounds solely based on a text description, even without pre-defined sound categories [12, 13, 14].

While prior research demonstrated the faithful sound synthesis by neural models for a target source, few focused on the timing and
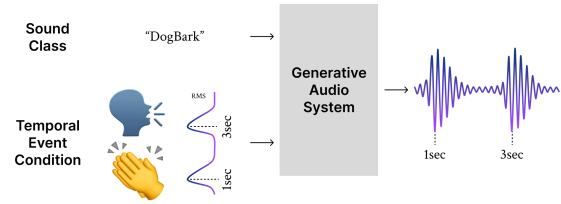
Fig. 1. Temporal-event-guided Foley sound synthesis.

envelope of sound events. Precisely locating sound events is crucial for practical Foley sound synthesis. Some studies generated implicit temporal features from input videos during synthesis [6, 9, 10, 15]. However, they lack explicit temporal event conditions for controllability and do not include quantitative analysis.

This research aims to address these challenges and produce realistic, timing-aligned Foley sound effects within a specific sound category. To the best of our knowledge, this is the first attempt to generate audio with explicit temporal event conditions. Our contributions are the following: First, we propose **T-Foley**, a **T**emporal-event-guided diffusion model with a conditioning sound class to generate high-quality **Foley** sound. For the temporal guidance, we introduce the temporal event feature to guide timing and envelope representation. To devise a conditioning method that reflects temporal informative condition, we introduce Block-FiLM, a modification of FiLM [16] for block-wise affine transformation. Second, we conduct extensive experiments to validate the performance and provide a comparative analysis of results on temporal conditioning methods. A metric is proposed to measure the temporal fidelity quantitatively. Lastly, we show the potential applications of T-Foley by demonstrating its performance on a human voice condition that mimics target sound events as an intuitive way to capture temporal event features.

2. T-FOLEY

As shown in Figure 1, the model is designed to input the sound class category and temporal event condition, representing "what" and "when" the sound should be generated, respectively.

2.1. Overall Architecture

The architecture of T-Foley is depicted in Figure 2. Our U-Net design, with an auto-regressive bottleneck, is based on the advanced DAG [11] model, which produces high-resolution audio without a pretrained vocoder. Compared with CRASH [5], which is the first

---

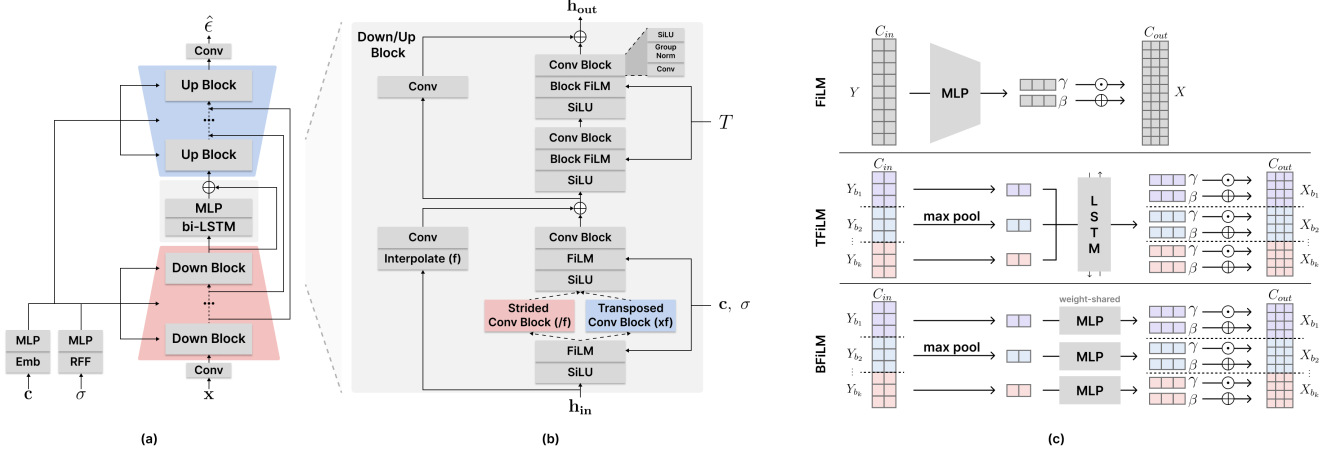[1]https://yoonjinxd.github.io/Event-guided_FSS_Demo.github.io

