

---

# YingSound: Video-Guided Sound Effects Generation with Multi-modal Chain-of-Thought Controls

---

Zihao Chen<sup>1</sup>, Haomin Zhang<sup>1</sup>, Xinhan Di<sup>1</sup>, Haoyu Wang<sup>1,3</sup>, Sizhe Shan<sup>1,3</sup>, Junjie Zheng<sup>1</sup>, Yunming Liang<sup>1</sup>, Yihan Fan<sup>1,4</sup>, Xinfu Zhu<sup>1,2</sup>, Wenjie Tian<sup>1,2</sup>, Yihua Wang<sup>1</sup>, Chaofan Ding<sup>1</sup>, and Lei Xie<sup>2</sup>

<sup>1</sup>AI Lab, Giant Network

<sup>2</sup>ASLP Lab, Northwestern Polytechnical University

<sup>3</sup>Zhejiang University

<sup>4</sup>East China University of Science and Technology

## Abstract

Generating sound effects for product-level videos, where only a small amount of labeled data is available for diverse scenes, requires the production of high-quality sounds in *few-shot* settings. To tackle the challenge of limited labeled data in real-world scenes, we introduce *YingSound*, a foundation model designed for video-guided sound generation that supports high-quality audio generation in few-shot settings. Specifically, YingSound consists of two major modules. The first module uses a conditional flow matching transformer to achieve effective semantic alignment in sound generation across audio and visual modalities. This module aims to build a learnable audio-visual aggregator (AVA) that integrates high-resolution visual features with corresponding audio features at multiple stages. The second module is developed with a proposed multi-modal visual-audio chain-of-thought (CoT) approach to generate finer sound effects in few-shot settings. Finally, an industry-standard video-to-audio (V2A) dataset that encompasses various real-world scenarios is presented. We show that YingSound effectively generates high-quality synchronized sounds across diverse conditional inputs through automated evaluations and human studies. Project Page: <https://giantailab.github.io/yingsound/>

## 1 Introduction

In recent years, audio, speech, and music generation development has advanced unprecedentedly. Speech generation has seen significant progress, particularly in two distinct areas: traditional text-to-speech (TTS) systems [1–9] and large-scale speech models [10–18]. Traditional TTS systems are typically trained on limited datasets recorded in studios, lacking high-quality zero-shot capabilities. In contrast, large-scale speech models, trained on extensive datasets, exhibit strong zero-shot performance and deliver superior prosody and naturalness. These advancements collectively contribute to a broad range of applications. Similarly, music generation has emerged as a vibrant field, encompassing both symbolic music generation based on scores or MIDI [19–24] and waveform generation conditioned on text or visual prompt [25–29]. These approaches have shown remarkable potential. However, the development of multi-modal based audio generation remains in its early stages.

Audio generation has immense potential for applications like Foley sound production, gaming, virtual reality (VR), and background audio creation. Advances in generative models have made it possible to synthesize audio from multi-modal (video & audio & text) inputs. For instance, latent diffusion models enable audio synthesis conditioned on text by controlling the diffusion process, while flow matching [30] and consistency models [31] offer alternative approaches to audio modeling.

Furthermore, treating audio generation as a language modeling task [32] has opened new avenues for research and application.

Synthetic audio can become more vivid and contextually rich when guided by images [33] and videos [34]. Moreover, the temporal co-occurrence between video and audio provides a natural alignment for audio synthesis. This makes video a powerful guide to generating audio, enabling applications such as AI-generated video dubbing and Foley sound production for films. These capabilities underline the promising and broad application potential of V2A generation.

In filmmaking, Foley reproduces everyday sound effects added to films, videos, and other media in post-production to enhance audio quality. Traditional Foley requires a significant amount of human effort and time. Recently, with the rapid development of video generation [35–40], designers can quickly generate video clips according to their imagination. However, the generated video clips are often silent without background sound. To address this issue, researchers have started using V2A techniques to generate audio associated with video. Moreover, V2A can be applied to create sound effects for short videos and professional film and television productions.

Initially, text-to-audio (T2A) technology [41–48] is commonly used to generate corresponding sound effects, which are then manually aligned with video. This approach still requires a lot of effort with prompt engineering and manual alignment (both semantically and temporally). In addition, other key factors include sound quality and fine-grained generation.

This paper presents YingSound, a video-guided sound effects generation approach with multi-modal CoT controls. It consists of a transformer-based flow matching module with a learnable AVA connector, a multi-modal visual-audio CoT module, and a corresponding training strategy enabling audio generation towards few-shot settings. Our key contributions are as follows:

- We design a learnable audio-visual aggregator (AVA), a dynamic module that integrates high-resolution visual features with corresponding audio features at multiple stages within the flow matching transformers.
- We propose a multi-modal visual-audio CoT module that generates high-quality audio in few-shot settings for industrial scenes optimized through reinforcement learning.
- We develop an industry-standard video-to-audio (V2A) dataset encompassing various real-world scenarios, including movies, games, and commercials.

