Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

FoleyMaster: High-Quality Video-to-Audio Synthesis via MLLM-Augmented Prompt Tuning and Joint Semantic-Temporal Adaptation

Liming Liang[1], *Luo Chen*[1], *Yuehan Jin*[2], *Xianwei Zhuang*[1], *Yuxin Xie*[1], *Yongkang Yin*[1], *Yuexian Zou**[1]

[1]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University, China; [2]South China University of Technology, China

limingliang@stu.pku.edu.cn,crystachen77@gmail.com,zouyx@pku.edu.cn

Abstract

We study video-to-audio (V2A) generation, a critical task for automatically creating high-quality sound effects synchronized with silent video. Current V2A methods face three limitations: (1) inadequate textual annotations in existing datasets, (2) over-reliance on global video features, and (3) coarse temporal synchronization. To address these, we propose FoleyMaster with three key innovations: 1) We introduce VGGSound Plus dataset with 197,955 videos annotated by Qwen2-VL-7B for fine-grained event descriptions; 2) We develop a cross-attention semantic adapter integrating token-level text embeddings with global video features via prompt learning, enabling precise alignment between visual events and sound; 3) We develop a probabilistic temporal adapter that adjusts audio generation based on action prominence replacing binary synchronization. Extensive experiments demonstrate that FoleyMaster achieves state-of-the-art V2A performance across all metrics. Demo and dataset are available.

Index Terms: Video-to-Audio Generation, Dataset Construction, Semantic and Temporal Alignment, Prompt Learning.

1. Introduction

Foley is the process of adding realistic and synchronized sound effects to videos [1], widely used in film production. Traditional Foley relies on professional sound artists who manipulate objects in a recording studio to create sound effects synchronized with video actions. Although this process is artistic, it is time-consuming and labor-intensive. In contrast, neural Foley leverages artificial intelligence to generate high-quality synchronized audio, significantly speeding up video production and reducing manual labor. This task, known as video-to-audio (V2A) generation, has gained increasing attention with recent advances in generative artificial intelligence [2, 3, 4, 5, 6, 7, 8, 9, 10].

Previous studies [11, 12, 13, 14, 15] have explored various approaches for video-to-audio generation. SpecVQGAN [12] uses a cross-modal Transformer to auto-regressively generate audio from video tokens, achieving the first end-to-end video-to-audio synthesis. Im2Wav [13] generates autoregressive audio tokens from CLIP [16] features, enabling image-guided open-domain audio generation. Diff-Foley [11] improves semantic and temporal synchronization by pre-training on aligned video-audio data through contrastive learning. FoleyCrafter [14] combines a semantic adapter and temporal controller with a pre-trained text-to-audio model to produce high-quality, video-synchronized sound effects. SVA [15] utilizes key frames to understand video semantics, generating creative audio schemes that guide text-to-audio models via natural language interfaces.

However, current V2A methods face several challenges:

1. Insufficient textual annotations: Publicly available audio-visual datasets often lack detailed semantic annotations. For example, VGGSound [17] provides labels such as “YouTube ID, start seconds, label, train/test split,” while AudioSet [18] uses coarse labels as well. This lack of detailed descriptions limits the semantic grounding of V2A models, hindering their ability to generate audio that aligns accurately with video content. For instance, a video labeled “waterfall burbling” may also contain additional events, like a bird flying by. Training with VGGSound labels would cause the loss of important semantic information, such as the presence of the “bird”.

2. Over-reliance on global features. Most existing V2A methods focus on global video representations, neglecting local event details. Diff-Foley [11] trains on randomly cropped audio-video segments, discarding fine-grained event information and resulting in coarse-grained audio synthesis. Similarly, FoleyCrafter [14] averages video features extracted using CLIP [16], which reduces sensitivity to specific visual events and weakens video-audio alignment.

3. Coarse temporal detection. Current V2A systems use simplistic time detection mechanisms. FoleyCrafter [14], for instance, applies a binary classifier (-1 or 1) to determine whether audio should be generated at a given time frame. This rigid approach fails to capture varying levels of action prominence, limiting the model’s ability to produce dynamically adjusted sound.

4. Disjoint training of semantic and temporal modules. Existing approaches, such as FoleyCrafter, train the semantic adapter and temporal adapter separately, preventing optimal synergy between the two components. When combined, their independently learned representations may not fully align, reducing the overall quality of generated audio.

