AvSyncDiff: Video-guided Audio Generation via Enhanced Multimodal Feature

Anonymous ACL submission

Abstract

Video-driven audio synthesis aims to generate synchronized and contextually appropriate audio based on visual content, with applications in multimedia, virtual reality, and film production. Existing methods often rely solely on visual cues, leading to suboptimal audio generation that lacks synchronization and semantic alignment. To address these challenges, we introduce a novel video-guided audio synthesis method, termed *AvSyncDiff*. Unlike traditional approaches, AvSyncDiff leverages both visual and textual inputs, along with an optional audio prompt, to achieve precise control over the audio generation, enhancing the quality and realism of the synthesized audio. Furthermore, we propose a Gaussian Mixture Diffusion Search (GMDS) algorithm, a test-time scaling strategy inspired by advancements in the text-to-image domain. GMDS employs a dual-scale sampling mechanism to adaptively explore the latent space, balancing local exploitation and global exploration through a combination of small and large step sizes. The experimental results demonstrate that AvSyncDiff significantly outperforms state-of-the-art methods in both quantitative metrics and qualitative evaluations, showcasing its potential for diverse applications in multimedia and beyond.

1 Introduction

Recent advancements in cross-modal learning have led to significant progress in audio generation. State-of-the-art models like AudioGen (Kreuk et al., 2022), Make-An-Audio (Huang et al., 2023), and AudioLDM (Liu et al., 2023a) have demonstrated remarkable capabilities in synthesizing audio from text descriptions. As text-to-video generation (Khachatryan et al., 2023; Ge et al., 2023) continues to show promising applications, there is increasing interest in tackling the inverse problem: generating audio from video.

Video-to-audio generation presents significantly greater challenges than text-to-audio generation, primarily due to the following reasons: (1) Multimodal Alignment Requirements: The generated audio must synchronize with both the semantic content and temporal dynamics of the video. (2) Complex Audio-Visual Relationships: The relationship between visual and audio elements is inherently complex, as audio can encompass various styles and dynamic characteristics that evolve over time. Video content alone often provides insufficient cues to determine the desired audio style and emphasis without additional guidance.

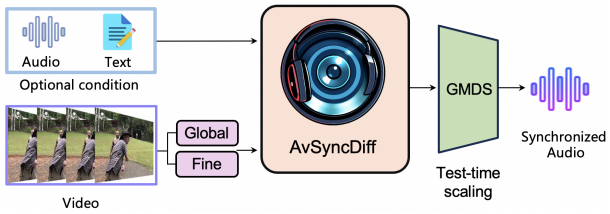

Figure 1: Our AvSyncDiff model supports multi-modal conditioning, incorporating not only video but also text and audio inputs.

Several methods have been proposed for video-to-audio generation. One approach uses diffusion-based architectures (Zhang et al., 2024b) where video frames are encoded using image encoders like CLIP (Radford et al., 2021) to condition audio generation. A distinct approach is Diff-Foley, which employs contrastive learning between video and audio to pre-train representations with audio-aware features before generation. Another approach is to use text as an intermediate modality to bridge the gap between video and audio (Wang et al., 2024). These methods first convert the video into textual descriptions and then leverage powerful text2audio models (Xie et al., 2024) to generate the corresponding audio. However, existing video-to-audio generation methods face three key challenges: First, high-quality video-audio paired datasets remain scarce compared to image-text data. While CLIP utilized 400 million image-text pairs

and CLAP incorporated data totaling over 5000 hours, video-audio datasets lack comparable scale and diversity, limiting model generalization in complex scenarios. Second, current approaches typically compress the input video into a single feature representation, resulting in substantial information loss. Videos contain rich temporal dynamics and contextual changes that single representations cannot adequately capture, leading to generated audio lacking precise temporal alignment with visual events. Finally, existing methods overlook the one-to-many nature of the video-to-audio relationship. For any video, multiple plausible audio outputs could authentically accompany the visual content. Current approaches implicitly assume a deterministic mapping between modalities, failing to capture the diverse space of possible audio interpretations. Incorporating reference audio as an additional conditioning signal would enable more precise navigation of this solution space while maintaining semantic coherence with the visual content.