## 2 Related Work

**Audio-visual learning.** Audio-visual learning has been a central focus in multi-modal representation research, with numerous studies exploring the relationship between audio and visual data. Some studies explore the semantic connections between sight and sound, focusing on learning shared audio-visual associations [49–51], audio-visual sound localization [52, 53], and audio-visual segmentation [54]. Other research emphasizes the temporal alignment of audio and video events [55, 56]. Similarly, we develop a learnable AVA, a module that integrates high-resolution visual features for V2A generation, which achieves semantic and temporal alignment between audio and vision progressively across multiple stages.

**Video-to-audio Generation.** The current state-of-the-art V2A models can be divided into two main approaches: one involves training end-to-end models directly on large-scale datasets, while the other builds upon the base capabilities of T2A models, enhancing them with temporal and video-semantic alignment. For end-to-end models, two common strategies have emerged. The first employs diffusion models to generate Mel spectrograms guided by visual features directly [57–63]. The second approach uses transformer architectures to autoregressively predict audio tokens, which are then fed into a decoder to generate audio [64–68]. Notable examples of these approaches include Frieren [63], which employs rectified flow matching with a non-autoregressive vector field estimator, integrating feedforward transformers and cross-modal feature fusion to achieve robust temporal alignment. Movie Gen Audio [35] explores generating audio conditioned on video and text. For the two-stage methods, current research [69–76] leverages the capabilities of T2A models, incorporating temporal or semantic alignment to achieve high-quality V2A. For example, SonicVisionLM [69] aims to generate diverse sound effects by utilizing vision-language models (VLMs). MaskVAT [70] proposes a full-band, high-quality general audio codec with a sequence-to-sequence masked generative model.

FoleyCrafter [71] and Rewas [72] leverage a pre-trained text-to-audio model to ensure high-quality audio generation. FoleyCrafter and Rewas adopt time and energy conditions, respectively, to guide sound effects generation. Draw-an-audio [73] proposes the mask-attention module (MAM) that focuses on regions of interest using masked video instructions. At the same time, the time-loudness module (TLM) aligns sound synthesis with the video in terms of loudness and timing. However, the above methods lack exploration of audio generation in few-shot settings. Therefore, we design a multi-modal visual-audio CoT module to facilitate audio generation in such settings.

### 3 Approach

#### 3.1 Overview

We present a V2A model, YingSound, designed for industrial-level audio generation from videos and address challenges such as long-form audio generation with limited annotations. YingSound adopts a few-shot learning approach and involves three key aspects. Firstly, a comprehensive data collection and annotation pipeline is established to build a high-quality task-specific dataset at the industrial-level. Secondly, a learnable AVA is developed, generating audio from video inputs. The AVA is a two-tower transformer network that leverages the capability of a pre-trained T2A model and progressively builds audio-visual alignment across multiple training stages. Thirdly, a CoT module is introduced to refine the audio generated by the AVA in few-shot settings, which produces high-quality, contextually appropriate audio through an expert module with multi-modal reasoning.

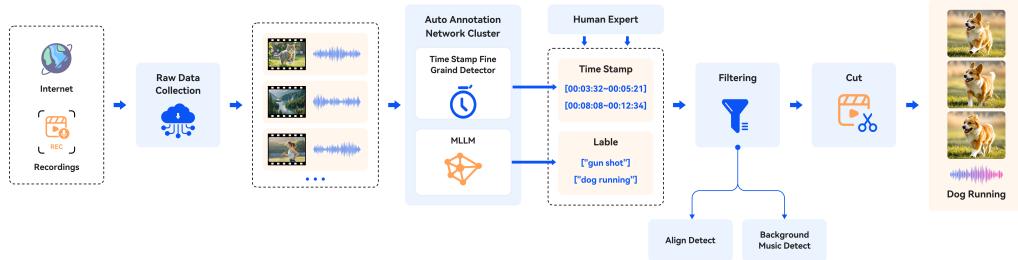


Figure 1: The data collection and processing pipeline with human-in-the-loop.

#### 3.2 Data-Pipeline

As illustrated in Figure 1, we design a comprehensive data processing pipeline that includes collection, annotation, filtering, and cutting. It is essential to integrate human expertise with AI, as the complexity of the entire process and the variation across different video types, relying solely on automation, are insufficient.