Fig. 2. (a) Overall architecture of the proposed model. (**c**: sound class, σ: diffusion timestep, T: temporal event feature) (b) A detailed structure of a Down/Up sampling block. Each Down block includes strided convolutional layer at first while Up blocks exploit the transposed one. ($\mathbf{h_{in}}/\mathbf{h_{out}}$: latent features) (c) Comparison of FiLM, TFiLM, and the proposed BFiLM. (Y: conditioning input, X: input activation)

diffusion model to generate waveform from scratch, DAG contains a sequential module at the bottleneck of deeper U-Net architecture to address the problem of inconsistent timbre within a generated sample. To predict the noise of each diffusion timestep, the model first downsamples the noised signal $\mathbf{x}$ into the latent vector and passes it to the bidirectional LSTM to maintain the timbre consistency within a sample. The output of the bottleneck layer is resized by linear projections and finally upsampled into the prediction of noise $\hat{\epsilon}$. In each downsample and upsample block, convolution layers are conditioned on diffusion time step embedding σ and class embedding $\mathbf{c}$ through FiLM, like in DAG, with the latter half employing Block-FiLM conditioned by temporal event feature T. The model is trained end-to-end to minimize the continuous-time L2 loss on predicted noise as proposed in CRASH [5].

2.2. Temporal Event Feature

As the primary objective of T-Foley is to generate audio upon a temporal event condition, the model needs to learn temporal information regarding the timing and envelope of sound events in time. This necessitates the appropriate conditioning temporal feature for the sound events. We used root-mean-square (RMS) of the waveform, which is widely used frame-level amplitude envelope features as below: for the i-th frame,

$$E_i(x) = \sqrt{\frac{1}{W} \Sigma_{t=ih}^{ih+W} x^2(t)} \qquad (1)$$

where $x(t)$ ($t \in [0, T]$) is the audio waveform, W is a window size and h is a hop size. In our experiment, we set $W = 512$ and $h = 128$. We also considered power (the square of RMS) and onset/offset (the start and end of a particular sound event) as candidates. After a preliminary experiment, we decided to use RMS because there was no significant difference between RMS and power, and some categories (e.g., rain, sneeze) do not have definite onsets and offsets by the nature of the sounds but temporal patterns with varying intensities (i.e., envelope).

2.3. Block FiLM

We propose Block FiLM as a neural module to condition the generation model on the temporal event feature. FiLM is one of the Conditional Batch Normalization (CBN) techniques to modulate the activations of individual feature maps or channels with affine transformation based on a conditioning input. It has been widely used in various tasks, including image synthesis, style transfer, and diffusion models for waveform generation [17, 5, 11]. Mathematically, with a conditioning input $\mathbf{y}_{C_{in}, L_{in}} \in \mathrm{R}^{C_{in} \times L_{in}}$, where C_{in} is the input channel size, and L_{in} is the input length, the FiLM modulation can be expressed as follows: $\mathrm{FiLM}(\mathbf{x}, \mathbf{y}, \gamma, \beta) = \gamma \odot \mathbf{x} + \beta$ where $\mathbf{x}$ represents the input activations, $\mathbf{c}$ is the conditioning input, $\gamma, \beta \in \mathrm{R}^{C_{out}}$ are normalizing parameters obtained as $\gamma, \beta = \mathrm{MLP}(\mathbf{y})$. The $\odot$ symbol denotes the Hadamard product(element-wise matrix multiplication).

