DIFF-FOLEY: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models

Simian Luo[1] Chuanhao Yan[1] Chenxu Hu[1] Hang Zhao[1,2*]
[1]IIIS, Tsinghua University [2]Shanghai Qi Zhi Institute

Abstract

The Video-to-Audio (V2A) model has recently gained attention for its practical application in generating audio directly from silent videos, particularly in video/film production. However, previous methods in V2A have limited generation quality in terms of temporal synchronization and audio-visual relevance. We present DIFF-FOLEY*, a synchronized Video-to-Audio synthesis method with latent diffusion model (LDM) that generate high quality audio with improved synchronization and audio-visual relevance. We adopt contrastive audio-visual pretraining (CAVP) to learn more temporally and semantically aligned features, then train an LDM with CAVP aligned visual features on spectrogram latent space. The CAVP aligned features enable LDM to capture the subtler audio-visual correlation via cross-attention module. We further significantly improve sample quality with 'double guidance'.* DIFF-FOLEY *achieves state-of-the-art V2A performance on current large scale V2A dataset. Our demos are available:* https://sinishell2.github.io/Diff-Foley.github.io/

1. Introduction

Recent advances in generative models have accelerated the development of AI-Generated Content. Progress has been made in various multi-modal generation tasks like Text-to-Image (T2I) [13, 14], Text-to-Audio (T2A) [10, 11], and Text-to-Video (T2V) [5]. This paper focus on Video-to-Audio (V2A) Generation, which has practical applications in video/film production and automatic foley.

Unlike text-based generative models requiring lots of hard-to-collect text-data pairs for training, audio-video pairs for V2A tasks are readily available with millions of new videos uploaded to YouTube daily. The challenge is to design a reliable and scalable generative model for V2A.

V2A generation has always been a challenging problem. First, the generated audio should match the video content. Second, the generated audio should be in sync with video because humans are sensitive to the synchronicity between audio and video. Although some progress [2, 8] has been made recently on V2A, most methods for generating audio focus only on the content relevance, neglecting crucial aspect of audio-visual synchronization. For example, given a video of playing drums, current methods can only generate drums sound, but cannot ensure the sounds match exactly with what's happening in the video (e.g hitting the snare drum or the crash cymbal at the right time)

RegNet [2] uses a pretrained (RGB + Flow) network as conditional inputs to GAN for synthesizing sounds. Meanwhile, SpecVQGAN [8] uses Transformer-based autoregressive model conditioned on pretrained ResNet50 or (RGB + Flow) visual features for better sample quality. These methods have limitations in generating audio that is both synchronized and relevant to video content as pretrained image and optical-flow features cannot capture the nuanced correlation between audio and video.

We present DIFF-FOLEY, a novel V2A generative framework based on LDM that synthesizes realistic and synchronized audio with strong audio-visual relevance. *Foley*, which adds synchronized and realistic sound effects to video, is a more challenging task in V2A. Our model overview is shown in Figure 1. It first learns more temporally and semantically aligned features via CAVP. By maximizing similarity of visual and audio features in the same video, it captures subtle audio-visual connections. Second, an LDM conditioned on CAVP visual features is trained on Spec. latent space. CAVP aligned visual features helps LDM in capturing audio-visual relationships. To further improve sample quality, we propose *'double guidance'*, using classifier-free and alignment classifier guidance simultaneously to guide reverse process. DIFF-FOLEY achieves state-of-the-art performance on large scale V2A dataset VGGSound [1] with IS of 60.39, outperforming SpecVQGAN [8] baseline (IS of 30.01) by a large margin.

2. Method

2.1. Audio-Visual Contrastive Pretraining

The audio and visual components from the same video are strongly correlated and complement each other. When

*Corresponding Author.