**Data Collection and Annotation.** We start by classifying video types and then download or record content from the internet based on these categories to improve diversity. Each video comes with its corresponding audio. To ensure data quality, we resample the audio and convert it to mono. We develop an automated annotation network cluster, comprising a fine-grained timestamp detector and multi-modal large language models (MLLMs) for timestamp extraction and text labeling. We utilize advanced video-large-language models (Video-LLMs) for complex video understanding, accurately annotating multiple sound events. Subsequently, the detected sound events from the previous step are fed into a fine-grained audio timestamp prediction model, producing precise timestamp for seamless integration into downstream tasks. In addition, certain tasks require manual annotation to ensure accuracy and handle complex cases effectively.

**Filtering and Cutting.** Data with low alignment quality, including temporal alignment and video-semantic alignment, or strong background music, is discarded to ensure the overall quality of the dataset. We use the AV-Align [77] metric to filter out data with low temporal alignment scores, ensuring that only high-quality audio-video pairs are retained for further processing. Next, we employ CLIP and CLAP [78] techniques to filter out semantically misaligned data. To eliminate pure background music and speech-involved segments, we utilize background music detection techniques

and automatic speech recognition (ASR) [79–81] to filter out audio with non-relevant content. Finally, after multiple rounds of filtering, we cut the data based on precise timestamp and label annotations, resulting in a high-quality audio-video-text dataset ready for subsequent training.

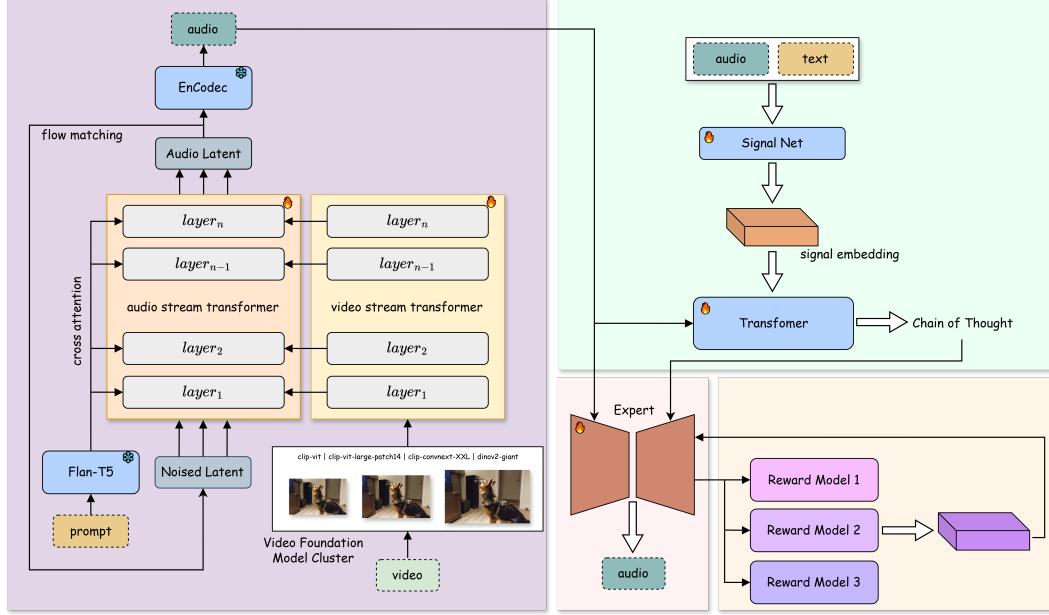


Figure 2: **The overview of the YingSound.** It comprises two key components: Conditional Flow Matching with Transformers and a Multi-modal Chain-of-Thought Based Audio Generation.

### 3.3 Conditional Flow Matching with Transformers

We perform audio generation based on flow matching [82] and a scalable diffusion transformer (DiT) architecture [83], which have been proven to be both high-quality and efficient in image and audio generation. Our model can be conditioned on both text prompts and videos. For text conditions, we adopt an instruction-tuned LLM FLAN-T5 [84] as a text encoder and apply cross attention in every audio DiT layer similar to Tango [85] and some other diffusion-based models. For video conditions, we construct a Video Foundation Model Cluster inspired by [86]. We use different input resolutions and different visual encoders pre-trained, including clip-vit, clip-vit-large-patch14, clip-convnext-XXL [87], dinov2-giant [88] to get frame-level features. A linear layer then projects the clip features to a suitable dimension for the video DiT.

We encode the raw waveform with Encoded [89], which is trained in 24 kHz monophonic audio in diverse domains, including speech, music, and general audio. In particular, we use the features extracted before the residual quantization layer [90, 91, 26].