To address these deficiencies, we first extract the textual annotations from the VGGSound dataset using a strong multimodal large language model (MLLM), Qwen2-VL-7B [19], introducing a new dataset, VGGSound Plus, to resolve the issue of insufficient textual annotations. Building on this, we proposed FoleyMaster. We train a new semantic and temporal adapter jointly on the VGGSound Plus dataset, which enables the model to understand the video from both local and global perspectives, improving the synergy between the two components. For the time detector, we replaced the binary approach with probability values for each frame, allowing the model to emphasize or attenuate the impact of the prominence of actions in the video. In addition, we propose in V2A tasks the use of MLLM-augmented prompt tuning strategy [20, 21] to improve text notation and video-and-audio consistency for the first time.

In summary, our contributions are as follows:

- We introduce and release **VGGSound Plus**, a dataset annotated using a powerful multimodal large language model, of-

[1] https://foleymaster.github.io/

*Yuexian Zou is the corresponding author.

10.21437/Interspeech.2025-300

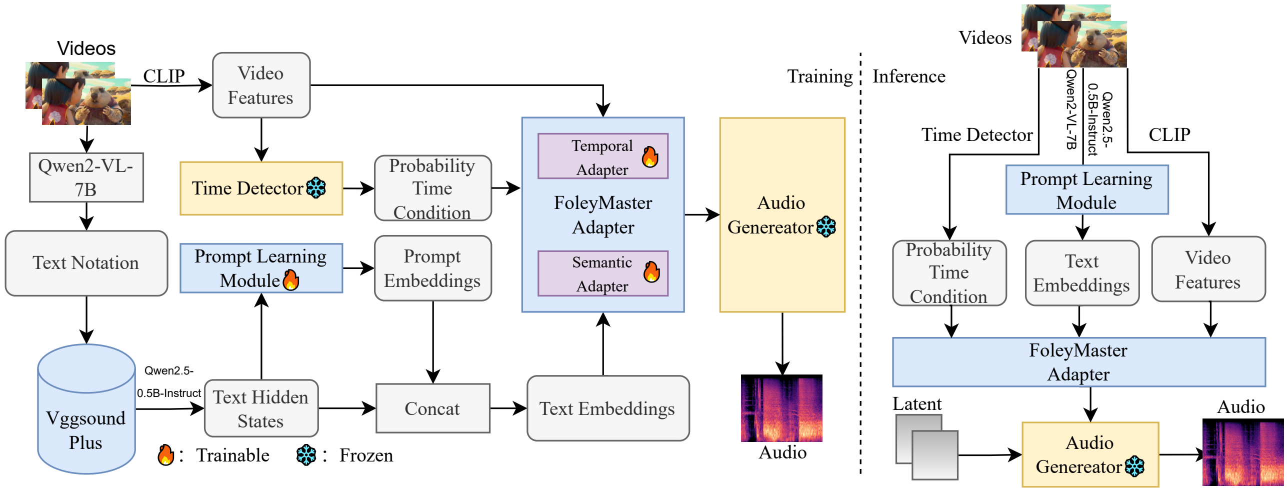

Figure 1: *Illustration of the proposed FoleyMaster Framework.*

Table 1: *Example of a dataset sample*

Field	**Example Content**
Video	NaDw52ggC7g_000209.mp4
Description	Person typing on a keyboard while speaking. Background noise includes faint electrical hum and ambient sound......

fering detailed textual descriptions for each video.

- We develop a new adapter tailored to the V2A problem, leveraging the VGGSound Plus dataset to significantly enhance V2A performance, improving both audio quality and the alignment between audio and video events.
- We implement a **prompt-learning** approach to optimize text-video interactions. Our experimental results demonstrate that this technique improves the model's ability to understand and generate semantically accurate audio, further boosting V2A performance.

2. Proposed Methods

2.1. Proposed VGGSound Plus Dataset

We first used `Qwen2-VL-7B-Instruct` to extract textual annotations for 197,955 videos from the VGGSound dataset. The newly annotated files, along with the original videos, together constitute the VGGSound Plus dataset. An example of the VGGSound Plus dataset is shown in Table 1. The creation of this dataset took approximately 144 NVIDIA A100 GPU hours. We make this dataset publicly available online for the convenience of future researchers.

2.2. Overall Model Architecture

The FoleyMaster framework, as illustrated in Figure 1, integrates multimodal learning for effective video-to-audio (V2A) generation. In training, the model processes video frames through the CLIP encoder to extract video features, while Qwen2-VL-7B provides textual annotations. These text embeddings are enhanced by a prompt learning module and fused with video features using a cross-attention mechanism [22, 23, 24] within the FoleyMaster Adapter, which includes both semantic and temporal adapters. The final audio is generated using a audio generator [25]for high-quality sound synthesis.