To address these limitations, we propose a novel video-to-audio generation framework called *AvSyncDiff*, which leverages multiple high-level control signals, including text, video, and audio. As illustrated in Figure 1, the AvSyncDiff model is built upon a U-Net diffusion-based architecture and incorporates three additional inputs: text, video, and audio. Unlike conventional approaches that rely on a single feature to represent the video input, AvSyncDiff innovatively partitions the video into two distinct components: a video global feature, which encapsulates the overall context of the video, and fine-grained frame-wise image features, which are derived from individual frames. This dual-feature representation mitigates the risk of information loss across the entire video and facilitates precise timing alignment at the frame level. Furthermore, in line with prior research, AvSyncDiff can integrate text input as an additional control modality, for example, using “loudly" to control the volume or giving a detailed description to assist generation. More significantly, our model introduces the capability to accept prompt audio input. This novel feature allows for more nuanced and refined control over the audio generation process, enabling the model to produce audio that is more closely aligned with the desired output characteristics.

Besides the model architecture design, we also improve the performance at the inference stage with a test-time scaling strategy inspired by advancements in the text-to-image domain (Ma et al., 2025). Specifically, we propose a Gaussian Mixture Diffusion Search (GMDS) algorithm, which employs a dual-scale sampling mechanism to explore the latent space adaptively. This approach balances local exploitation and global exploration by combining small and large step sizes, ensuring more efficient and effective optimization during inference. We conduct experiments to demonstrate the superiority of our method, and our contributions can be summarized as follows:

- We identify that maintaining both global context and temporal dynamics is crucial for video-to-audio generation. Based on this insight, we propose a simple yet effective solution that uses a complementary representation of global video features and frame-wise image features, significantly improving audio-video synchronization quality.

- We introduce a multi-modal controlled framework that integrates text, video, and audio inputs, allowing for fine-grained control over the generated audio. Including prompt audio as a conditioning signal enables the model to navigate the diverse space of possible audio interpretations while maintaining semantic coherence with the visual content.

- We develop a test-time scaling method specifically designed for video-to-audio generation, which optimizes the model’s output quality during inference. This method ensures high fidelity and temporal alignment of the generated audio with the input video.

2 Related Work

2.1 Text-to-audio Generation

Text-to-audio generation has emerged as a prominent research direction in recent years. Contemporary approaches can be systematically categorized into two predominant frameworks: transformer-based autoregressive models and diffusion-based architectures.

AudioGen (Kreuk et al., 2022) pioneered an autoregressive framework for audio synthesis, capable of generating acoustic content conditioned on either textual descriptions or audio prompts. Expanding this paradigm, MusicGen (Copet et al., 2024) implements a sophisticated language modeling approach utilizing efficient token interleaving

strategies to enable high-fidelity music generation from textual inputs.

Within the diffusion-based domain, AudioLDM (Liu et al., 2023a) advances the field by implementing contrastive language-audio pretraining (CLAP) embeddings as latent conditional variables within a VAE framework for audio synthesis. Make-an-Audio (Huang et al., 2023) introduces an innovative approach through a spectrogram autoencoder that predicts self-supervised representations instead of direct waveforms, thereby facilitating enhanced compression efficiency and deeper semantic interpretation. Through the integration of CLAP embeddings (Wu et al., 2023) with high-fidelity diffusion architectures, Make-an-Audio achieves sophisticated language comprehension capabilities alongside superior audio generation quality.

2.2 Video-to-audio Generation

Existing methods in video-to-audio generation aim to learn joint representations of visual and auditory modalities to produce audio synchronized with visual content. SpecVQGAN (Iashin and Rahtu, 2021) presents an efficient framework for multiclass, visually guided sound synthesis using a transformer decoder trained to sample from a codebook-based prior. Diff-foley (Luo et al., 2024) employs a latent diffusion model (LDM) for audio synthesis, utilizing contrastive audio-visual pre-trained (CAVP) features derived from a large-scale video dataset. V2A-Mapper (Wang et al., 2024) leverages pre-trained foundation models to bridge multimodal gaps, enabling the transfer of pre-trained knowledge to better handle open-domain challenges. FoleyCrafter (Zhang et al., 2024b) uses an IP-adapter (Ye et al., 2023) connection to capture video semantics while employing an audio event detection model to identify temporal information.