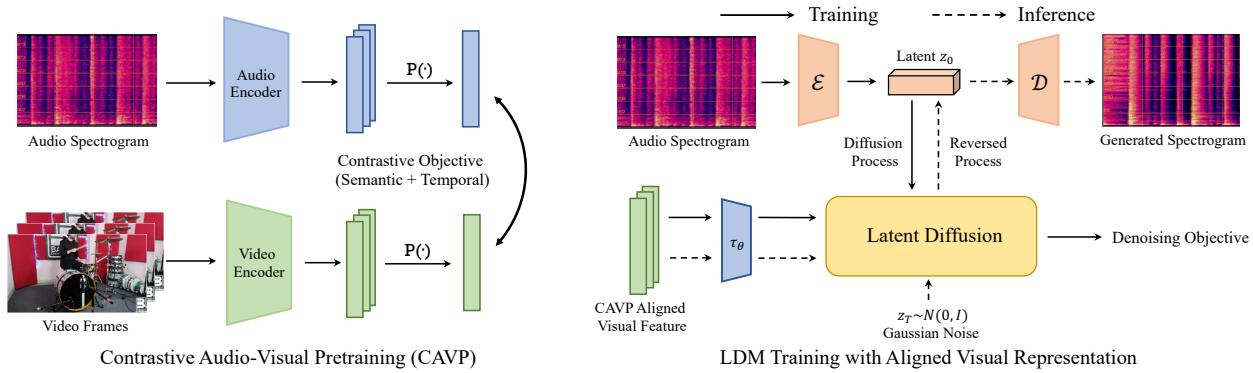

Figure 1. Overview of DIFF-FOLEY: First, it learns more semantic and temporal aligned audio-visual features by CAVP, capturing the subtle connection between audio-visual modality. Second, a LDM conditioned on the aligned CAVP visual features is trained on the spectrogram latent space. DIFF-FOLEY can synthesize highly synchronized audio with strong audio-visual relevance. $\mathcal{P}(\cdot)$ denotes pooling layer.

watching a silent drumming video, viewers can easily imagine corresponding sounds, and expect different sounds when drummers hit different parts of the drums. Unfortunately, Current image and optical flow backbones (ResNet, CLIP, etc.) struggle to reflect the strong alignment relationship between audio and visual. To overcome this limitation, we propose Contrastive Audio-Visual Pretraining (CAVP) to align audio-visual features at the outset, facilitating the subsequent generation process.

Given a audio-video pair (x_a, x_v), where $x_a \in \mathbb{R}^{T' \times M}$ is a Mel-Spec. with M mel basis and T' is the time dimension. $x_v \in \mathbb{R}^{T'' \times 3 \times H \times W}$ is a video clip with T'' frames. An audio encoder $f_A(\cdot)$ and a video encoder $f_V(\cdot)$ are used to extract audio feature $E_a \in \mathbb{R}^{T \times C}$ and video feature $E_v \in \mathbb{R}^{T \times C}$ with same temporal dim T. We adopt the design of audio encoder in PANNs [9], and SlowOnly architecture for video encoder. Using temporal pooling layer $P(\cdot)$, we obtain the temporal-pooled audio/video features, $\bar{E}_a = P(E_a) \in \mathbb{R}^C, \bar{E}_v = P(E_v) \in \mathbb{R}^C$. We then use the cross-entropy symmetric objective similar in CLIP [12] to contrast $\bar{E}_a$ and $\bar{E}_v$. To improve semantic and temporal alignment of audio-video features, we use two objectives: Semantic contrast $\mathcal{L}_\mathcal{S}$ and Temporal contrast $\mathcal{L}_\mathcal{T}$.

For $\mathcal{L}_\mathcal{S}$, we maximize the similarity of audio-visual pairs from the same video, and minimize the similarity of audio-visual pairs from different videos. It encourages learning semantic alignment for audio-visual pairs across different videos. In specific, we extract audio-visual features pairs from *different* videos, $\mathcal{B}_\mathcal{S} = \{(\bar{E}_a^i, \bar{E}_v^i)\}_{i=1}^{N_S}$, where N_S is the number of different videos. We define the per-sample pair semantic contrast objective: ${\mathcal{L}_\mathcal{S}}^{(i,j)}$, where $sim(\cdot)$ is the cosine similarity.

$$
\begin{aligned}
{\mathcal{L}_\mathcal{S}}^{(i,j)} = &-\frac{1}{2}\log\frac{\exp\left(sim(\bar{E}_a^i, \bar{E}_v^j)/\tau\right)}{\sum_{k=1}^{N_S}\exp\left(sim(\bar{E}_a^i, \bar{E}_v^k)/\tau\right)} \\
&-\frac{1}{2}\log\frac{\exp\left(sim(\bar{E}_a^i, \bar{E}_v^j)/\tau\right)}{\sum_{k=1}^{N_S}\exp\left(sim(\bar{E}_a^k, \bar{E}_v^j)/\tau\right)}
\end{aligned} \tag{1}
$$

