High-Quality Visually-Guided Sound Separation from Diverse Categories

Chao Huang[1], Susan Liang[1], Yapeng Tian[1],
Anurag Kumar[2], and Chenliang Xu[1]

[1] University of Rochester, Rochester NY 14627, USA
[2] Meta Reality Labs Research, Redmond WA 98052, USA

Abstract. We propose DAVIS, a **D**iffusion-based **A**udio-**VI**sual **S**eparation framework that solves the audio-visual sound source separation task through generative learning. Existing methods typically frame sound separation as a mask-based regression problem, achieving significant progress. However, they face limitations in capturing the complex data distribution required for high-quality separation of sounds from diverse categories. In contrast, DAVIS leverages a generative diffusion model and a Separation U-Net to synthesize separated sounds directly from Gaussian noise, conditioned on both the audio mixture and the visual information. With its generative objective, DAVIS is better suited to achieving the goal of high-quality sound separation across diverse sound categories. We compare DAVIS to existing state-of-the-art discriminative audio-visual separation methods on the AVE and MUSIC datasets, and results show that DAVIS outperforms other methods in separation quality, demonstrating the advantages of our framework for tackling the audio-visual source separation task. Our project page is available here: https://wikichao.github.io/data/projects/DAVIS/.

1 Introduction

Visually-guided sound source separation, also referred to as audio-visual separation, is a pivotal task for assessing a machine perception system's ability to understand multisensory signals [**?**, **?**]. It aims to separate individual sounds from a complex audio mixture by utilizing visual cues about the objects that are producing the sounds, *e.g.*, separate the "barking" sound from the mixture by querying the "dog" object. An effective separation model should be capable of handling a ***diverse*** range of sounds and producing ***high-quality*** separations that can deliver a realistic auditory experience. The community has devoted considerable effort to tackling this task [**?**,**?**,**?**,**?**,**?**,**?**,**?**,**?**], developing more powerful separation frameworks [**?**, **?**, **?**, **?**], proposing more effective training pipelines [**?**], and incorporating additional visual cues [**?**] to enhance the performance. Conventional approaches usually employ discriminative learning through mask regression [**?**] or spectrogram reconstruction [**?**] as training objectives.

While these methods have shown promising separation performance, they are inherently limited in dealing with diverse time-frequency structures and separating sounds in challenging situations. For instance, different sounds can interact

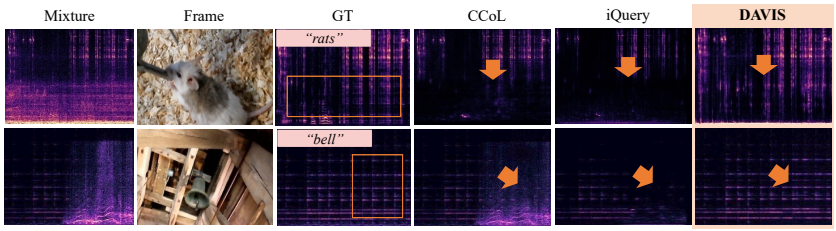


Fig. 1: Separation results on diverse time-frequency structures are shown for SOTA discriminative methods and our proposed DAVIS. Each row displays the audio mixture, reference visual frame, ground truth magnitude, and predicted magnitudes from DAVIS, iQuery [**?**], and CCoL [**?**]. DAVIS successfully recovers suppressed time-frequency structures (highlighted in the box), where mask-regression methods fail.

in complicated ways, and sometimes the desired sound is heavily suppressed by the others (*e.g.*, examples shown in Fig. 1), posing significant hurdles in regressing the mask from hidden sound patterns. Therefore, a natural question arises: *is there an effective approach to model complex data distributions, capture precise audio-visual associations, and generate high-quality separated sounds?*

We answer the question by introducing a generative framework for audio-visual separation. A new class of generative models called denoising diffusion probabilistic models (DDPMs) [**?**, **?**, **?**], also known as diffusion models, has emerged recently and demonstrated remarkable abilities in generating diverse and high-quality images [**?**] and audio [**?**]. Nonetheless, at first glance, whether diffusion models can be effectively repurposed for audio-visual separation remains unclear. The challenges of this task stem from the necessity of specialized architecture to capture audio-visual correspondences coupled with modeling the unique characteristics of magnitude spectrograms. Generic diffusion models may not be well-suited to this task without addressing the above challenges.

In this paper, we present DAVIS, a novel diffusion model-based audio-visual separation framework. Unlike conventional methods that regress masks, DAVIS tackles separation as a conditional generation process, iteratively "growing" the magnitude spectrogram from Gaussian noise to the desired sound. This approach distributes the burden of recovering complex time-frequency patterns across multiple steps, making DAVIS versatile for serving various scenarios, even the challenging cases shown in Fig. 1. A key component to train the diffusion model for audio-visual separation is the network architecture. Given the complexities of spectrograms and audio-visual correlations, we propose a Separation U-Net, well-suited for capturing these features. In particular, we find that modeling long-range dependencies is important as similar but distant time frames commonly exist. We, therefore, introduce a Convolution-Attention (CA) block in the Separation U-Net to capture both local and non-local contexts. Furthermore, to enhance the audio-visual association learning, we explore different interaction manners and devise a Feature Interaction module to facilitate the injection of visual cues into the separation task.