Temporal FiLM (TFiLM) was proposed to overcome the limitation of FiLM in time-varying information in conditioning signals, which is crucial for processing audio signals. Given a sequential conditioning input $\mathbf{y}_{\mathbf{C_{in}}, \mathbf{L_{in}}} \in \mathrm{R}^{C_{in} \times L_{in}}$, where C_{in} is the input channel size, and L_{in} is the sequential length in the time domain, TFiLM firstly splits $\mathbf{y}$ into N blocks. THe i-th block $Y_{b_i} \in \mathrm{R}^{C_{in} \times L_{in}/N}$ ($i = 1, ..., N$) is max-pooled in the time dimension as $Y_{b_i}^{\mathrm{pool}} = \mathrm{Max\text{-}Pool}_{1:L_{in}/N}(Y_{b_i}) \in \mathrm{R}^{C_{in}}$ and followed by an RNN as sequential modeling to obtain normalizing parameters $(\gamma_i, \beta_i), h_i = \mathrm{RNN}(Y_{b_i}^{\mathrm{pool}}; h_{i-1})$. Finally, a linear modulation is applied channel-wise according to the normalizing parameters:

$$X_{b_i}^{\mathrm{output}} = (\mathbf{1}_{L_{out}/N}\gamma_i^T) \odot X_{b_i} + (\mathbf{1}_{L_{out}/N}\beta_i^T) \in \mathrm{R}^{C_{out} \times L_{out}/N} \qquad (2)$$

where $\gamma_i, \beta_i \in \mathrm{R}^{C_{out}}$, $\mathbf{1}_d = [1, ..., 1]^T \in \mathrm{R}^d$. TFiLM was originally proposed for self-conditioning to modulate intermediate features. Therefore, $C_{in} = C_{out}$ and $L_{in} = L_{out}$ as $[X_{b_i}\text{'s}]^T = \mathbf{x} = \mathbf{y} = [Y_{b_i}\text{'s}]^T$. In this paper, we generalize it to the case where conditioning signal $\mathbf{y}_{C_{in}, L_{in}}$ and modulating signal $\mathbf{x}_{C_{out}, L_{out}}$ are different. However, incorporating TFiLM in every conditioning layer can significantly increase computational complexity.

Block FiLM (BFiLM) is a simplified version of TFiLM motivated by the characteristics of RMS. The temporal events embedded in sequential information of RMS are weakly dependent (e.g.,

Model	#params↓	infer.t ↓	E-L1↓	FAD-P↓	FAD-V↓	IS↑
Real data	-	-	0.0	22.81	4.06	2.18
DAG[11] (w/o temp. c.)	87M	12s	0.2212	53.94	36.10	1.46
T-Foley (FiLM[16])	83M	**6.3s**	0.0772	54.59	36.06	**1.94**
T-Foley (TFiLM[18])	101M	13s	0.0469	49.44	36.10	1.74
T-Foley (BFiLM)	**74M**	9.5s	**0.0367**	**41.59**	36.09	1.79

Table 1. Objective evaluations of reproduced DAG(which lacks temporal condition) and our T-Foley conditioned by FiLM, TFiLM and BFiLM. (#params: Number of trainable parameters, infer.t: Approximate inference time for predicting 1 sample, E-L1: Event-L1 Distance, FAD-P, and FAD-V: FADs based on PANNs and VGGish, IS: Inception Score.)

a sound event at t=1.3sec does not affect another sound event at t=3sec.) Therefore, we suggest adopting block-wise transformations from TFiLM by replacing the unnecessary sequential modeling layer with a simple MLP layer as the following:

$$(\gamma_i, \beta_i) = \text{MLP}(Y_{b_i}^{\text{pool}}). \tag{3}$$

T-Foley is designed under the assumption that the LSTM layer located at the bottleneck of U-Net architecture is capable of handling the sequence modeling among the blocks. We demonstrate the performance and efficiency of BFiLM in Section 4.1.

3. EXPERIMENTAL SETUP

3.1. Datasets

We utilized the Foley sound dataset from the Foley sound synthesis task of 2023 DCASE challenge [1]. The given dataset comprises approximately 5k class-labeled sound samples (5.4 hours) sourced from three different datasets. The dataset covers 7 Foley sound classes: *DogBark, Footstep, GunShot, Keyboard, MovingMotorVehicle, Rain, Sneeze_Cough*. All audio samples were constructed in mono 16-bit 22,050 Hz format, each lasting for 4 seconds. We used approximately 95% of the development dataset for training and 5% for validation in this work.