3 Method

3.1 Text-Video Condition

Text condition We build upon Tango-2 (Majumder et al., 2024), a strong pre-trained text-to-audio diffusion model, as our backbone architecture. For the text input processing, we follow the original pipeline of Tango-2. Specifically, we employ the pre-trained FLAN-T5 text encoder (Chung et al., 2024) to extract text embeddings, which are kept frozen during training.

Given that the original captions in the VGGSound dataset are relatively simple and limited in semantic richness, we enhance them using Qwen2-Audio (Chu et al., 2023). As illustrated in Figure 2(c), we designed a system prompt and fed both the audio and its corresponding original caption to Qwen2-Audio, obtaining a more descriptive and semantically enriched version of the prompt.

Then, we further introduce diversity into the prompts by generating perturbed versions using the DeepSeek-V3 language model. For each rewritten prompt, we generate three variations. To ensure semantic consistency, we include specific instructions in the input prompt to guarantee that the perturbed texts (c') remain conceptually or semantically close to the original (c). This approach not only enriches the textual content but also effectively expands the training data, leading to improved generalization and robustness of the model.

Video condition Accurate video representation is essential for capturing the semantic and temporal content of dynamic scenes. While previous methods like FoleyCrafter (Zhang et al., 2024b) rely on CLIP-based image encoders to extract averaged frame features, such approaches lack explicit modeling of motion and temporal structure due to CLIP's static-image pre-training.

To extract global video features, we utilize the video encoder from the contrastive video-language pre-training framework InternVid (Wang et al., 2023). InternVid is built upon CLIP, and it undergoes continued pre-training of CLIP encoder on a large-scale dataset of video-text pairs, which enables it to better understand temporal relationships and motion dynamics across video frames. For the video sequence $\{v_1, v_2, ..., v_N\}$, we use the visual encoder of InternVid to obtain a global feature vector $f_g \in R^d$, where d is the feature dimension. Then, f_g is passed through a projection layer followed by a LayerNorm to match the dimensionality required by our diffusion model. As shown in Figure 2(a), to effectively incorporate this video information into the diffusion process, we adopt a decoupled cross-attention mechanism inspired by the IP-Adapter(Ye et al., 2023). Specifically, given the query features Z from the U-Net, text features f_t, and global video features f_g, the decoupled cross-attention is formulated as:

$$Z_{sem} = Softmax(\frac{QK^\top}{\sqrt{d'}})V + Softmax(\frac{Q(K')^\top}{\sqrt{d'}})V' \quad (1)$$

where d' is the dimensionality of the diffusion model's latent space.,$Q = ZW_q$ is the shared query