Another challenge arises from the frequent occurrence of silent time frames in magnitude spectrograms, where the values are almost zero. This skewed data distribution renders the conventional $\mathcal{L}_2$ loss in diffusion models susceptible to error. However, these silent parts also provide valuable information during inference, indicating the overlap between the mixed and target sounds. Therefore, they can be effectively utilized in the sampling process. To address these issues, we propose using a more robust $\mathcal{L}_1$ loss for training and a silence mask-guided sampling strategy to refine the results at the inference stage.

Experiments on the AVE [**?**] and MUSIC [**?**] datasets demonstrate that DAVIS consistently outperforms the state-of-the-art methods in terms of separation quality. Our contributions are summarized as follows:

- We approach the audio-visual separation task as a conditional generation process with generative diffusion models.
- We highlight the importance of network design and propose a Separation U-Net, equipping with an Audio-Visual Feature Interaction Module to capture multimodal association effectively.
- We identify an issue in magnitude spectrograms that can be leveraged to enhance the inference process and propose a novel silence mask-guided sampling strategy as a solution.
- Our framework is competitive with or surpasses previous methods on datasets with diverse categories, validating the effectiveness of our approach.

2 Related Work

Audio-Visual Sound Source Separation. In this section, our focus is on modern audio-visual sound source separation approaches while acknowledging the prolonged research efforts dedicated to signal processing-based separation [**?**, **?**] and other multimodal research [**?**, **?**, **?**]. Recent deep learning-based audio-visual sound source separation methods have been applied to a variety of audio categories, including speech signals [**?**, **?**, **?**, **?**], musical instrument sounds [**?**, **?**, **?**, **?**, **?**, **?**, **?**], and universal sound sources [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. These methods typically employ a learning regime that involves mixing two audio streams from different videos to provide supervised training signals. A sound separation network, often implemented as a U-Net, is then used for mask regression [**?**] conditioned on the associated visual features. In recent years, research in this area has focused on both domain-specific and open-domain sound source separation [**?**, **?**, **?**, **?**, **?**]. However, existing methods often require additional information, such as text queries [**?**], motion cues [**?**, **?**], or class labels [**?**], to achieve satisfactory performance. In this paper, we propose a novel generative audio-visual separation approach that demonstrates competitive or outperforms existing methods in separating both specific and open-domain sound sources.

Diffusion Models. Diffusion models [**?**, **?**, **?**] fall under the category of deep generative models that start with a sample in a random distribution and gradually restore the data sample through a denoising process. Recently, diffusion models

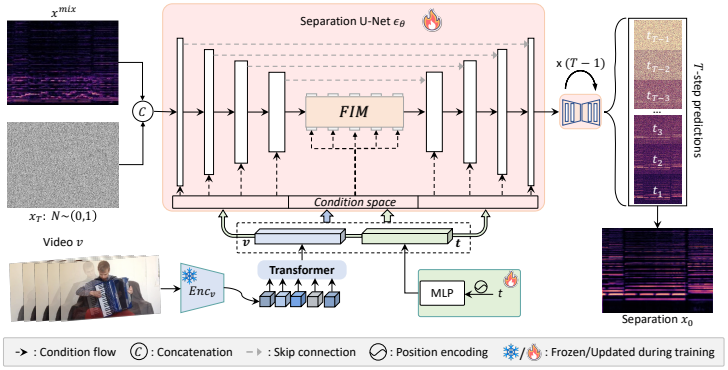


Fig. 2: Overview of the DAVIS framework. We aim to synthesize x_0 from the mixture x^{mix}, visual stream v, and timestep t. Starting with x_T from a standard distribution, we encode $v = \{I_j\}_{j=1}^{K}$ and t into the embedding space. A temporal transformer generates the visual feature $\boldsymbol{v}$, which, along with $\boldsymbol{t}$, conditions the Separation U-Net $\epsilon\theta$ to iteratively denoise x_T into x_0. $\boldsymbol{v}$ is used only in the Feature Interaction Module for audio-visual association, while $\boldsymbol{t}$ is used throughout.

have exhibited remarkable performance across various domains, including computer vision [?,?,?,?,?,?,?,?,?,?], natural language processing [?,?,?,?], audio applications [?,?,?,?,?,?,?], and audio-visual content generation [?]. Furthermore, there has been a growing interest in utilizing diffusion models for discriminative tasks. Some pioneer works have explored the application of diffusion models to image segmentation [?,?,?] and object detection [?]. Despite the significant interest in this direction, successful applications of generative diffusion models to audio-visual scene understanding remain limited. A few recent works have attempted to use diffusion-based approaches for audio-visual speech separation and enhancement [?,?]. However, these works limit themselves to speech separation or enhancement and do not study the more challenging audio-visual sound separation problem. The major blockers are the inherent challenges in designing effective network architectures to capture audio-visual correspondence and adapting diffusion models to handle unique data distributions. In this paper, we address this gap by employing a diffusion model for audio-visual sound separation. Our novel separation architecture empowers the diffusion-based model to effectively learn the intricate relationships between audio and visual modalities, leading to superior separation performance.