For $\mathcal{L}_\mathcal{T}$, we sample video clips at different times within the *same* video. It aims to maximize the similarity of audio-visual pairs from the same time segment, and mininize the similarity of audio-visual pairs across different time segments. In details, we sample different time segements in the *same* video to extract audio-visual features pairs. $\mathcal{B}_\mathcal{T} = \{(\bar{E}_a^i, \bar{E}_v^i)\}_{i=1}^{N_T}$, where N_T is the number of sampled video clip within same video. We define the per sample pair temporal contrast objective: ${\mathcal{L}_\mathcal{T}}^{(i,j)}$

$$
\begin{aligned}
{\mathcal{L}_\mathcal{T}}^{(i,j)} = &-\frac{1}{2}\log\frac{\exp\left(sim(\bar{E}_a^i, \bar{E}_v^j)/\tau\right)}{\sum_{k=1}^{N_T}\exp\left(sim(\bar{E}_a^i, \bar{E}_v^k)/\tau\right)} \\
&-\frac{1}{2}\log\frac{\exp\left(sim(\bar{E}_a^i, \bar{E}_v^j)/\tau\right)}{\sum_{k=1}^{N_T}\exp\left(sim(\bar{E}_a^k, \bar{E}_v^j)/\tau\right)}
\end{aligned} \tag{2}
$$

The final objective is defined as the weighted sum of semantic and temporal objective: $\mathcal{L} = \mathcal{L}_\mathcal{S} + \lambda\mathcal{L}_\mathcal{T}$, where $\lambda = 1$. After training, CAVP encodes an audio-video pair into embedding pair: $(x_a, x_v) \rightarrow (E_a, E_v)$ where the (E_a, E_v) are highly aligned, with the visual features E_v containing rich information for audio. The aligned and strongly correlated features E_v and E_a facilitate subsequent audio generation.

2.2. LDM with Aligned Visual Representation

LDMs [14] are probabilistic models that fit the data distribution $p(x)$ by denoising on the data latent space. LDMs first encode the origin high-dim data x into low-dim latent $z = \mathcal{E}(x)$ for efficient training. The forward and reverse process are performed in the compressed latent space. In V2A generation, our goal is to generate synchronized audio x_a given video clip x_v. Using similar latent encoder $\mathcal{E}_\theta$ in [14], we compress Mel-Spec x_a into a low-dim latent $z_0 = \mathcal{E}_\theta(x_a) \in \mathbb{R}^{C' \times \frac{T'}{r} \times \frac{M}{r}}$, where r is the compress rate. With pretrained CAVP model to align audio-visual features, the visual features E_v contain rich audio-related information. This enable to synthesize highly synchronized and relevant audio using LDMs conditioned on E_v. In forward process, origin data distribution transforms into Gaussian

Model	Visual Features	Guidance	Metrics			
			IS ↑	FID ↓	KL ↓	Align Acc (%) ↑
SpecVQGAN [8]	RGB + Flow	✘	30.01	8.93	6.93	52.94
Diff-Foley (Ours)	Audio-Visual Contrastive	CFG Guidance (✔)	52.07	11.61	**6.33**	92.35
Diff-Foley (Ours)	Audio-Visual Contrastive	Double Guidance (✔✔)	**60.39**	10.73	6.42	**94.78**

Table 1. Video-to-Audio generation evaluation results with CFG scale $\omega = 4.5$, CG scale $\gamma = 50$.

by adding noise gradually with a fixed schedule $\alpha_1, \ldots, \alpha_T$, where T is the total timesteps, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

$$\begin{aligned} q(z_t|z_{t-1}) &= \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, (1-\alpha_t)\mathbf{I}) \\ q(z_t|z_0) &= \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_0, (1-\bar{\alpha}_t)\mathbf{I}) \end{aligned} \tag{3}$$

The denoising objective [6] of LDM is:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,t,\epsilon} \|\epsilon - \epsilon_\theta(z_t, t, E_v)\|_2^2 \tag{4}$$

After LDM is trained, we generate audio latent by sampling through the reverse process with $z_T \sim \mathcal{N}(0, \mathbf{I})$, conditioned on the given visual-features E_v.

$$\begin{aligned} p_\theta(z_{t-1}|z_t) &= \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, E_v), \sigma_t^2 \mathbf{I}) \\ \mu_\theta(z_t, t, E_v) &= \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t, E_v) \right) \\ \sigma_t &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t) \end{aligned} \tag{5}$$

At last, the Mel-spectrogram is obtained through decoding the generated latent z_0 with a decoder $\mathcal{D}$, $\hat{x}_a = \mathcal{D}(z_0)$.