**Audio-Visual Aggregator Module.** Video frames contain rich information about the audio content of each frame. We use an audio video mapping Module (AVMM) between every DiT layer to perform cross-modality mapping [92]. Given an output of audio/video DiT layer  $i$   $y_a^i, y_v^i$ , the input of the next layer  $x_a^{i+1}, x_v^{i+1}$  is computed as:

$$x_a^{i+1} = y_a^i + \text{Linear}_a(\text{concat}(y_a^i, y_v^i))$$

$$x_v^{i+1} = y_v^i + \text{Linear}_v(\text{concat}(y_a^i, y_v^i))$$

**Multi-Stage Training for the T2A/V2A Generation.** We train our model from T2A to V2A step by step to make the training process more stable.

- First, the model is trained only using the T2A data, which means that the video condition module is omitted, including the video DiT. In this stage, we obtain a sufficiently good T2A model.
- Videos are then added to train the video condition module, and video descriptions are retained during training. The ratio of T2A and T&V2A training data is kept at 1:1 during this stage.
- Finally, we randomly drop some text prompts for videos so that the model can generate audio purely from video. The T2A, T&V2A, and V2A training data ratio is set to 1:1:2 during this stage.

So, the probability of keeping the text conditions for each stage is 100%/100%/50%. And the probability of keeping the video conditions is 0%/50%/75%. Finally, we get a model capable of generating audio based on texts, videos, or both.

### 3.4 Multi-modal Chain-of-Thought Based Audio Generation

To enable audio generation from guided video at an industrial-level, we address the challenge of the high cost associated with fully supervised training samples. We propose a multi-modal CoT generation module [93–96] tailored for few-shot settings. This module comprises two key components. The first component is the multi-modal CoT module, which generates multi-modal CoT based on coarse-level audio outputs. The second component focuses on finer-level audio generation, leveraging the outputs from the coarse audio and the corresponding CoT with the application of reward models [97–103].

**Multi-modal Chain-of-Thought Module.** The multi-modal CoT module [104–106] consists of two components. The first component is a signal module, which receives multi-modal inputs and outputs signal embeddings. The second component is the generation module, which inputs coarse-level audio and signal embeddings and produces the corresponding CoT for finer-level audio. Both components are built with multiple layers of transformers.

**Finer-level Audio Generation.** The audio generation module at the finer-level consists of two components. The first component is a reward module, composed of a set of reward models to generate different rewards based on the audio. The second component is an expert module that generates finer-level audio that aligns with the rewards, corresponding CoT, and coarse-level audio.

**Multi-Stage Training for Adaptive Generation.** To enable adaptive generation in few-shot settings, the training process follows these steps:

- First, we train the two components in a few-shot supervised setting as an initial training step, using supervised samples.
- Second, we proceed with the training stage using generated CoT in an iterative preference learning process to minimize the generation of incorrect CoT.
- Third, we transition to joint training in an intrinsic self-correction setting to improve audio quality iteratively.

## 4 Experiments

### 4.1 Experiment Settings

**Dataset.** We use about 1.2M text-audio pairs, including AudioCaps [107], WavCaps [108], Tango-PromptBank [85], MusicCaps [25], and AF-AudioSet [109], about 3.3k hours, to span the entire three-stage training process. We only use the VGGSound [110] dataset as video-related input in the second and third training stages. Datasets are detailed in Table 1.

**Implementation details.** We train the first stage model using the entire T2A dataset for a total of 250k steps. In the second stage, we utilize the complete VGGSound dataset along with its corresponding T2A data, where both text and audio are used for training. This phase involves a total of 50k training steps. In the third stage, we use a combination of T2A, T&V2A, and V2A datasets in a ratio of 1:1:2,

Table 1: Dataset Details.

Dataset	Traning Stages	Modality	Clip Numbers
AudioCaps [107]	1/2/3	T/A	49k
WavCaps [108]	1/2/3	T/A	402k
TangoPromptBank [85]	1/2/3	T/A	37k
MusicCaps [25]	1/2/3	T/A	5k
AF-AudioSet [109]	1/2/3	T/A	695k
VGGSound [110]	2/3	T/V/A	173k

Table 2: Objective results of VGGSound-Test regarding audio quality, semantic alignment, and temporal alignment. w. text denotes audio generation with text as a guiding condition, and w/o text denotes audio generation without text input, using only the video content.