3 Method

In this section, we introduce DAVIS, our novel diffusion model-based audio-visual separation framework. We begin by providing a brief recap of diffusion models in Sec. 3.1. Then, we describe our task setup and give a method overview in Sec. 3.2. Next, we present our proposed separation framework in Sec. 3.3. Furthermore,

we discuss the training pipeline in Sec. 3.4. Finally, we introduce the silence mask-guided sampling strategy in Sec. 3.5.

3.1 Preliminaries: Diffusion Models

Diffusion models [**?**] typically consist of a forward and a reverse process. The forward process is defined as a Markov chain that gradually adds noise to the data sample x_0 according to a variance schedule $\beta_1, ..., \beta_T$. To sample x_t at an arbitrary timestep t, we have:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. Note that the variance schedule is fixed for both forward and reverse processes. If the total number of T goes to infinity, the diffusion process will finally lead to pure noise, *i.e.*, the distribution of $p(x_T)$ is $\mathcal{N}(x_t; \mathbf{0}, \mathbf{I})$ with only Gaussian noise.

The reverse process aims to recover samples from Gaussian distribution by removing the noise gradually, which is a Markov chain parameterized by θ:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{2}$$

where at each iteration, the transition is formulated as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \boldsymbol{\mu_\theta}(x_t, t, \boldsymbol{c}), \tilde{\beta}_t\mathbf{I}). \tag{3}$$

Note that we set the variances to untrained constants $\tilde{\beta}_t\mathbf{I}$ following [**?**], and $\boldsymbol{\mu_\theta}(x_t, t, \boldsymbol{c})$ is typically implemented as neural networks. Unlike vanilla diffusion models, we include the conditional context $\boldsymbol{c}$ as additional network inputs, which represent audio mixture and visual information in our task.

To train the network, a simplified way is to penalize the ϵ-prediction with the $\mathcal{L}_2$ loss, which is equal to predict $\boldsymbol{\mu_\theta}$ according to its parameterization [**?**]:

$$\mathcal{L}_2(\theta) = \mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \boldsymbol{c}, t)||^2]. \tag{4}$$

Here, ϵ_θ represents a function approximator used to predict the noise added at each iteration, while t is a uniformly sampled value ranging from 1 to T.

3.2 Task Setup and Method Overview

Given an unlabeled video clip V, we can extract an audio-visual pair (a, v), where a and v are the audio and visual streams, respectively. In real-world scenarios, the audio stream can be a mixture of N individual sound sources, denoted as $a = \sum_{i=1}^{N} s_i$, where each source s_i can be of various categories. Meanwhile, the visual stream v is typically a synchronized video of K frames, denoted as $v = \{I_j\}_{j=1}^{K}$. The visually-guided sound source separation task aims to utilize visual cues from v to help separate a into N individual sources s_i. Since no labels are

provided to distinguish the sound sources s_i, prior works [?, ?, ?] have commonly used a "mix and separate" strategy, which involves mixing audio streams from two different videos and manually create the mixture: $a^{mix} = a^{(1)} + a^{(2)}$. In practice, audio is usually transformed into magnitude spectrogram by short-time Fourier transform $x = \mathbf{STFT}(a) \in \mathbb{R}^{T \times F}$, allowing for manipulations in the 2D Time-Frequency domain. Here, F and T are the numbers of frequency bins and time frames, respectively. Consequently, the goal of training is to learn a separation network capable of mapping $\boldsymbol{f : (x^{mix}, v^{(i)}) \rightarrow x^{(i)}}$. For simplicity, we will omit the video index notation in the subsequent sections[3].

In contrast to conventional approaches that perform the mapping through regression, our proposed DAVIS framework is built on a diffusion model with a T-step diffusion and reverse processes. The diffusion process is determined by a fixed variance schedule as described in Eq. (1), which gradually adds noises to the magnitude spectrogram x_0 and converts it to latent x_T. As depicted in Fig. 2, the reverse process (according to Eq. (2) and Eq. (3)) of DAVIS is specified by our proposed Separation U-Net ϵ_θ. This reverse process iteratively denoises a latent variable x_T, which is sampled from a uniform distribution, to obtain a separated magnitude conditioned on the magnitude of the input sound mixture x^{mix} and the visual stream v. Consequently, the objective of the Separation U-Net ϵ_θ is to predict the noise ϵ added at each diffusion timestep during forward.