2.3. Temporal Split & Merge Augmentation

Using large-scale text-image pairs datasets like LAION-5B is crucial for the success of current T2I models [14]. However, for V2A generation task, large scale and high quality datasets are still lacking. Further, we expect V2A model to generate highly synchronized audio based on visual content, such temporal audio-visual correspondence requires a large amount of audio-visual pairs for training. To overcome this limitation, we propose using *Temporal Split & Merge Augmentation* to facilitate model training by incorporating prior knowledge for temporal alignment into training process. During training, we randomly extract video clips of different time lengths from two videos (*Split*), denoted as (x_a^1, x_v^1), (x_a^2, x_v^2), and extract visual features E_v^1, E_v^2 with pretrained CAVP model. We then create a new audio-visual feature pairs for LDM training with:

$$z_a^{new} = \mathcal{E}_\theta([x_a^1;\, x_a^2]) \quad , \quad E_v^{new} = [E_v^1;\, E_v^2] \tag{6}$$

, where $[\cdot;\cdot]$ represent temporal concatenation (*Merge*). Split and merge augmentation greatly increase the number of audio-visual pairs, preventing overfitting and facilitating LDM to learn temporal correspondence.

2.4. Double Guidance

Guidance techniques is widely used in diffusion model reverse process for controllable generation. There are currently two main types of guidance techniques: classifier guidance [3] (CG), and classifier-free guidance [7] (CFG). For CG, it additionally train a classifier (e.g class-label classifier) to guide the reverse process at each timesteps with gradient of class label loglikelihood $\nabla_{x_t} \log p_\phi(y|x_t)$. For CFG, it does not require an additional classifier, instead it guides the reverse process by using linear combination of the conditional and unconditonal score estimates [7], where the c is the condition and ω is the guidance scale.

$$\tilde{\epsilon_\theta}(z_t, t, c) \leftarrow (1+\omega)\epsilon_\theta(z_t, t, c) - \omega\epsilon_\theta(z_t, t, \varnothing) \tag{7}$$

Although CFG is currently the mainstream approach used in diffusion models, the CG method offers the advantage of being able to guide any desired property of the generated samples given true label. In V2A setting, the desired property refers to semantic and temporal alignment. Moreover, we discover that these two methods are not mutually exclusive. We propose a *double guidance* technique that leverages the advantages of both CFG and CG methods by using them simultaneously at each timestep in the reverse process. In specific, for CG we train an alignment classifier $P_\phi(y|z_t, t, E_v)$ that predict whether a audio-visual pair is a real pair in terms of semantic and temporal alignment. For CFG, during training, we randomly drop condition E_v with prob. 20%, to train conditional and unconditional likelihood $\epsilon_\theta(z_t, t, E_v), \epsilon_\theta(z_t, t, \varnothing)$. Then *double guidance* is achieved by improved noise estimation:

$$\begin{aligned} \hat{\epsilon}_\theta(z_t, t, E_v) \leftarrow & (1+\omega)\epsilon_\theta(z_t, t, E_v) - \omega\epsilon_\theta(z_t, t, \varnothing) \\ & - \gamma\sqrt{1-\bar{\alpha}_t}\nabla_{z_t} \log P_\phi(y|z_t, t, E_v) \end{aligned} \tag{8}$$

, where ω, γ is the CFG, CG guidance scale.