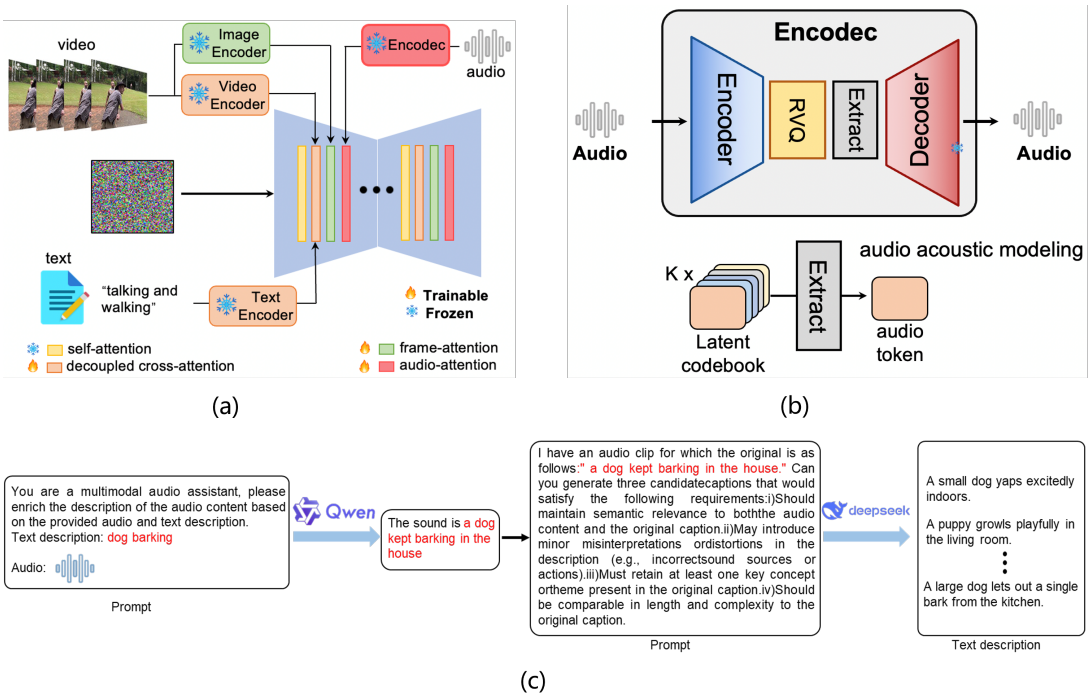

Figure 2: Figure (a) illustrates the architecture of our AvSyncDiff model. Figure (b) depicts the audio features extracted by the Encodec model. Figure(c) shows the pipeline of our text generation.

transformation, $K = f_t W_k, V = f_t W_v$ are the key and value projections for text features, and $K' = f_g W'_k, V' = f_g W'_v$ are the key and value projections for global video features.

Image frame conditon While the global video features capture high-level semantic content and temporal dynamics, fine-grained frame-level details are crucial for maintaining precise temporal alignment between the generated audio and visual events. Therefore, we propose a complementary frame-level processing stream that operates parallel to global feature extraction.

As illustrated in Figure 2(a), we uniformly sample N frames $\{v_1, v_2, \ldots, v_N\}$ and process each frame through the CLIP visual encoder to obtain frame-wise embeddings $e_i \in R^d$. To effectively integrate these frame-level features into the diffusion model, we first transform them through a projection layer with RMSNorm(Zhang and Sennrich, 2019) to match the required dimensionality. These adapted frame-wise embeddings $H = \{h_1, h_2, \ldots, h_N\}$ are then integrated into the diffusion process through a dedicated cross-attention layer called frame-attention in the U-Net architecture. Importantly, we incorporate Rotary Position Embeddings (RoPE)(Su et al., 2024) on key and query vector to the frame-wise cross-attention layer (frame-attention) to encode the temporal ordering of frames, enabling the model to better understand and utilize sequential relationships. Formally, we have

$$Z_{fine} = Softmax\left(\frac{RoPE(Q)RoPE(K_f)^\top}{\sqrt{d'}}\right)V_f \quad (2)$$

where $Q = ZW_q$ is the query transformation, $K_f = HW_{k_f}, V_f = HW_{v_f}$ are the key and value projections for fine-grained visual features, and W_q, W_{k_f}, W_{v_f} are learnable projection matrices.

3.2 Prompt Audio condition

To guide the audio generation process, we use a pre-trained EnCodec model (Défossez et al., 2022) to extract acoustic-aware embeddings from the prompt audio. These embeddings capture the essential acoustic characteristics of the input audio, providing a strong conditioning signal for the generation process.

As illustrated in Figure 2(b), EnCodec employs a Residual Vector Quantization (RVQ) mechanism to quantize the output of the encoder. To integrate this embedding into the diffusion model, we add a cross-attention layer called audio-attention in the U-Net architecture. The audio-attention is formulated as:

$$Z_{audio} = Softmax\left(\frac{QK_a^\top}{\sqrt{d'}}\right)V_a \quad (3)$$

where $Q = ZW_q$ is the query transformation, $K_a = \mathbf{a}W_{k_a}, V_a = \mathbf{a}W_{v_a}$ are the key and value projections for the audio embedding, and W_q, W_{k_a}, W_{v_a} are learnable projection matrices.

By incorporating the prompt audio embedding, our model can generate audio that aligns with both the visual content and the acoustic characteristics of the prompt, enhancing the overall quality and realism of the output.