3.3 Proposed DAVIS Framework

Diffusion models often use U-Net-like [?] architectures, which excel at capturing multi-level feature representations and maintaining the output shape identical to the input. These properties also make them well-suited for the audio-visual separation task in the network aspect. However, naively applying existing conditional diffusion models to audio-visual separation is ineffective, as they are typically designed for image-to-image translation [?] or text-to-image synthesis [?, ?]. These models utilize different condition mechanisms than those required for audio-visual tasks, and they are not tailored to address the unique characteristics of audio-visual data. Therefore, the development of a specialized audio-visual separation network for diffusion models is essential. In this context, we revisit the challenges that need to be addressed: (1) Similar frequency patterns commonly exist even in temporally distant time frames, which necessitates the network to capture both long-range dependencies across time and frequency dimensions, and thus pure convolution [?, ?] may fall short. (2) Real-world videos often have mismatched visual and audio content. Extracting visual condition [?, ?] without considering the possible unrelated audio-visual content can potentially lead to less discriminative visual cues. (3) Establishing precise audio-visual associations is crucial, but directly concatenating visual and audio embeddings at the bottleneck [?] lacks the ability to foster further interactions between the two modalities.

[3] In this paper, superscripts denote video indices, while subscripts usually refer to diffusion timesteps.

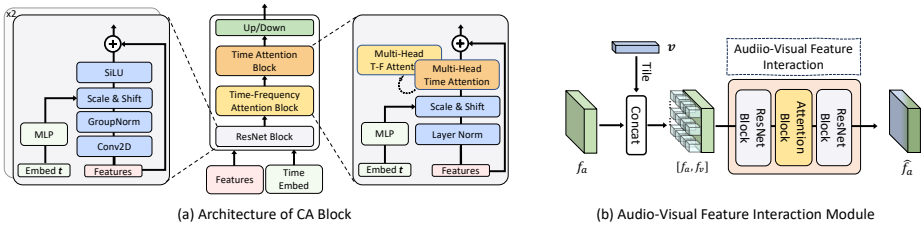


Fig. 3: Illustrations on (a) CA block: It operates by taking audio feature maps and a time embedding $\boldsymbol{t}$ as inputs. Each sub-block, except the up/down sampling layer, is conditioned on $\boldsymbol{t}$. ResNet and attention blocks are stacked to capture local and non-local audio contexts; **(b) Audio-Visual Feature Interaction Module:** It functions by replicating and concatenating $\boldsymbol{v}$ with $\boldsymbol{f_a}$, and uses two identical ResNet blocks and an attention block to process the concatenated features.

To address these challenges, we propose a novel and specialized Separation U-Net in our diffusion model that incorporates Convolution-Attention blocks to learn both local and global time-frequency associations, introduce a simple yet effective temporal transformer to aggregate the frame features and devise an audio-visual feature interaction module to enhance association learning by enabling interactions between audio and visual modalities.

Encoder/Decoder Designs. Our proposed Separation U-Net architecture consists of an encoder and a decoder, linked by an audio-visual feature interaction module. Both the encoder and decoder comprise five CA blocks. Initially, we concatenate the latent variable x_T with the mixture x^{mix} along the channel dimension and use a 1x1 convolution to project it into the feature space (another 1x1 convolution to convert the decoder output back to magnitude). As depicted in Fig. 3(a), each CA block consists of a ResNet block, a Time-Frequency Attention block, and a Time Attention block. Following this, a down-sample or an up-sample layer with a scale factor of 2 is used. Concretely, we build the ResNet block using WeightStandardized 2D convolution [**?**] along with GroupNormalization [**?**] and SiLU activation [**?**]. To incorporate the time embedding $\boldsymbol{t}$ as a condition, a Multi-Layer Perceptron (MLP) is used to generate $\boldsymbol{t}$-dependent scaling and shifting vectors for feature-wise affine transformation [**?**]. We also adopt an efficient form of attention mechanism [**?**] for implementing the Time-Frequency Attention block. To enhance the long-range time dependency modeling, a Time Attention block is then appended. In practice, we follow the design in [**?**], which includes Pre-Layer Normalization and Multi-Head Attention along the time dimension within the residual connection. The down-sample and up-sample layers are simply 2D convolutions with a stride of 2. As a result, we can obtain audio feature maps $\boldsymbol{f_a} \in \mathbb{R}^{C \times \frac{T}{32} \times \frac{F}{32}}$ at the bottleneck, where C represents the number of channels.

Timestep Embedding. In a diffusion model, the timestep embedding serves to inform the model about the current position within the Markov chain. As

shown in Fig. 2, timestep t is specified by the sinusoidal positional encoding [**?**] and further transformed by an MLP, which will be passed to each CA block as a timestep condition.

Visual Condition Aggregation. Not all frames in a video will be attributable to the synchronized audio. To account for unaligned visual content, we incorporate a shallow transformer to effectively aggregate the visual condition. Concretely, we extract frame features $\{\boldsymbol{I}_j\}_{j=1}^{K}$ from the visual stream v using a pre-trained visual backbone $\mathbf{Enc_v}$, where $\boldsymbol{I}_j \in \mathbb{R}^C$. We apply a self-attention temporal transformer $\phi(\cdot)$ to aggregate raw visual frame features, resulting in $\{\hat{\boldsymbol{I}}_j\}_{j=1}^{K} = \phi(\{\boldsymbol{I}_j\}_{j=1}^{K})$. For the transformer design, we empirically find that a shallow transformer with three encoder layers and one decoder layer works well. The global visual embedding $\boldsymbol{v}$ is then computed by averaging the temporal dimension of $\boldsymbol{v} = \frac{1}{K}\sum_{j=1}^{K}\hat{\boldsymbol{I}}_j$.