While controllable Foley sound synthesis holds great potential, expressing desired temporal events can be non-trivial for users. For more intuitive conditioning, we used human voices that mimic Foley sounds as a reference to extract temporal event conditions. In particular, we used subsets of two vocal mimicking datasets paired with the original target sounds: Vocal Imitation Set [19] and VocalSketch [20]. We adjusted the duration of each audio sample representing the 6 sound classes (excluding Sneeze_Cough) to match the training data.

3.2. Experimental Details

We train our model to estimate the reparameterized score function of a normal transition kernel with variance-preserving cosine scheduling as proposed in [5]. For conditional sampling, we employ DDPM-like discretization of SDE [21] with classifier-free guidance [22]. During the 500-epoch of training, we randomly dropped the conditions $p = 0.1$ for training in the unconditional scenario.

3.3. Objective Evaluation

Objective Evaluation of audio generation models relies on three metrics. FAD and IS measure that the generated sounds align with

Model	Category Fidelity↑	Temporal Fidelity↑	Audio Quality↑
FiLM[16]	3.85(±0.12)	4.11(±0.10)	3.28(±0.11)
TFiLM[18]	4.02(±0.11)	4.00(±0.13)	3.75(±0.11)
BFiLM	**4.22**(±0.11)	**4.41**(±0.09)	**4.06**(±0.10)

Table 2. Comparison of Mean Opinion Scores(MOS). Mean and 95% confidence interval are reported.

the given class condition and their diversity. For FAD, we exploit two classifiers from VGGish model[23](FAD-V, 16kHz) and PANNs[24](FAD-P, 32kHz). IS also utilizes PANNs. To verify the effectiveness of the temporal condition, we exploit ***Event-L1 Distance (E-L1)***. E-L1 assesses how well the generated sounds adhere to the given temporal event condition. We employed the L1 distance between the event timing feature of the target sample and the corresponding generated sample as follows:

$$E\text{-}L1 = \frac{1}{k}\Sigma_{i=1}^{k}||E_i - \hat{E}_i|| \tag{4}$$

where E_i is the ground-truth event feature of i-th frame, and $\hat{E}_i$ is the predicted one. We average the class-wise scores in each case.

3.4. Subjective Evaluation

A total of 23 participants conducted subjective evaluations, providing assessments for two types of generated samples. These types include: 1) samples generated with temporal conditions from Foley sound test dataset, and 2) samples generated with temporal conditions from human vocals mimicking Foley sounds. Participants rated the generated samples on a scale from 1 to 5, in 0.5-point increments based on three criteria: *Temporal Fidelity (TF)* for alignment of the generated samples with the temporal event condition of the target sample, *Category Fidelity (CF)* for suitability to the given category, and *Audio Quality (AQ)* for the overall quality of the generated sample. We report the Mean Opinion Score (MOS) derived from the participants' ratings.

4. RESULTS

4.1. Temporal Event Conditioning Methods

We compare the performance of different conditioning methods for the temporal event condition. Table 1 and 2 present the objective scores and MOS (Mean Opinion Score) from subjective evaluation. Overall, TFiLM and BFiLM, which consider the temporal aspect of the event condition, received higher scores in most of the objective and subjective metrics. Notably, BFiLM demonstrated markedly superior performance in most of the results, particularly achieving improvements with approximately 0.7 times fewer parameters and less inference time compared to TFiLM. These results validate the hypothesis on the efficiency of BFiLM (Section2.3). FiLM may have high IS value due to generating diverse low-quality audio different from ground-truths. In other experiments, we only use BFiLM.