3.3 Test-time Scaling

To enhance the inference performance of our video-to-audio diffusion model, we propose a Gaussian Mixture Diffusion Search (GMDS) strategy. This approach introduces a dual-scale sampling mechanism that adaptively explores the latent space by combining both small and large step sizes, thereby balancing local exploitation and global exploration.

The GMDS algorithm maintains a population of candidate solutions that evolve over multiple iterations. Initially, the population P_0 is sampled from a standard normal distribution. At each iteration, for every candidate in the population, we generate two potential updates using different diffusion scales: a conservative step with parameter β_s for local exploitation and a more aggressive step with parameter $\beta_l$for global exploration ($\beta_l > \beta_s$). Specifically, given a candidate solution, we generate two new candidates:

$$x_s = \beta_s x + \sqrt{1-\beta_s^2} \cdot \eta \tag{4}$$

$$x_l = \beta_l x + \sqrt{1-\beta_l^2} \cdot \eta \tag{5}$$

where η is sampled from $\mathcal{N}(0,1)$.

To evaluate the quality of the generated audio, we employ a zero-shot evaluation approach. Specifically, we utilize the CLIP score (Wu et al., 2022) to measure the audio-visual correspondence and the CLAP score to assess the audio-text alignment. These metrics provide a comprehensive evaluation of the generated audio in terms of both visual and textual relevance. We show the algorithm in the 1

3.4 Training Method

During the training process, we keep all feature extractors frozen, including the text encoder, CLIP image encoder, InternVid video encoder, and Encodec. We only train the projection layers, which are designed to align the dimensions of different modalities, as well as the additional cross-attention mechanisms integrated into each U-Net diffusion block. The training objective is based on the standard diffusion loss, which ensures that the model learns to generate high-quality audio by progressively denoising the latent representations.

Algorithm 1 Gaussian Mixture Diffusion Search

Require: $D, T, N, \beta_s, \beta_l, \mathcal{F}$
Ensure: **best_solution**, **best_score**
Initialize population $P_0 \sim \mathcal{N}(0,1)^D$ with N samples
Find initial $x^* \leftarrow \arg\max_{x \in P_0} \mathcal{F}(x)$
best_score $\leftarrow \mathcal{F}(x^*)$, **best_solution** $\leftarrow x^*$
for $t \leftarrow 1$ to T **do**
  for $i \leftarrow 1$ to N **do**
    Sample noise vector $\eta \sim \mathcal{N}(0,1)^D$
    Generate candidates: $\mathrm{x}_s \leftarrow \beta_s P_{t-1}[i] + \sqrt{1-\beta_s^2} \cdot \eta$
    $\mathrm{x}_l \leftarrow \beta_l P_{t-1}[i] + \sqrt{1-\beta_l^2} \cdot \eta$
    Select $x \leftarrow \arg\max\{\mathcal{F}(x_s), \mathcal{F}(x_l)\}$
    Update population: $P_t[i] \leftarrow x$
    if $\mathcal{F}(x) >$ **best_score then**
      best_score $\leftarrow \mathcal{F}(x)$
      best_solution $\leftarrow x$
    end if
  end for
end for=0

4 Experiments

4.1 Experiment Setup

Dataset To train our proposed AvSyncdiff, we utilize VGGSound (Chen et al., 2020) dataset, adhering to its original train/test splits. For evaluation, we assess our method using both the VGGSound and AVSync15 (Zhang et al., 2024a) datasets.

Implementation Details Given the scarcity of large-scale audio-visual paired data, we build our video-to-audio generation framework upon a strong text-to-audio diffusion model, Tango-2 (Majumder et al., 2024). Inspired by vision-language models such as LLaVA (Liu et al., 2023b), we adapt the pre-trained language model through modality projection and alignment. We use OpenCLIP ViT-H/14 (Cherti et al., 2023) for image encoding and ViCLIP-L-14 (Wang et al., 2023) for video encoding. For audio encoding, we utilize the Encodec 24 kHz model with a bandwidth of 6 kbps. The training process uses the AdamW optimizer with a constant learning rate of 1×10^{-4}, a batch size of 128 visual-audio embedding pairs, and a dropout

rate of 0.1 for classifier-free guidance.