Audio-Visual Feature Interaction Module. The key to audio-visual separation lies in effectively utilizing visual information to separate visually-indicated sound sources. Therefore, the interaction between audio and visual modalities at the feature level becomes crucial. Existing approaches often concatenate audio and visual features at the bottleneck [**?**, **?**] and pass them to the decoder for further fusion. This design, however, imposes a dual task on the decoder: to integrate visual cues while simultaneously reconstructing the audio signal. We hypothesize that enabling further audio-visual interaction at the bottleneck could potentially enhance the separation performance. To this end, we explore different interaction manners and propose an audio-visual feature interaction module to improve this capability (see Tab. 3). We spatially tile $\boldsymbol{v}$ to match the shape of $\boldsymbol{f_a}$, resulting in visual feature maps $\boldsymbol{f_v}$. Subsequently, the audio and visual feature maps are concatenated along channel dimension and fed into the feature interaction module (FIM): $\hat{\boldsymbol{f_a}} := \mathbf{FIM}([\boldsymbol{f_a}, \boldsymbol{f_v}])$, where $\hat{\boldsymbol{f_a}} \in \mathbb{R}^{C\times\frac{T}{32}\times\frac{F}{32}}$. The details of the module are illustrated in Fig. 3(b), including two ResNet blocks and a Time-Frequency Attention block to facilitate capturing audio-visual associations within both local and global regions.

3.4 Training Pipeline

Given the sampled audio-visual pairs from the dataset, we first adopt the "mix and separate" strategy and compute the magnitudes $x^{(1)}, x^{(2)}, x^{mix}$ with STFT. To align with the frequency decomposition of the human auditory system, we apply a logarithmic transformation to the magnitude spectrogram, converting it to a log-frequency scale. Additionally, we ensure consistent scaling by multiplying log-frequency magnitudes with a scale factor σ and clipping the values to fall within the range [0, 1].

$\mathcal{L}_1$ **Denosing Loss.** The visual frames are encoded to embeddings $\boldsymbol{v^{(1)}}$,$\boldsymbol{v^{(2)}}$. Taking video (1) as an example, we sample ϵ from a standard Gaussian distribution and t from the set $\{1, ..., T\}$. Then, we input $x_t^{(1)}, x^{mix}, \boldsymbol{v}^{(1)}, t$ to the Separation U-Net ϵ_θ and optimize the network by taking a gradient step on Eq. (4). While $\mathcal{L}_2$ loss is well-suited for Gaussian noise estimation, the distribution of magnitude spectrograms is usually left-skewed due to the silent time

frames and various frequency patterns, resulting in numerous regions with near-zero values. We hypothesize that the $\mathcal{L}_1$ loss is more robust to this type of data distribution and, therefore, be beneficial for training the denoising neural network. In practice, we use both videos (1) and (2) for training, and the final loss term is formulated as $\mathcal{L} = \mathcal{L}_1^{(1)}(\theta) + \mathcal{L}_1^{(2)}(\theta)$.

3.5 Silence Mask-Guided Sampling

Our inference process starts from a sampled latent variable x_T, and ends with a sample from the target distribution $p(x_0|x^{mix}, \boldsymbol{v})$, conditioned on the mixture and visual frame embedding. As the goal of separation is to predict the individual sound from the mixture, an observation is drawn: silent time frames in the mixture should also be silent in the separated sound, *i.e.*, the network can leverage portions of the mixture x^{mix} to sample an individual sound x_0.

Given a sampling step from time t to $t-1$, the transition distribution can be rewrited using Eq. (1) on the x^{mix} and Eq. (3) on the x_t:

$$x_{t-1}^{mix} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}x^{mix}, (1-\bar{\alpha}_{t-1})\mathbf{I}), \tag{5a}$$

$$x_{t-1} \sim \mathcal{N}(\boldsymbol{\mu_\theta}(x_t, t, \boldsymbol{c}), \tilde{\beta}_t\mathbf{I}), \tag{5b}$$

$$m = [x^{mix} < \delta_{silence}], \tag{5c}$$

$$\hat{x}_{t-1} = m \odot x_{t-1}^{mix} + (1-m) \odot x_{t-1}. \tag{5d}$$

The refined output $\hat{x}_{t-1}$ is fed into the subsequent sampling iteration, further reducing the distribution gap between the ground truth target sound and the prediction. We introduce the hyperparameter $\delta_{silence}$ as the threshold for determining the silence region, setting it to a fixed value of 0.002. With the above modification, the model is enforced to predict individual sounds as parts of the mixture akin to the conventional mask-based approaches. Thus, the model refrains from "hallucinating" content as generation tasks, and the separation performance is improved as well (see Tab. 4). The waveform s_i for the separated sound can be reconstructed by applying inverse STFT on the magnitude prediction and the original mixture phase.