4.2. Effect of Block Numbers

Block Number N in Section 2.3 is an important hyperparameter that influences the resolution of the condition. Fewer blocks lead to sparser and smoother conditional information in the temporal axis. We compare the performance of different block numbers in Figure 3. For accuracy, E-L1 decreases as there are more blocks as expected.
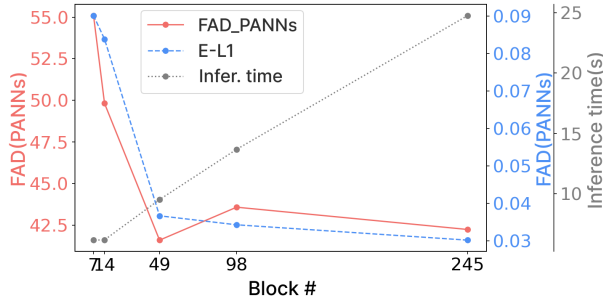
6822

Fig. 3. Tradeoff between performance(E-L1, FAD-P) and efficiency(inference time) among different number of blocks.

Block #	Vocal Imitation		VocalSketch	
	E-L1↓	FAD-P↓	E-L1↓	FAD-P↓
245	0.0228	74.94	0.0186	59.33
98	0.0300	69.13	0.0274	56.51
49	0.0306	64.55	0.0299	49.62
14	0.0764	59.56	0.0635	58.93
7	0.0935	66.23	0.0914	60.41

Table 3. Evaluation on Vocal Mimicking Datasets

FAD-P also decreases, without significant difference among 49, 98, and 245. In terms of efficiency, inference time increases along with the block number, as it requires more computation. Considering the tradeoff between accuracy and efficiency, we stick to 49 blocks in other experiments.

4.3. Evaluation on Vocal Mimicking Datasets

Table 3 summarizes the results among different block numbers, showing comparable performance in vocal sound. E-L1 decreases for larger block numbers, which is consistent with Section 4.2. On the other hand, FAD-P is the lowest around 14, 49 number of blocks. This result may arise from the differences between Foley sound and vocal, as the two sounds exhibit distinct RMS curves due to discrepancies in timbre and energy envelope. Therefore, choosing the right block number is crucial for adjusting the smoothness of RMS to match the conditioning feature with the target sound's characteristics. MOS for model with $N = 49$ was measured as CF= 4.41($\pm$0.09), TF= 4.40($\pm$0.10), and AQ= 4.34($\pm$0.08), providing a competitive result with Table 2.

4.4. Case Study

To showcase the performance and usability of T-Foley, we present two case studies. Firstly, we compare the output of our proposed BFiLM method with that of the FiLM and TFiLM methodologies. In Figure 4, BFiLM exhibits the highest alignment with the mel-spectrogram of the target conditioning sound. Both FiLM and TFiLM generate unclear and undesirable sound events in the Footstep sound class. For Gunshot, only BFiLM seems to reflect the sustain and decay of the attack in sound well.

Furthermore, the generation of Foley sounds using temporal event conditions yields considerably more realistic results when compared to manual Foley sound manipulation. We exemplify two specific scenarios in Figure 5. The first scenario involves consecutive machine gunshots. Manually adjusting and copying individual
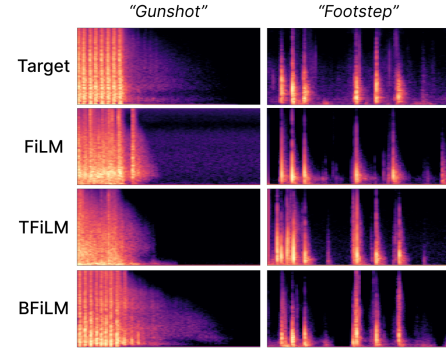

Fig. 4. The first row shows the control sounds used to extract the target event feature. Subsequent rows show three classes of Foley sounds generated with different conditioning blocks (FiLM, TFiLM, and BFiLM), all represented as mel-spectrograms.
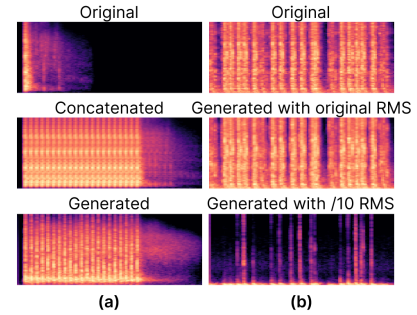