Metrics To evaluate the performance, we employ a comprehensive set of metrics that assess various aspects of the generated audio, including: Mean KL Divergence (MKL)(Iashin and Rahtu, 2021) to measure the sample-level similarity between the generated audio and the ground truth; CLIP Similarity to evaluate the semantic coherence between the input video and the generated audio embeddings, using Wav2CLIP(Wu et al., 2022) as the audio encoder and CLIP as the video encoder, as done in previous works (Wang et al., 2024; Zhang et al., 2024b); Frechet Distance (FD)(Heusel et al., 2017) and Frechet Audio Distance (FAD)(Kilgour et al., 2018) with VGGish(Hershey et al., 2017) to assess the fidelity and distribution similarity of the generated audio; CLAP Similarity to evaluate the cross-modal alignment between text and generated audio; and Onset Acc (onset detection accuracy) and Onset AP (onset detection average precision) (Xie et al., 2024) to evaluate the generated audios, using the onset ground truth from the datasets.

Baseline For comparison, we use current state-of-the-art method SpecVQGAN (Iashin and Rahtu, 2021), Diff-Foley (Luo et al., 2024), V2A-Mapper (Wang et al., 2024), Seeing-and-hearing (Xing et al., 2024) and Foley-crafter (Zhang et al., 2024b). SpecVQGAN generates audio tokens autoregressively from video tokens, while Diff-Foley applies contrastive learning to achieve synchronized audio synthesis via its CAVP encoders. V2A-Mapper aligns image representations with audio embeddings in CLAP space, facilitating video-based audio generation through a pre-trained model. Seeing-and-hearing uses ImageBind (Girdhar et al., 2023) as a connector between visual and audio domains. FoleyCrafter introduces both semantic and temporal adapters to enhance video2audio generation.

4.2 Qualitative Evaluation

Audio quality and cross-modal alignment We evaluate the proposed AvSyncDiff method on both the V2A (Video-to-Audio) and TV2A (Text-Video-to-Audio) tasks, as shown in Table 1. To ensure a fair comparison, we do not include results using audio prompts, as most baseline methods do not leverage such information. For GMDS, we set the number of candidates to be 6 in our test.

Our experimental results demonstrate that AvSyncDiff achieves strong performance on all evaluation metrics. When using video embeddings from a pre-trained video encoder and fine-grained frame-level features, our model generates more realistic and temporally coherent audio, especially under the challenging V2A and TV2A settings.

Compared to existing methods, AvSyncDiff consistently outperforms them in terms of both audio quality (MKL and FD) and cross-modal alignment (CLIP and CLAP), indicating better semantic and temporal synchronization with input video.

We further evaluate the effectiveness of our Gaussian Mixture Diffusion Search (GMDS) strategy by comparing results with and without it during inference. As shown in Table 1, applying GMDS significantly improves generation quality, highlighting its importance in enhancing the diffusion sampling process.

Time alignment To evaluate temporal synchronization, we conduct experiments on the AVSync15 dataset. This dataset is constructed from high-quality videos sourced from VGGSound, which have been carefully filtered and segmented to retain only the most precise video-audio pairs while discarding irrelevant or inactive segments. As shown in Table 2, our method achieves state-of-the-art performance in temporal synchronization, demonstrating its effectiveness in aligning audio with visual.

4.3 Quantitative Evaluation

Visualization Results We demonstrate the effectiveness of AvSyncDiff by comparing it with the state-of-the-art method FoleyCrafter through qualitative results. For both methods, we set the text prompt to None to simulate the scenario where no additional text guidance is provided.

Figure 3 presents audio-visual results from the AVSync-15 dataset. We extract frames at regular intervals and generate corresponding mel-spectrograms that span the full duration between adjacent frames. Unlike other methods that only show partial spectrograms, we visualize the complete time-frequency structure for a more comprehensive comparison.

In the gunshot video, FoleyCrafter generates gunshots that are intermittent and misaligned with the single firing event in the video. Additionally, the low- and high-frequency components are separated, leading to an unnatural sound. Our method produces sharper and temporally aligned gunshots, concentrated in the mid and high-frequency ranges, resulting in a more realistic auditory match.