During inference, the framework follows a similar flow, generating audio by applying the learned prompt learning module and the adapter with denoising model to the input latent, which is sampled from a standard Gaussian distribution, ensuring high-quality and synchronized sound.

2.3. MLLM-Augmented Prompt Tuning

To enhance the effectiveness of textual prompts in the video-to-audio (V2A) generation process, we employ a learnable prompt learning approach. This method enables the model to dynamically optimize the text prompt representations during training, ensuring a more effective fusion of textual and visual features.

Learnable Prompt Representation: Inspired by previous works in prompt learning, such as CoOp [20], we introduce a learnable prompt matrix P_{learn} that is prepended to the token-level features extracted by `Qwen2.5-0.5B-Instruct` [26]. Given a textual description T of length L, we obtain the token feature:

$$T_{\text{HiddenStates}} = \text{Qwen}(T) \in \mathbb{R}^{L \times D}, \tag{1}$$

where D is the hidden states dimension. We then introduce a learnable prompt matrix:

$$P_{\text{learn}} \in \mathbb{R}^{N \times D}, \tag{2}$$

where N is the number of learnable prompt tokens. The updated textual representation is:

$$T_{\text{emb}} = \text{concat}(P_{\text{learn}}, T_{\text{HiddenStates}}) \in \mathbb{R}^{(N+L) \times D}. \tag{3}$$

2.4. FoleyMaster Adapter

Visual Encoder. The CLIP encoder has proven to be a powerful tool for extracting semantic information from visual data [27]. In our method, we adopt the strategies from prior studies [14] to obtain visual embeddings from each frame in the input video using the CLIP image encoder. To ensure these embeddings are compatible with the text-to-audio generator, we utilize multiple learnable projection layers. This process can be represented as follows:

$$V_{\text{emb}} = \text{MLP}(\text{AvgPooling}(\tau_{\text{vis}}(v))).$$

Here, v represents the input video, τ_{vis} denotes the CLIP image encoder, and AvgPooling refers to the average pooling of the extracted CLIP features across frames.

Semantic Adapter: To improve the interaction between token-level textual prompts (local) and video features (global), we employ a Cross-Attention mechanism.

$$Attn = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{4}$$

$$T'_{\text{emb}} = Attn + T_{\text{emb}}, \tag{5}$$

where $Q = T_{\text{emb}}$ represents the text embedding after prompt learning, and $K, V = V_{\text{emb}}$ are the visual features. T'_{emb} is the final fused feature that fully represents the semantic information of the video.

Temporal Adapter: The temporal adapter follows the same structure as the UNet encoder [28] in the text-to-audio generator, inspired by the design of ControlNet [29]. In our method, we adopt the time detector pre-trained in [14]for event detection,but using its probability values as time indicators instead of binary timestamps in order to dynamically adjusts audio generation intensity based on action prominence. Specifically, the temporal adapter leverages the predicted timestamp condition probability to guide the sound generation process at the desired timestamp. It takes both the timestamp probability and the identical latent input as the original UNet, and the resulting output is added as a residual to the output of the original UNet.

2.5. Training and Optimization

During training, the learnable prompt P_{learn} and the FoleyMaster Adapter(both semantic and temporal adapters)are jointly optimized using a diffusion loss function. The objective is to minimize the difference between the predicted noise and the true noise, considering both the temporal and semantic alignments:

$$L = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,c}\left[\left|\left|\epsilon - \epsilon_\theta(z_t, t, T'_{\text{emb}}, C_{\text{time}}, V_{\text{emb}})\right|\right|\right], \tag{6}$$

where:

- ϵ represents the noise added to the latent variable z_t at time step t,
- $\epsilon_\theta(z_t, t, T'_{\text{emb}}, C_{\text{time}}, V_{\text{emb}})$ is the predicted noise from the model, based on the input latent variable z_t, the timestep t, the attention-enhanced text-video representation T'_{emb}, probability time condition C_{time} and video embeddings V_{emb}.
- $\mathcal{N}(0, 1)$ denotes a standard normal distribution.

3. Experiments and Results

3.1. Datasets, Metrics, and Implementation Details

Datasets We use the VGGSound Plus dataset we proposed, an enhanced large-scale audio-visual dataset built upon VGGSound. VGGSound Plus comprises over 190K videos across 310 audio classes with detailed textural annotations, ensuring strong audio-visual correspondence.