4 Experiments

4.1 Experimental Setup

Datasets. Our model demonstrates the ability to handle mixtures of diverse sound categories. To evaluate our approach, we use AVE [**?**] and MUSIC [**?**] datasets, which cover musical instruments and open-domain sounds. The evaluation settings are described in detail below: AVE [**?**] contains 4143 10-second videos, including 28 diverse sound categories, such as *Church Bell*, *Barking*, and *Frying*, among others. The AVE dataset presents greater challenges as the audio in these videos may not span the entire duration and can be noisy, including off-screen sounds (*e.g.*, human speech) and background noise. In addition to

Methods	AVE [?]			MUSIC [?]		
	SDR ↑	SIR ↑	SAR ↑	SDR ↑	SIR ↑	SAR ↑
NMF-MFCC† [?]	-	-	-	0.92	5.68	6.84
Sound-of-Pixels† [?]	1.21	7.08	6.84	4.23	9.39	9.85
Co-Separation† [?]	-	-	-	6.54	11.37	9.46
Sound-of-Motions† [?]	1.48	7.41	7.39	-	-	-
Minus-Plus† [?]	1.96	7.95	8.08	-	-	-
Cascaded Filter† [?]	2.68	8.18	8.48	-	-	-
CCoL† [?]	-	-	-	7.74	13.22	11.54
AMnet† [?]	3.71	9.15	11.00	-	-	-
iQuery† [?]	5.02	8.21	12.32	11.17	15.84	14.27
DAVIS (ours)	4.86	9.13	9.92	**11.61**	**18.36**	**14.70**

Table 1: Comparison of our method to other audio-visual separation approaches on the AVE and MUSIC test set. The top three results are highlighted in red, orange, and yellow, respectively. The results noted by † are obtained from [?, ?]. Note that audio in AVE could include off-screen sounds and background noise, which may reduce the accuracy of the reported metrics.

the AVE dataset, we also evaluate our proposed method on the widely-used MUSIC [?] dataset, which includes 11 musical instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin, and xylophone. All the videos are clean solo and the sounding instruments are visible. For both datasets, we follow the same train/validation/test splits as in [?, ?].

Baselines. To the best of our knowledge, we are the first to adopt a generative model for the audio-visual source separation task. Thus, we compare DAVIS against the following discriminative methods: *NMF-MFCC* [?] which is an audio-only separation method; *Sound of Pixels* [?] and *Sound of Motions* [?] that learn ratio mask predictions with a 1-frame-based model or with motion as condition; *Multisensory* [?] that separates mixtures based on learning discriminative audio-visual representations; *Minus-Plus* [?] that separates sounds by recursively eliminating high-energy components from the sound mixture; *Cascaded Filter* [?] which separates sounds in a multi-stage manner; *Co-Separation* [?] that takes a single visual object as the condition to perform mask regression; *Cyclic Co-Learn* (CCoL) [?] which jointly trains the model with sounding object visual grounding and visually-guided sound source separation tasks; *AMnet* [?] which is a two-stage framework modeling both appearance and motion; *iQuery* [?] that adapts the maskformer architecture for audio-visual separation and achieves the current state-of-the-art (SOTA) results.

Evaluation Metrics. To quantitatively evaluate the audio-visual sound source separation performances, we use the standard metrics [?, ?, ?], namely: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). We adopt the widely-used mir_eval library [?] to report