Method	Task	MKL ↓	CLIP ↑	FD ↓	FAD ↓	CLAP ↑
V2A Results						
SpecVQGAN	V2A	3.40	5.876	32.01	5.79	-
Diff-Foley	V2A	3.32	9.172	29.03	6.23	20.5
V2A-Mapper	V2A	2.85	9.720	24.16	**1.34**	24.5
FoleyCrafter	V2A	2.56	10.70	19.67	2.78	25.3
AvSyncDiff (Ours)	V2A	<u>2.38</u>	**11.52**	**11.24**	2.14	<u>26.8</u>
AvSyncDiff (GMDS)	V2A	**2.33**	**12.98**	<u>10.13</u>	<u>1.95</u>	**30.3**
TV2A Results						
FoleyCrafter	TV2A	2.28	**14.80**	19.16	2.59	26.0
AvSyncDiff (Ours)	TV2A	<u>1.96</u>	<u>12.71</u>	<u>10.24</u>	2.17	26.0
AvSyncDiff (GMDS)	TV2A	**1.89**	12.54	**9.85**	**2.07**	**31.20**

Table 1: Comparison of results on VGGSound dataset. The results show that our method can achieve better results with less data. The number underlined indicates the second best result.

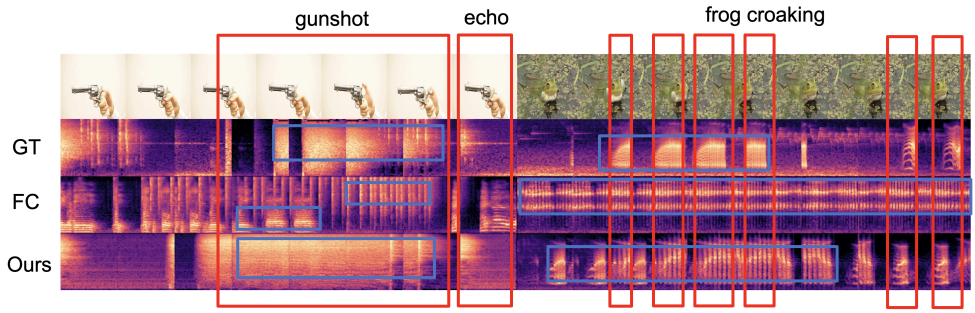

Figure 3: Visualiation results on Avsync15 dataset.

Method	Onset ACC ↑	Onset AP ↑
SpecVQGAN	26.74	63.18
Diff-Foley	21.18	66.55
Seeing and Hearing	20.95	60.33
FoleyCrafter	28.48	68.14
AvSyncDiff (Ours)	**30.72**	**69.28**

Table 2: Comparison of results on AVSync15 dataset.

In the frog croaking video, FoleyCrafter generates two constant-frequency sounds throughout the clip, failing to capture the natural frequency shift seen in the original spectrum. In contrast, AvSyncDiff successfully reproduces the dynamic progression from low to high frequency, while maintaining accurate temporal alignment with the visual content.

4.4 Audio Transfer

To evaluate the effectiveness of our audio condition block, we conducted a human evaluation focusing on two aspects: (i) **Overall Quality (OVL)** (Liu et al., 2023a), and (ii) **Relevance to the Input Audio (REL)** (Liu et al., 2023a).

For the evaluation, we selected two samples per category from the AVSync15 test set. For each sample, we collected three distinct 10-second audio clips from the internet as input prompts and randomly selected one during generation. The text inputs were also randomly sampled from multiple GPT-4-generated variations to enhance diversity and generalization.

Evaluators rated each generated audio clip on a scale from 1 to 5 based on the two criteria. We then computed average scores for each dimension and used them to compare model performance. We tested configurations using the first 1, 2, 4, and 8 layers of the Encodec codebook. Results are summarized in Table 3.