Metrics We evaluate the model performance using the CLIP Score (CS) from [30], onset detection accuracy (Onset Acc) and onset detection average precision (Onset AP) from [31], Inception Score (IS), Frechet Distance (FID), and Mean KL Divergence (MKL) from [12]. The CLIP Score first extracts features from multiple consecutive frames of the original video using CLIP[16], and from the generated audio using Wav2CLIP[32]. Then, the cosine similarity between the image and audio CLIP features is calculated, as done in previous works[30, 33]. Onset Acc and Onset AP assess whether the generated sound includes the correct number of onsets and whether their timings match those in the input video[31]. IS evaluates the quality and diversity of the generated audio samples, FID measures the similarity of distributions, and MKL assesses the similarity at the paired sample level[11].

Implementation Details In our experiments, we first sample the training audio at a rate of 16kHz. To extract the audio features, we compute the mel-spectrogram using the following parameters: $n_{\text{fft}} = 2048$, num_mels $= 256$, hop_size $= 160$, win_size $= 1024$, $f_{\min} = 0$, and $f_{\max} = 8000$. These mel-spectrograms are then passed through a pretrained variational autoencoder [34] from the Affusion model[25] to generate the corresponding latent representations, which are used in the subsequent training of the latent diffusion model.

For training, we set the size of the prompt learning vector $N = 16$, which corresponds to the number of context tokens that are used for text-to-audio alignment. During inference, we employ the vocoder as described in [34] to decode the generated latent representations back into audio, ensuring the high-quality reconstruction of the audio from the learned representations.

3.2. Main Experiment

Table 2 and table 3 present a comparison of different models on both semantic and temporal alignment aspects. The proposed **FoleyMaster** model outperforms previous methods in all key metrics.

For **semantic alignment**, Table 2 shows that FoleyMaster achieves the highest CLIP Score (CS) of 12.371, significantly surpassing FoleyCrafter (9.821) and Diff-Foley (8.126). This indicates that our model generates audio that is better aligned with the visual content. Moreover, FoleyMaster obtains the best Inception Score (IS) of 62.073, demonstrating higher diversity and quality in the generated audio. Additionally, FoleyMaster achieves the lowest Frechet Distance (FID) of 9.832 and Mean KL Divergence (MKL) of 4.031, suggesting that our method produces more natural and realistic sound effects compared to previous approaches.

For **temporal alignment**, Table 3 shows that FoleyMaster significantly improves onset detection accuracy (Onset ACC) and onset detection average precision (Onset AP). FoleyMaster attains best performance on Onset ACC (34.36) and Onset AP(79.83), demonstrating superior synchronization between generated audio and video events. These results highlight the effectiveness of our probabilistic temporal adapter in capturing finer temporal variations compared to the binary mask method used in prior works.

Overall, FoleyMaster demonstrates state-of-the-art performance across all evaluated metrics, achieving superior semantic and temporal alignment while maintaining high-quality and diverse audio generation.

Table 2: *Comparison of Different Models in Semantic Aspect*

Model	CS↑	IS↑	FID↓	MKL↓
SpecVQGAN	2.703	12.665(0.416)	23.148	8.544
Diff-Foley	8.126	51.311(0.450)	13.402	6.667
FoleyCrafter	9.821	34.809(2.288)	20.008	5.969
FoleyMaster	**12.371**	**62.073(1.765)**	**9.832**	**4.031**

3.3. User Study

To better evaluate the model's performance, we conducted a user study employing a subjective evaluation method similar to previous approaches[14]. We randomly selected 50 videos from

Table 3: *Comparison of Different Models in Temporal Aspect*

Model	Onset ACC↑	Onset AP↑
SpecVQGAN	29.52	74.82
Diff-Foley	22.97	61.72
FoleyCrafter	28.23	64.04
FoleyMaster(ours)	**34.36**	**79.83**

Table 4: *User Study of Different Models*

Model	Semantic	Temporal	Quality
SpecVQGAN	28.13	37.50	20.63
Diff-Foley	18.75	26.88	16.88
FoleyCrafter	21.25	13.13	30.63
FoleyMaster(ours)	**78.75**	**81.25**	**66.25**

the VGGSound Plus test set and generated corresponding audio samples using different models for a survey. The outputs were anonymized and presented to 20 participants who were unfamiliar with the project. Participants were asked to choose the audio sample that demonstrated better semantic alignment, temporal alignment, and generation quality. The preference score was then calculated as follows: Score $= \frac{S}{A}$. Where S represents the number of times a method was chosen, and A represents the number of times the method appeared.