Method	FAD↓	FD ↓	KL-sigmoid ↓	IS ↑	CLIP ↑	AV ↑
Diff-Foley [62]	6.05	23.38	7.86	10.95	9.40	0.21
FoleyCrafter w/o text [71]	2.38	26.70	6.24	9.66	15.57	<b>0.25</b>
FoleyCrafter w. text [71]	2.59	20.88	5.38	13.60	14.80	0.24
V2A-Mapper [111]	0.82	13.47	6.57	10.53	15.33	0.14
Frieren [63]	1.36	12.48	6.58	12.34	11.57	0.21
YingSound w/o text (ours)	0.80	8.66	5.12	12.08	16.14	<b>0.25</b>
YingSound w. text (ours)	<b>0.78</b>	<b>6.28</b>	<b>3.97</b>	<b>14.02</b>	<b>16.86</b>	<b>0.25</b>

with a total of 230k training steps. The training is carried out with a batch size of 128 on 8 Nvidia A800 GPUs. We use the Adam optimizer with a learning rate of 3e-5. We also clip the gradient norm to 0.2 for training stability. During inference, we use a sway sampling strategy [112] with NFE = 64 to improve our model’s performance.

**Evaluation Metrics.** For evaluation, we employ several evaluation metrics to assess semantic alignment, temporal alignment, and audio quality, namely Inception Score (IS) [113], CLIP score (CLIP), Fréchet Distance (FD) [114], Fréchet Audio Distance (FAD), AV-align (AV) and KL Divergence-sigmoid (KL-sigmoid) [115] by calculating all data in the VGGSound-test set. The CLIP measures the similarity between the input video and the generated audio embeddings within the same representation space. To achieve this, we utilize Wav2CLIP [116] as the audio encoder and CLIP as the video encoder, following the approach adopted in prior studies [33]. FD assesses the distribution similarity to evaluate the fidelity of the generated audio, using PANNs [117] as the feature extractor. FAD evaluates the fidelity of the generated audio by assessing the similarity of its distribution by using VGGish [118] as the feature extractor. KL-sigmoid computes the average KL-divergence across all classes in the test set to assess the similarity at the paired sample level. IS assesses sample quality and diversity between ground truth and generated audio. AV is based on the detection and comparison of energy peaks in both modalities to assess the temporal alignment of the input audio and generated video.

**Baselines.** We conduct comprehensive evaluations of YingSound by comparing it with state-of-the-art approaches, namely Diff-Foley, FoleyCrafter, V2A-Mapper [111], and Frieren. Diff-Foley adopts contrastive audio-visual pre-training (CAVP) [119] to learn more temporally and semantically aligned features, then trains an LDM with CAVP-aligned visual features in the spectrogram latent space. FoleyCrafter adopts a novel framework that leverages a pre-trained text-to-audio model to ensure high-quality audio generation. FoleyCrafter comprises two key components: the semantic adapter for semantic alignment and the temporal controller for precise audio-video synchronization. We evaluate both with and without (FoleyCrafter w. or w/o text) text as a condition. Like FoleyCrafter, our model also supports two types of sound effects generation results: one with text input, where text serves as the condition for generating audio based on video content, and video-based audio generation (YingSound w. or w/o text). V2A-Mapper, evaluated on only 7,590 WAV clips, introduces a module to translate visual inputs for high-fidelity sound generation, achieving superior performance with fewer parameters compared to existing approaches. Frieren uses rectified flow matching to synthesize high-quality, temporally aligned audio from video. It features a non-autoregressive vector

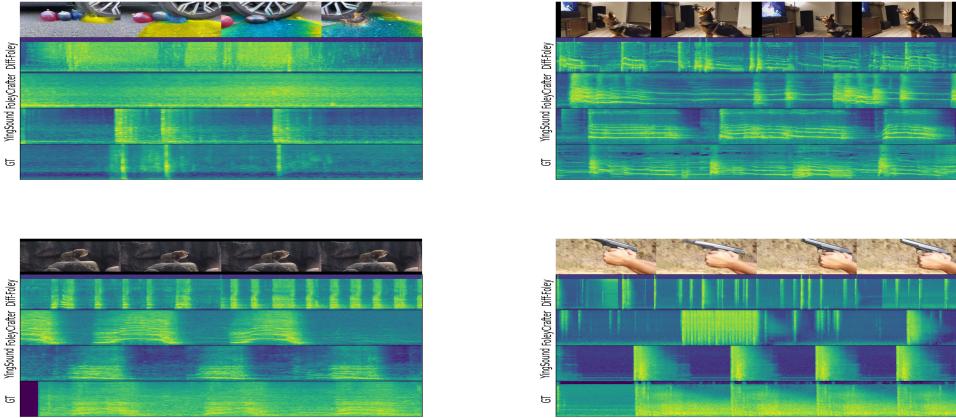


Figure 3: **Temporal Alignment** comparison.

field estimator and channel-level cross-modal feature fusion, enabling efficient audio generation with superior synchronization and quality.