In terms of relevance to the input audio (REL), the Encodec-4 configuration achieved the highest score of 3.73, indicating superior alignment with the input prompts in both style and content. When considering the overall average (AVG) across both metrics, the Encodec-1 configuration performed

Method	Encodec-1	Encodec-2	Encodec-4	Encodec-8
OVL↑	**3.83**	3.60	3.47	3.65
REL↑	3.53	3.67	**3.73**	3.65
AVG	**3.68**	3.64	3.60	3.65

Table 3: Comparison of results. We use user study scores to show the performance.

Method	MKL ↓	CLIP ↑	FID ↓
AvSyncDiff(CLIP)	2.540	11.83	41.53
AvSyncDiff (Ours)	**1.732**	**12.58**	**33.46**

Table 4: We evaluated the performance using both CLIP average features and InternVid features for the given video. The InternVid features can produce better performance.

best with an average score of 3.68, making it the most balanced and effective.

4.5 Ablation Study

Effect of video condition We replace the InternVid video encoder with the CLIP image encoder and use the same average pooling method as FoleyCrafter. We fine-tune the projection layer and our trained frame-attention on the VGGSound dataset and report the results on the AVSync-15 dataset in Table 4. Compared to average pooling image embeddings, video embeddings from InternVid yield a noticeable performance improvement. Specifically, an increase of 0.75 in CLIP score is observed when using Internvid. For MKL and FID scores, we achieve $+0.808$ and $+8.07$, respectively.

Effect of image frame condition To investigate the effectiveness of our image frame condition, we conduct an ablation study by nullifying its contribution through zero-vector initialization. Due to the residual architecture of the diffusion model, this modification effectively eliminates the frame-attention layer while maintaining network stability. We evaluate this variant on the AVSync-15 dataset, with results presented in Table 5.

When the frame-attention mechanism is neutralized, we observe substantial degradation across all metrics. Specifically, our frame-attention achieves a 1.8 improvement in MKL score, indicating a significant enhancement in audio-visual synchronization quality. The CLIP score demonstrates a notable improvement of 2.84 points, suggesting stronger semantic coherence between the visual and audio modalities. Furthermore, we achieve an

Method	MKL ↓	CLIP ↑	FD ↓
AvSyncDiff(Zero)	3.532	9.74	45.37
AvSyncDiff (Ours)	**1.732**	**12.58**	**33.46**

Table 5: Comparison of results. We use zero feature vectors and the fine-grained visual features.

improvement of 11.91 points in FD metric, reflecting enhanced audio generation quality and naturalness.

5 Conclusion

In this paper, we present AvSyncDiff, a novel diffusion-based framework for video-to-audio synthesis that generates realistic audio content synchronized with visual input. Our approach uniquely integrates three modalities—video, audio, and text—into a unified architecture to ensure both semantic alignment and acoustic consistency in the generated output. Our framework leverages the InternVid video encoder with a decoupled cross-attention mechanism to extract comprehensive global video features, achieving semantic alignment between visual and audio content. The direct incorporation of frame-level conditions into the U-Net architecture enables precise temporal alignment without requiring additional training data. Furthermore, the introduction of prompt audio conditions provides fine-grained control over audio characteristics. Additionally, we introduce GMDS(Gaussian Mixture Diffusion Search) a novel inference-time optimization algorithm that enhances generation quality through test-time scaling. Through comprehensive experimental evaluation, we demonstrate that AvSyncDiff consistently outperforms existing approaches in generating high-quality, temporally synchronized audio content that aligns with input videos.

6 Limitation

While AvSyncDiff demonstrates significant advancements in video-to-audio synthesis, there are several areas for future improvement. One potential avenue for enhancing performance is the incorporation of higher-quality and larger audio-visual datasets. The largest available dataset, VGGSound, contains fewer than 200,000 samples, with each video limited to just 10 seconds in length. This constraint restricts our model's ability to generalize to longer videos and capture temporal dynamics.

References

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. 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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

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

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and 1 others. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR.

Vladimir Iashin and Esa Rahtu. 2021. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2024. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and 1 others. 2025. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International*

conference on machine learning, pages 8748–8763. PMLR.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. 2024. 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.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, and 1 others. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. 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.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. 2024. Sonicvisionlm: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. 2024. 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*, pages 7151–7161.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. 2024a. Audio-synchronized visual animation. *arXiv preprint arXiv:2403.05659*.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. 2024b. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*.