3. Experiments

Datasets In this paper, we use two datasets VGGSound [1] and AudioSet [4]. VGGSound consists of ∼200K 10-seconds videos. We follow the original VGGSound train/test splits. AudioSet comprises 2.1M videos with 527 sound classes, but it is highly imbalanced, with most of the videos labeld as Music and Speech. Since generating meaningful speech directly from visual features is not expected in V2A tasks (not necessary either), we download a subset of the Music tag data and all other tags except Speech, resulting in a new dataset named AudioSet-V2A with about 80K music tagged videos and 310K other tagged videos. We use

Model	Stage1 Pretrained Dataset	CFG Guidance	CG Guidance	Metrics			
				IS ↑	FID ↓	KL ↓	Align Acc (%)↑
Diff-Foley (Ours)	VGGSound	✘	✘	19.86	18.45	6.41	67.59
	VGGSound	✔	✘	51.42	11.48	6.48	85.88
	VGGSound	✔	✔	53.45	**10.67**	6.54	89.08
	VGGSound + AudioSet-V2A	✘	✘	22.07	18.20	6.52	69.41
	VGGSound + AudioSet-V2A	✔	✘	52.07	11.61	**6.33**	92.35
	VGGSound + AudioSet-V2A	✔	✔	**60.39**	10.73	6.42	**94.78**

Table 2. Abalation Study: The effect of different Stage1 pretrained dataset and different guidance strategies. $\omega = 4.5$, $\gamma = 50$

VGGSound and AudioSet-V2A for Stage1 contrastive pre-training, while for Stage2 LDM training and evaluation, we only use VGGSound, which is consistent with the baseline.

Evaluation Metrics For evaluation, we adopt Inception Score (IS), Frechet Distance (FID) and Mean KL Divergence (MKL) proposed in [8]. IS evaluates both sample quality and diversity, FID evaluates distribution-level similarity between generated and ground-truth samples, and MKL measures paired sample level similarity. We also introduce a new metric, Alignment Accuracy (Align Acc), to assess synchronization and audio-visual relevance quality. We train an alignment classifier to predict whether the audio-visual pairs is the real pairs, During training, we use three different types of pairs: 50% of the pairs are real audio-visual pairs (*true pair*) labeled as 1, 25% are audio-visual pairs from the same video but temporally shifted (*temporal shifted pair*) labeled as 0, and the remaining 25% are audio-visual pairs from different videos (*wrong pair*) labeled as 0. Our alignment classifier can reach 90% accuracy on test set. To better evaluate the generated sample quality, we select IS and Align Acc as the primary evaluation metrics in this paper. For each video sample in testing set, we generate 10 audios for evaluation, resulting in around 145K generated audio samples.

Baseline We compare Diff-Foley to SpecVQGAN [8], a state-of-the-art V2A model.We used pre-trained SpecVQGAN models trained on VGGSound and chose the best-performing visual feature setting (RGB+Optical Flow).

3.1. Video-to-Audio Generation Results

Quantitative evaluation results on VGGSound test set are in Table 1. Diff-Foley significantly outperforms baseline method in IS, MKL and Align Acc, while maintaing comparable performance on FID. Notably, Diff-Foley achieves twice the performance of baseline on IS (*60.39 v.s 30.01*). Moreover, Diff-Foley achieves an impressive 94.78% Align Acc, compared to baseline's 52.94% Align Acc. Generated results are available at the demo link [1].

[1] https://sinishell2.github.io/Diff-Foley.github.io/, recommending to use Chrome browser.

3.2. Ablation Study

We conduct extensive ablation study on Diff-Foley in Table 2, exploring various Stage 1 pretrained datasets and guidance techniques. 1). Performance improves with more pre-training data, as shown by rows 1-4, 2-5, and 3-6. 2). Guidance techniques significantly enhance model performance in all metrics except KL. 3). *Double guidance* techniques achieve the best performance on IS and Align Acc, the primary audio quality metrics of interest.

4. Conclusion

We introduce Diff-Foley, a V2A approach for generating highly synchronized audio with strong audio-visual relevance. We empirically demonstrate the superiority of our method in terms of generation quality. Moreover, we show that using *double guidance* technique to guide the reverse process in LDM can further improve the audio-visual alignment of generated audio samples. Additionally, we conduct an ablation study, analyzing the effect of pretrained dataset size and various guidance techniques.

References

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 3

[2] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29:8292–8302, 2020. 1

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[4] 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. 3

[5] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen

video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

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

[8] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021. 1, 3, 4

[9] 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. 2

[10] 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

[11] 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. 1

[12] 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. 2

[13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[14] 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. 1, 2, 3