## 4.2 Video-to-Audio Experiments

We present a quantitative comparison of semantic alignment and audio quality in the VGGSound-test dataset. The VGGSound-test dataset comprises 15,446 videos collected from YouTube, covering a diverse range of genres. The results show that the YingSound model, particularly when integrated with text input, exhibits superior performance across multiple objective metrics, including semantic alignment, temporal alignment with visual input, and better audio fidelity.

### 4.2.1 Audio Fidelity

As shown in Table 2, the YingSound w. text model achieves the lowest FAD of 0.78, indicating a high level of audio realism, and the lowest FD of 6.28, suggesting excellent audio quality. It also achieves the lowest KL-sigmoid with a score of 3.97, which indicates the overall high-quality of sound effects generated based on video. Furthermore, the YingSound w. text model stands out with an IS of 14.02, which is a testament to the diversity and quality of the audio content it generates.

### 4.2.2 Semantic Alignment

It is important to note that when it comes to the CLIP, which measures semantic alignment, the YingSound w. text model achieves the highest score of 16.86, indicating a very strong correlation between the generated audio and the corresponding video. This highlights that both models, despite differences in their input, perform exceptionally well in ensuring that the generated audio is semantically aligned with the accompanying text. The YingSound w. text model’s advantage in the CLIP suggests that it may have a slight edge in capturing the nuances of the text description for semantic alignment.

### 4.2.3 Temporal Alignment

The AV metric evaluates the synchronization between each segment of the input audio and its corresponding video segment. In Table 2, the YingSound w. text, YingSound w/o text, and FoleyCrafter w/o text models all achieve an impressive AV score of 0.25, tying for the highest score among the compared models. This demonstrates that YingSound models are capable of generating audio with excellent temporal alignment to the video, highlighting their strong performance in this critical aspect of cross-modal generation quality.

As shown in Figure 3, we compare ground truth (GT), Diff-Foley, FoleyCrafter, and our own YingSound models. It can be observed that the audio generated by YingSound exhibits nearly identical temporal alignment to the ground truth, demonstrating that our model achieves exceptional temporal alignment performance. However, Diff-Foley frequently generates either more or fewer sound effects than the ground truth. The results of the FoleyCrafter generation exhibit some misalignment in terms of timing.

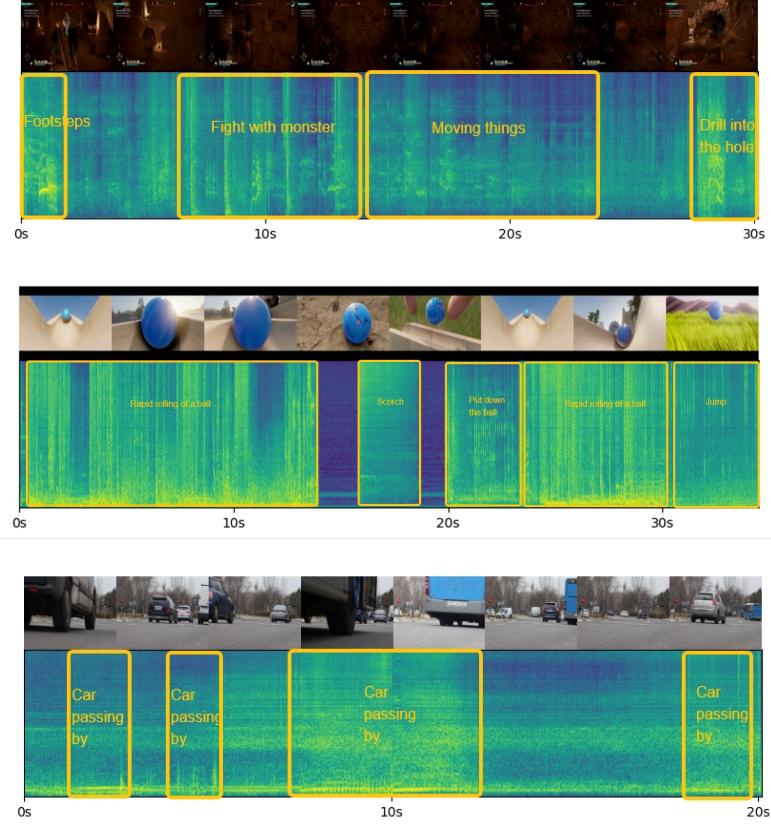


Figure 4: Application visualization results of YingSound.

## 5 Applications

In few-shot settings, it is promising to apply industrial-level applications for long-duration scenes, including synchronized sound effects generation for the motions of game characters, animation videos, and short-term videos. Below are three application samples that represent the corresponding initial industrial-level applications. Examples of long-duration video inference and additional cases can be listened to on our Project Page.