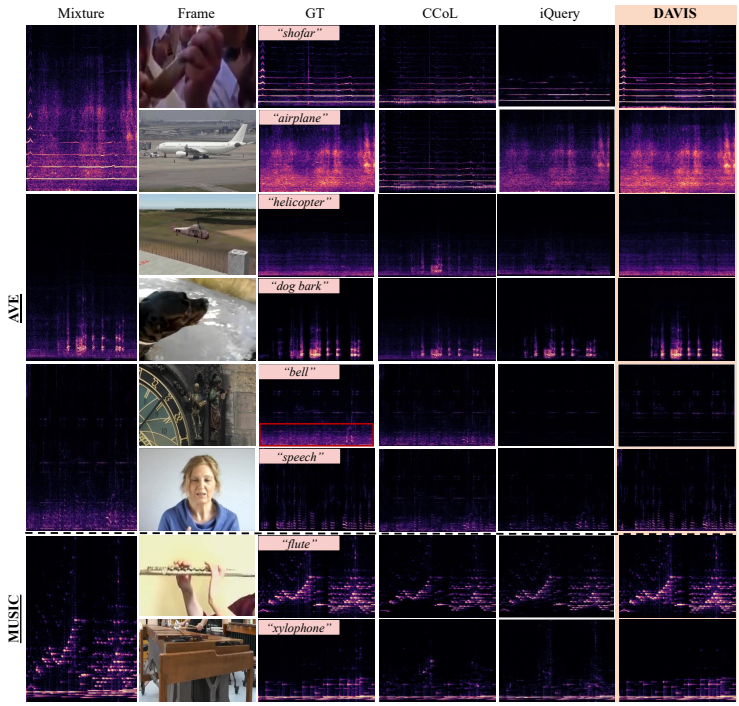


Fig. 4: Visualizations of audio-visual separation results on the AVE (the top three mixtures) and MUSIC (the last mixture) datasets. Two sounds are mixed, and reference frames are provided to guide the separation. The comparison is shown between the predictions made by DAVIS (ours), iQuery [**?**], and CCoL [**?**] with the ground truth. DAVIS can effectively separate sound mixtures from various categories, such as *airplane*, *rats*, and *dog barking*.

the standard metrics. Note that SDR and SIR evaluate the accuracy of source separation, whereas SAR specifically measures the absence of artifacts [**?**].

4.2 Comparisons with State-of-the-art

To assess the effectiveness of our method, we have compared DAVIS with state-of-the-art approaches on the AVE and MUSIC datasets. The comparison results are presented in Tab. 1. Our results demonstrate the benefits of using generative modeling for audio-visual separation. DAVIS achieves comparable SDR results to the strong baseline iQuery [**?**] while improving the SIR scale by **0.9 dB**. It is worth noting that the results on AVE are not as clear-cut as those on MUSIC. We believe there are two reasons behind this: firstly, iQuery uses ground truth class labels to choose a mask in both training and inference, thereby yielding stronger conditional signals. In contrast, DAVIS only uses video frames as the condition. Secondly, the original AVE audio clip often contains off-screen

Block	SDR↑	SIR↑	SAR↑	# Params (M)
{R, R, R}	9.03	14.05	13.20	51.76
{R, R, T}	11.78	17.91	15.44	43.85
{R, R, TF}	11.50	18.01	15.21	42.95
{R, TF, T}	11.88	17.52	16.12	35.04

Table 2: Ablation on CA block design. R, TF, and T denote ResNet, Time-Frequency, and Time Attention blocks, respectively. We highlight the setting used in this paper in gray.

Fusion	SDR↑	SIR↑	SAR↑
Concat	10.85	17.62	15.52
FIM (Point-wise)	11.06	17.37	15.44
FIM (Local)	11.56	17.02	16.28
FIM (Global)	11.23	17.56	15.84
FIM (Local&Global)	11.88	17.52	16.12

Table 3: Ablation study on Feature Interaction Module. We explore different ways of integrating audio and visual features.

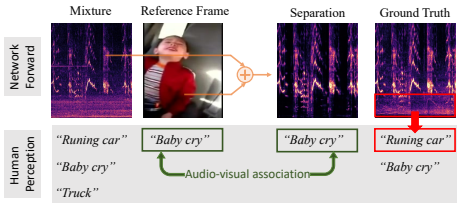


Fig. 5: A visualization example showing that our DAVIS model can capture accurate audio-visual association.

$\delta_{silence}$	SDR↑	SIR↑	SAR↑
0 (baseline)	11.53	18.30	14.74
0.01	11.15	18.21	14.69
0.001	11.50	18.26	14.74
0.002	11.61	18.36	14.70

Table 4: The effect of silence mask-guided sampling strategy on the MUSIC test set.

sounds and background noise, which the conventional metrics (*e.g.*, SDR, SIR, and SAR) cannot accurately capture, even if the separation results have removed the background noise simultaneously. To address this issue, we provide qualitative visualization in Fig. 4, which reinforces the advantages of DAVIS. On the clean MUSIC dataset, DAVIS consistently outperforms existing methods across various evaluation metrics, surpassing the next best approach iQuery. These results clearly demonstrate DAVIS's versatility across diverse datasets with varying categories, while the subjective test in Sec. 4.3 also supports our claim.

4.3 Experimental Analysis

We conduct ablations on the MUSIC validation set (unless specified) to examine the different components of DAVIS. *For more ablations, please refer to the supplementary materials.*

Block Design. We validate the effectiveness of our proposed CA block (shown in Fig. 3) by designing the following baselines: (a) using three consecutive ResNet blocks within the CA block, which only captures local time-frequency patterns; (b) replacing the last ResNet block with a Time Attention block; (c) replacing the last ResNet block with a Time-Frequency Attention block; and (d) replacing the last two ResNet blocks with Time-Frequency and Time attention blocks to enhance the capability of modeling long-range dependency. The results in Tab. 2 underscore the importance of learning both local and global contexts across time

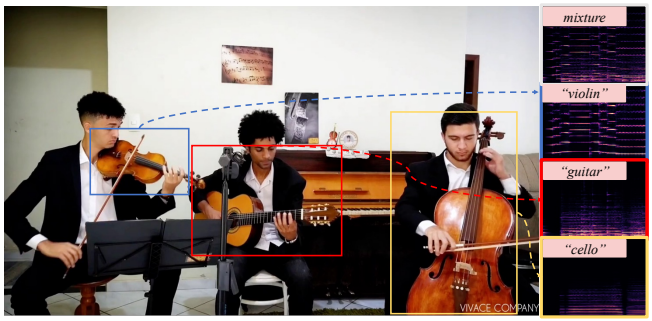


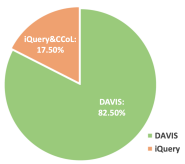
Fig. 6: Separation results on a real-world challenging three-source example. YouTube ID: *R1DCTNEMibw*.

and frequency dimensions. Furthermore, the comparison of model sizes confirms that the improvements are not attributable to increased network capacity.