Table 4 presents the results of the user study. FoleyMaster achieved the highest preference scores across all three evaluation criteria.

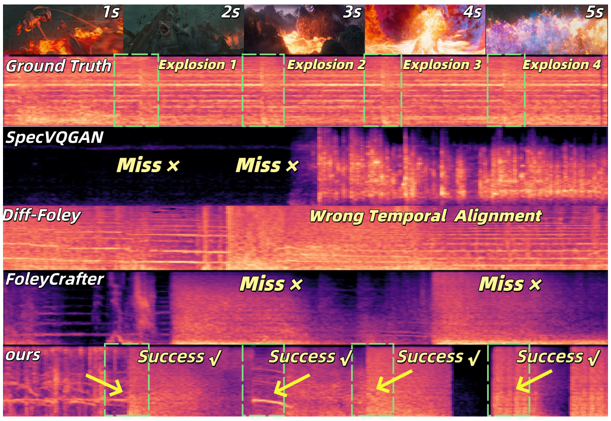

Figure 2: *Qualitative Comparison of Different Models*

3.4. Qualitative Comparison

We provide the visualization of generated audio for qualitative comparison. It can be observed from Figure 2 that FoleyMaster generates sound at the most accurate time aligned with visual cues, closely resembling the pattern of the ground truth audio. In addition, we can see from the figure that FoleyMaster can better understand the semantic information of the video, such as the understanding of explosion sounds, and therefore can generate explosion simulation sounds from high to low frequencies at the 2nd second.

3.5. Ablation Study

To further analyze the impact of key components in our proposed **FoleyMaster** model, we conduct an ablation study focusing on **prompt learning** and **the probabilistic time condition**. The results are presented in Table 5 and Table 6.

Effect of Prompt Learning. To evaluate the effectiveness of prompt learning, we compare FoleyMaster with and without prompt learning applied to the text-video feature fusion process. As shown in Table 5, removing prompt learning module leads to a decrease in CLIP Score (CS) from 12.371 to 10.006, indicating weaker semantic alignment between the generated audio and the video. Additionally, the Inception Score (IS) drops from 62.073 to 53.871, suggesting reduced diversity and quality in the generated audio. Furthermore, the Frechet Distance (FID) increases from 9.832 to 14.315, and the Mean KL Divergence (MKL) rises from 4.031 to 5.546, both implying a deterioration in the naturalness and distribution alignment of the generated audio. These results demonstrate that prompt learning significantly enhances the model's ability to generate audio that is semantically rich and aligned with the video content.

Effect of the Probabilistic Time Condition. We further investigate the impact of replacing the binary time condition with a probabilistic approach. As shown in Table 6, using the binary time condition results in an Onset Accuracy (Onset ACC) of 32.78 and an Onset AP of 75.64. By incorporating a probabilistic time detector, FoleyMaster achieves an Onset ACC of 34.36 and an Onset AP of 79.83, leading to a notable improvement in temporal synchronization. This demonstrates that our proposed method enables finer control over when audio should be generated, effectively enhancing the model's ability to align sound events with visual cues.

Overall, these ablation studies confirm the effectiveness of both prompt learning and the probabilistic time condition, highlighting their contributions to improving FoleyMaster's audio generation performance.

Table 5: *Ablation Study for Prompt Learning*

Methods	CS↑	IS↑	FID↓	MKL↓
w/o prompt learning	10.006	53.871	14.315	5.546
with prompt learning	**12.371**	**62.073**	**9.832**	**4.031**

Table 6: *Ablation Study for Probabilistic Time Condition*

Methods	Onset ACC↑	Onset AP↑
binary time condition	32.78	75.64
FoleyMaster	**34.36**	**79.83**

4. Conclusion

We propose **FoleyMaster**, a novel video-to-audio generation framework that enhances semantic alignment, temporal synchronization, and audio quality. Our contributions include (1) **VGGSound Plus**, a large-scale dataset with detailed multimodal annotations, (2) **A jointly trained semantic and temporal adapter** for precise audio-video alignment, and (3) **Prompt learning** to optimize text-video interactions. Experiments demonstrate that FoleyMaster surpasses state-of-the-art methods across all evaluated metrics. Ablation studies confirm the effectiveness of prompt learning and the probabilistic time condition methods. Our approach advances neural Foley generation, and future work will focus on exploring more diverse soundscapes and further improving model performance.