### 5.1 Sound Effects for Game Character Motions

In game-play, sound effects are essential for enhancing the realism and immersion of character movements, with the challenge lying in achieving precise synchronization between sound and actions at a finer level. Synchronizing audio with character actions provides players with a more dynamic and immersive gaming experience. The following samples showcase our demonstration of this promising application. We support the essential requirements for mid-duration sound effects generation for up to 30 seconds with coarse time alignment to character motions (Figure 4).

### 5.2 Sound Effects for Animation Videos

Sound effects are essential for enhancing generated videos by adding dynamic and contextually relevant audio that complements the visuals. The challenge lies in achieving semantic alignment between the generated audio and both the local and global dynamic visual content. Our model demonstrates high-quality sound effects generation for up to 30 seconds for a single subject (Figure 4).

### 5.3 Sound Effects for Short Videos

Sound effects are essential in enabling millions of short video creators and entrepreneurs to share their lives. A key factor is the adaptive production of sound effects that align with the varying rhythms of short videos, enhancing their impact and transmission potential towards zero-shot settings. YingSound demonstrates adaptive sound effects generation for up to 20 seconds to some extent (Figure 4).

## 6 Conclusion

In this paper, we introduce YingSound for video-guided sound effects generation. We introduce a novel V2A model approach, which is the first version of a foundation model and features a learnable AVA structure. This model dynamically integrates high-resolution visual features with the corresponding audio features across multiple stages. To generate high-quality sound effects, we propose a multi-modal visual-audio CoT module that produces high-quality audio in few-shot settings. The various metrics on the benchmark, along with the temporal alignment charts, demonstrate our model’s overall strong generation performance, high video semantic understanding, and exceptional temporal alignment capability. Additionally, we have built an industry-standard V2A dataset that encompasses a wide range of durations, including data from movies, games, and advertisements.

We are committed to continuously optimizing the existing V2A model, improving its performance across a wide range of application scenarios. Looking ahead, our team will focus on advancing its audio-related technical capabilities to provide an even more immersive and seamless user experience. One key area of development will be sound effects generation for game-play character motions, with a focus on sentimentally adaptive audio production including a variety of key elements: synchronization of action sequences, environment-specific adaptation, character-specific audio generation, layered audio design, and more.

## References

- [1] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.
- [2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.
- [3] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.
- [4] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [5] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR, 2018.
- [6] Younggun Lee and Taesu Kim. Robust and fine-grained prosody control of end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915. IEEE, 2019.
- [7] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.
- [8] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

- [9] Zhen Ye, Zeqian Ju, Haohe Liu, Xu Tan, Jianyi Chen, Yiwen Lu, Peiwen Sun, Jiahao Pan, Weizhen Bian, Shulin He, et al. Flashspeech: Efficient zero-shot speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6998–7007, 2024.
- [10] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023.
- [11] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.
- [12] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.
- [13] Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.
- [14] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- [15] Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, et al. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*, 2023.
- [16] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Audiochatllama: Towards general-purpose speech abilities for llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532, 2024.
- [17] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.
- [18] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023.
- [19] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [20] Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [21] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021.
- [22] Shangda Wu, Yue Yang, Zhaowen Wang, Xiaobing Li, and Maosong Sun. Generating chord progression from melody with flexible harmonic rhythm and controllable harmonic density. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):4, 2024.
- [23] Ang Lv, Xu Tan, Peiling Lu, Wei Ye, Shikun Zhang, Jiang Bian, and Rui Yan. Getmusic: Generating any music tracks with a unified representation and diffusion framework. *arXiv preprint arXiv:2305.10841*, 2023.
- [24] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

- [25] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.
- [26] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [27] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045, 2021.
- [28] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.
- [29] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. Efficient neural music generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [30] Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingshang Hong, and Yong Qin. Lafma: A latent flow matching model for text-to-audio generation. *arXiv preprint arXiv:2406.08203*, 2024.
- [31] Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. Audiolcm: Text-to-audio generation with latent consistency models. *arXiv preprint arXiv:2406.00356*, 2024.
- [32] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolum: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.
- [33] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [34] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021.
- [35] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.
- [36] 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.
- [37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
- [38] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.
- [39] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [40] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

- [41] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.
- [42] Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner, 2024.
- [43] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.
- [44] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.
- [45] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation, 2023.
- [46] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. M<sup>2</sup>ugen: Multi-modal music understanding and generation with the power of large language models, 2024.
- [47] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.
- [48] Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation. *arXiv preprint arXiv:2407.02869*, 2024.
- [49] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.
- [50] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12486, 2021.
- [51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.
- [52] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.
- [53] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022.
- [54] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024.
- [55] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronisation in the wild. *arXiv preprint arXiv:2112.04432*, 2021.
- [56] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