Fig. 5. (a) Comparing manually synthesized consecutive gunshot sounds with sounds generated through temporal event feature. (b) Generated sounds with the original temporal event features and those with a reduced gain by 10.

gunshot sound snippets can result in an unnatural audio sequence. Conversely, employing T-Foley to concatenate temporal event conditions leads to a seamless and lifelike sound. The second is a key-typing sound with two contrasted examples: typing vigorously on a typewriter and softly pressing keys on a plastic keyboard. T-Foley can generate these two sounds by adjusting the gain of the temporal event feature. This indicates that the level of temporal event feature controls not only the amplitude but also the timbre texture. In addition to the showcased examples, you can explore a wider array of demonstration samples and real-world scenarios (such as claps and voices) on our accompanying website introduced in the abstract.

5. CONCLUSION

This study presents T-Foley, a Foley sound generation system addressing controllability in the temporal domain. By introducing the temporal event feature and the E-L1 metric, we show that Block-FiLM, our proposed conditioning method, is effective and powerful in terms of quality and efficiency. We also demonstrate the performance of T-Foley on vocal mimicking datasets to claim its power in usability. Foley sounds can be broadly categorized into two groups: transient event-based sounds and continuous ambient sounds, each with distinct temporal characteristics. While we have not specifically differentiated between the two, we believe acknowledging and addressing this distinction could enhance performance.

6823

6. REFERENCES

[1] Keunwoo Choi, Jaekwon Im, Laurie M. Heller, Brian McFee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinnosuke Takamichi, "Foley sound synthesis at the DCASE 2023 challenge," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 16–20.

[2] Marco Comunità, Huy Phan, and Joshua D. Reiss, "Neural synthesis of footsteps sound effects with generative adversarial networks," in *Audio Engineering Society Convention 152*, May 2022.

[3] M Mehdi Afsar, Eric Park, Étienne Paquette, Gauthier Gidel, Kory W Mathewson, and Eilif Muller, "Generating diverse realistic laughter for interactive art," *arXiv preprint arXiv:2111.03146*, 2021.

[4] Javier Nistal, Stefan Lattner, and Gaël Richard, "DRUMGAN: synthesis of drum sounds with timbral feature conditioning using generative adversarial networks," in *Proc. of the 21th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2020, pp. 590–597.

[5] Simon Rouard and Gaëtan Hadjeres, "Crash: raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2021, pp. 579–585.

[6] Sanchita Ghose and John Jeffrey Prevost, "Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1895–1907, 2020.

[7] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens, "Conditional generation of audio from video via foley analogies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2426–2436.

[8] Xubo Liu, Turab Iqbal, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang, "Conditional sound generation using neural discrete time-frequency representation learning," in *IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6.

[9] Sanchita Ghose and John J Prevost, "Foleygan: Visually guided generative adversarial network-based synchronous sound generation in silent videos," *IEEE Transactions on Multimedia*, 2022.

[10] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan, "Generating visually aligned sound from videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.

[11] Santiago Pascual, Gautam Bhattacharya, Chunghsin Yeh, Jordi Pons, and Joan Serrà, "Full-band general audio synthesis with score-based diffusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[13] 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," *International Conference on Machine Learning*, 2023.

[14] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "Audiogen: Textually guided audio generation," in *The Eleventh International Conference on Learning Representations*, 2022.

[15] Chenye Cui, Zhou Zhao, Yi Ren, Jinglin Liu, Rongjie Huang, Feiyang Chen, Zhefeng Wang, Baoxing Huai, and Fei Wu, "Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[16] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[17] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan, "Wavegrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, 2020.

[18] Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei W Koh, and Stefano Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations.," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan, "Vocal Imitation Set: a dataset of vocally imitated sound events using the audioset ontology," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018, pp. 148–152.

[20] Mark Cartwright and Bryan Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 43–46.

[21] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21696–21707, 2021.

[22] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.