Audio-Visual Feature Interaction. To validate the importance of effective audio-visual association learning for this task, we conduct an ablation study on the Feature Interaction Module. Specifically, we explore different ways of feature interaction: (a) direct concatenation of visual and audio features, (b) a three-layer MLP for point-wise fusion, (c) three ResNet blocks, (d) three attention blocks, and (e) a combination of ResNet and attention blocks. The results presented in Tab. 3 show that naive concatenation of audio and visual features performs significantly poorly while enabling further interaction between them improves the results. Among all the designs, our proposed module achieves the best results by considering both local and non-local contexts.

Learned Audio-Visual Association. To showcase the accuracy of our model's learned audio-visual associations, we mixed a "Baby crying" video clip with a "Truck" video clip from the AVE dataset. As shown in Fig. 5, the original baby video, as perceived by human listeners, also contains a running car sound, thus establishing a complicated audio-visual relationship. Our model successfully extracts the baby's crying sound while eliminating all irrelevant sounds, demonstrating DAVIS's ability to learn accurate audio-visual associations.

Effects of Silence Mask-Guided Sampling. As shown in Tab. 4, we experiment with different thresholds for the silence mask-guided sampling method, which determines the proportion of re-used information from the mixture. While a high threshold may introduce leakage from non-silent regions (*e.g.*, from the second sound), we show that carefully selecting the threshold value can boost the separation performance in the post-training stage compared to the baseline.



Qualitative Visualization on Natural Sound Mixture. To further demonstrate our model's effectiveness, we present a challenging real-world example of testing it on a natural sound mixture with multiple sounds (see Fig. 6).

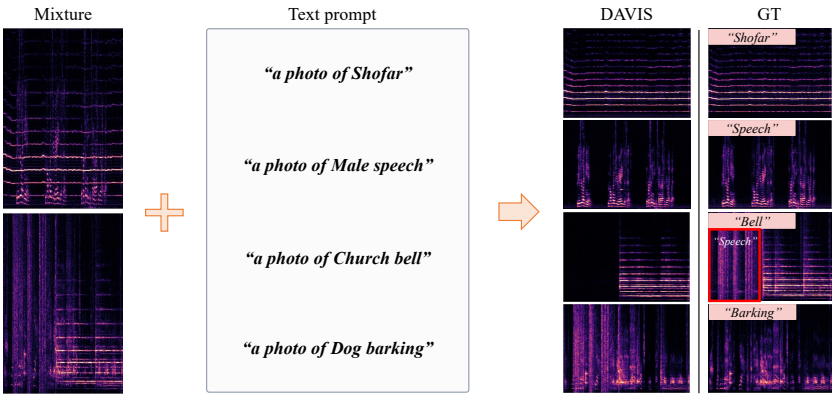


Fig. 7: Qualitative examples of zero-shot text-guided source separation. Notably, in *the third row* example, we observe the model's ability to capture precise audio-text correspondence by successfully filtering out the *"speech"* sound.

Subjective test. We conduct a subjective test of separation results of our model and strong baselines iQuery/CCoL. 11 participants are asked to answer *"Which separation result is closer to the ground truth audio and better matches the frame content?"*, with GT sound and frame as a reference. DAVIS outperformed the other baselines with a **winning rate of 82.5%** from 176 results, as shown in the figure to the right.

4.4 Application: Zero-Shot Text-guided Separation

Our model trained to capture the conditional distribution $p(x|\mathbf{v})$ can be employed for zero-shot inference from $p(x|\mathbf{t})$ where $\mathbf{t}$ represents the text description corresponding to the image $\mathbf{v}$. We achieve this by leveraging the well-established shared image-text embedding space from the CLIP [**?**] model. We qualitatively evaluate the results of utilizing replacement conditioning for separation and find it to be surprisingly effective, as shown in Fig. 7.

5 Conclusion

In this paper, we propose DAVIS, a diffusion model-based audio-visual separation framework designed to address the problem in a generative manner. Unlike conventional discriminative methods, DAVIS is built upon a T-step diffusion model, enabling the iterative synthesis of the separated magnitude spectrogram conditioned on the visual input. By proposing a specialized Separation U-Net coupled with a novel sampling strategy, we successfully apply diffusion model to this new task, yielding high-quality sound separation results. Extensive experiments on the MUSIC and AVE datasets validate DAVIS's effectiveness in separating sounds within specific and open domains, as well as its ability to deal with diverse time-frequency structures.

References