- [57] Marco Comunità, Riccardo F. Gramaccioni, Emilian Postolache, Emanuele Rodolà, Danilo Comminiello, and Joshua D. Reiss. Syncfusion: Multimodal onset-synchronized video-to-audio foley synthesis, 2023.
- [58] Manjie Xu, Chenxing Li, Xinyi Tu, Yong Ren, Rilin Chen, Yu Gu, Wei Liang, and Dong Yu. Video-to-audio generation with hidden alignment, 2024.
- [59] Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. TiVA: Time-aligned video-to-audio generation. In *ACM Multimedia 2024*, 2024.
- [60] Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, and Ruihua Song. Lova: Long-form video-to-audio generation, 2024.
- [61] Yong Ren, Chenxing Li, Manjie Xu, Wei Liang, Yu Gu, Rilin Chen, and Dong Yu. Sta-v2a: Video-to-audio generation with semantic and temporal alignment, 2024.
- [62] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023.
- [63] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*, 2024.
- [64] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. *arXiv preprint arXiv:2409.13689*, 2024.
- [65] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation, 2021.
- [66] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation, 2023.
- [67] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies, 2023.
- [68] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaocheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation, 2023.
- [69] Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. Sonicvisionlm: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875, 2024.
- [70] Santiago Pascual, Chunghsin Yeh, Ioannis Tsiamas, and Joan Serrà. Masked generative video-to-audio transformers with enhanced synchronicity. *arXiv preprint arXiv:2407.10387*, 2024.
- [71] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhenning Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024.
- [72] 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.
- [73] Qi Yang, Binjie Mao, Zili Wang, Xing Nie, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shimeng Xiang. Draw an audio: Leveraging multi-instruction for video-to-audio synthesis. *arXiv preprint arXiv:2409.06135*, 2024.
- [74] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners, 2024.
- [75] 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, 2023.
- [76] Mingjing Yi and Ming Li. Efficient video to audio mapper with visual scene detection, 2024.

- [77] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6639–6647, 2024.
- [78] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [79] Zhipu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.
- [80] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547*, 2021.
- [81] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- [82] Ling Yang, Zixiang Zhang, Zhilong Zhang, Xingchao Liu, Minkai Xu, Wentao Zhang, Chenlin Meng, Stefano Ermon, and Bin Cui. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024.
- [83] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
- [84] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.
- [85] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.
- [86] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.
- [87] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [88] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [89] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.
- [90] 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.
- [91] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.
- [92] Mikhail Burtsev and Anna Rumshisky. Multi-stream transformers. *arXiv preprint arXiv:2107.10342*, 2021.

- [93] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105, 2024.
- [94] Hongcheng Guo, Wei Zhang, Junhao Chen, Yaonan Gu, Jian Yang, Junjia Du, Binyuan Hui, Tianyu Liu, Jianxin Ma, Chang Zhou, et al. Iw-bench: Evaluating large multimodal models for converting image-to-web. *arXiv preprint arXiv:2409.18980*, 2024.
- [95] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*, 2024.
- [96] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.
- [97] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. *arXiv preprint arXiv:2310.12921*, 2023.
- [98] Yuwei Fu, Haichao Zhang, Di Wu, Wei Xu, and Benoit Boulet. Furl: Visual-language models as fuzzy rewards for reinforcement learning. *arXiv preprint arXiv:2406.00645*, 2024.
- [99] Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Murun Yang, Qiaozhi He, Tong Xiao, Chunliang Zhang, Tongran Liu, Quan Du, et al. Rovrm: A robust visual reward model optimized via auxiliary textual preference data. *arXiv preprint arXiv:2408.12109*, 2024.
- [100] Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. *arXiv preprint arXiv:2402.03681*, 2024.
- [101] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.
- [102] Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennherag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, et al. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*, 2023.
- [103] Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16426–16435, 2024.
- [104] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18798–18806, 2024.
- [105] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibeи Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.
- [106] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergrt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.
- [107] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

- [108] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [109] 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.
- [110] 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.
- [111] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15492–15501, 2024.
- [112] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.
- [113] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [114] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [115] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.
- [116] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.
- [117] 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.
- [118] Eunjeong Koh and Shlomo Dubnov. Comparison and analysis of deep audio embeddings for music emotion recognition. *arXiv preprint arXiv:2104.06517*, 2021.
- [119] Bill Jung, Yan Li, and Tamir Bechor. Cavp: A context-aware vulnerability prioritization model. *Computers & Security*, 116:102639, 2022.