5. Acknowledgements

This work is supported by Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology. (Grant No. 2024B1212010006)

6. References

[1] V. T. Ament, *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. New York: Routledge, 2014.

[2] Y. Zhang, X. Xu, and M. Wu, "Smooth-foley: Creating continuous sound for video-to-audio generation under semantic guidance," *arXiv preprint arXiv:2412.18157*, 2024.

[3] Z. Chen, P. Seetharaman, B. Russell, O. Nieto, D. Bourgin, A. Owens, and J. Salamon, "Video-guided foley sound generation with multimodal controls," *arXiv preprint arXiv:2411.17698*, 2024.

[4] X. Mei, V. Nagaraja, G. L. Lan, Z. Ni, E. Chang, Y. Shi, and V. Chandra, "Foleygen: Visually-guided audio generation," in *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2024, pp. 1–6.

[5] Y. Ren, C. Li, M. Xu, W. Liang, Y. Gu, R. Chen, and D. Yu, "Sta-v2a: Video-to-audio generation with semantic and temporal alignment," *arXiv preprint arXiv:2409.08601*, 2024.

[6] H. K. Cheng, M. Ishii, A. Hayakawa, T. Shibuya, A. Schwing, and Y. Mitsufuji, "Taming multimodal joint training for high-quality video-to-audio synthesis," *arXiv preprint arXiv:2412.15322*, 2024.

[7] Z. Chen, P. Seetharaman, B. Russell, O. Nieto, D. Bourgin, A. Owens, and J. Salamon, "Video-guided foley sound generation with multimodal controls," *arXiv preprint arXiv:2411.17698*, 2024.

[8] X. Cheng, X. Wang, Y. Wu, Y. Wang, and R. Song, "Lova: Long-form video-to-audio generation," *arXiv preprint arXiv:2409.15157*, 2024.

[9] Q. Yang, B. Mao, Z. Wang, X. Nie, P. Gao, Y. Guo, C. Zhen, P. Yan, and S. Xiang, "Draw an audio: Leveraging multi-instruction for video-to-audio synthesis," *arXiv preprint arXiv:2409.06135*, 2024.

[10] B. Li, F. Yang, Y. Mao, Q. Ye, H. Chen, and Y. Zhong, "Tri-ergon: Fine-grained video-to-audio generation with multi-modal conditions and lufs control," *arXiv preprint arXiv:2412.20378*, 2024.

[11] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] V. Iashin and E. Rahtu, "Taming visually guided sound generation," *arXiv preprint arXiv:2110.08791*, 2021.

[13] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen, "Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7151–7161.

[14] Y. Zhang, Y. Gu, Y. Zeng, Z. Xing, Y. Wang, Z. Wu, and K. Chen, "Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds," *arXiv preprint arXiv:2407.01494*, 2024.

[15] G. Chen, G. Wang, X. Huang, and J. Sang, "Semantically consistent video-to-audio generation using multimodal language large model," *arXiv preprint arXiv:2404.16305*, 2024.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[17] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216522760

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[19] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[20] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *CoRR*, vol. abs/2109.01134, 2021. [Online]. Available: https://arxiv.org/abs/2109.01134

[21] Y. Gu, X. Han, Z. Liu, and M. Huang, "PPT: Pre-trained prompt tuning for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8410–8423. [Online]. Available: https://aclanthology.org/2022.acl-long.576/

[22] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[23] Y. Xie, Z. Zhu, X. Zhuang, L. Liang, Z. Wang, and Y. Zou, "Gpa: global and prototype alignment for audio-text retrieval," in *Proc. Interspeech 2024*, 2024, pp. 5078–5082.

[24] X. Zhuang, Y. Xie, Y. Deng, D. Yang, L. Liang, J. Ru, Y. Yin, and Y. Zou, "Vargpt-v1. 1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning," *arXiv preprint arXiv:2504.02949*, 2025.

[25] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," 2024. [Online]. Available: https://arxiv.org/abs/2401.01044

[26] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[29] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2302.05543

[30] R. Sheffer and Y. Adi, "I hear your true colors: Image guided audio generation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[31] Y. Du, Z. Chen, J. Salamon, B. Russell, and A. 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.

[32] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4563–4567.

[33] H. Wang, J. Ma, S. Pascual, R. Cartwright, and W. Cai, "V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 492–15 501.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114