Supplementary Materials for "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

1 Demo Page: the "demo.html" file in the folder

We have created a demo page to illustrate our method and showcase our separation results. **We strongly encourage readers to visit this webpage**. Please note that the webpage may not be fully compatible with the Safari browser; therefore, we recommend using Google Chrome for an optimal viewing experience. To access the contents of demo, please visit https://wikichao.github.io/data/projects/DAVIS/.

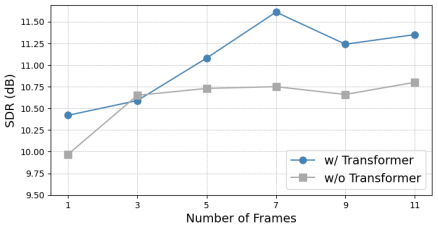

Fig. 1: Ablation on varying the number of frames to validate the effect of our proposed temporal transformer.

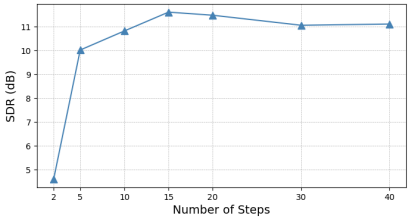

Fig. 2: Ablation study of the number of sampling steps with DDIM [5] acceleration technique. T: {2, 5, 10, 15, 20, 30, 40}.

2 More Ablations

Aggregated Visual Condition. To study the impact of the visual condition, we vary the number of sampled frames and compare models with and without the temporal transformer, as shown in Fig. 1. The results reveal that increasing the number of frames achieves a more informative visual condition and boosts separation quality. However, as the number of frames increases, noisy information

Table 1: Overview of our constructed **VGGSound-Animal**10 and **VGGSound-Vehicle**10 datasets. Each dataset contains 1000/100 videos for training/testing.

VGGSound-Animal10	VGGSound-Vehicle10
parrot talking	*tractor digging*
dog growling	*driving snowmobile*
cow lowing	*reversing beeps*
cat hissing	*helicopter*
gibbon howling	*train whistling*
wood thrush calling	*airplane flyby*
snake hissing	*railroad car, train wagon*
baltimore oriole calling	*driving motorcycle*
bull bellowing	*engine accelerating, revving, vroom*
snake rattling	*fire truck siren*

may be introduced, leading to a decline in performance. We show that adopting a temporal transformer effectively alleviates this issue, resulting in better separation performance.

Analysis on Sampling Step. We analyze the effect of the number of sampling steps in Fig. 2. We select a set of different $T = \{2, 5, 10, 15, 20, 30, 40\}$ with DDIM [5] for sampling acceleration. Our results reveal that a limited number of steps (*e.g.*, $<$10) is insufficient for effectively separating the sounds. On the other hand, the curve tends to converge with a higher number of sampling steps (*e.g.*, $>$15). This suggests that our model is capable of achieving satisfactory separation results without requiring a large number of inference steps. In practice, we set $T = 15$ for our model.

3 More Qualitative Visualizations from Diverse Categories

To demonstrate the effectiveness of our method in separating sounds across diverse categories, we conducted experiments on VGGSound [1]. VGGSound is a large-scale audio-visual dataset encompassing a wider range of sounds compared to commonly used datasets like AVE [6] and MUSIC [7]. However, since VGGSound isn't a standard benchmark for assessing audio-visual separation performance yet, we thus constructed two smaller datasets focusing on the two main subcategories within VGGSound: Animals and Vehicles.

We named these two datasets **VGGSound-Animal**10 and **VGGSound-Vehicle**10. For each subcategory, we randomly selected 10 classes and sampled 100 videos for training and 10 videos for testing within each class. This resulted in a training/testing split of 1000/100 videos for each dataset, exceeding the scale of the frequently used MUSIC dataset (11 categories, 468/26 videos for training/testing). Therefore, we believe our custom datasets provide a meaningful evaluation for audio-visual separation tasks, and we leave the exploration

of constructing a larger, more comprehensive dataset with additional categories and more comprehensive comparison for future work.

The table in Tab. 1 displays information about the classes in each dataset. We present qualitative results in Fig. 3 and Fig. 4, which demonstrate that DAVIS can separate high-quality sounds from diverse categories.

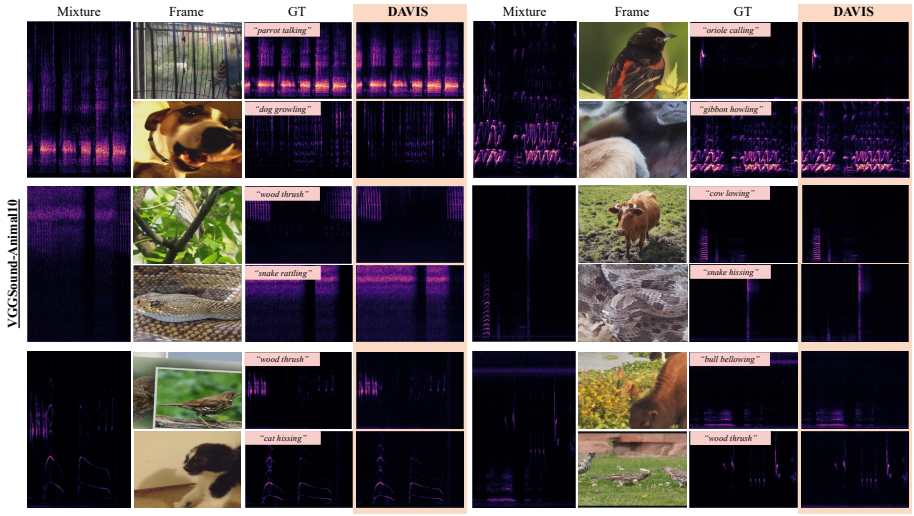

Fig. 3: Qualitative visualizations on VGGSound-Animal10.

4 Implementation Details

In our experimental setup, we down-sample audio signals at 11kHz. For the MUSIC dataset, the video frame rate is set to 8 fps. Each video is approximately 6 seconds and we uniformly select 11 frames per video. The pre-trained ResNet18 [2] is used as the image encoder. As for the AVE dataset, we set the video frame rate to 1 fps (following the setup of [6]). We use the entire 10-second audio as input and use 10 frames to train the model. The pre-trained CLIP image encoder [4] is used. During training, the frames are first resized to 256×256 and then randomly cropped to 224×224. We set the total diffusion time step $T = 1000$ to train our DAVIS model. During inference, all the frames are directly resized to the desired size without cropping. To accelerate the separation process, we use DDIM [5] with a sampling step of 15. The audio waveform is transformed into a spectrogram with a Hann window of size 1022 and a hop length of 256. The obtained magnitude spectrogram is subsequently resampled to 256×256 to feed into the separation network. We set the number of audio and visual feature channels C as 512 and empirically choose the scale factor $\sigma = 0.15$. Our model

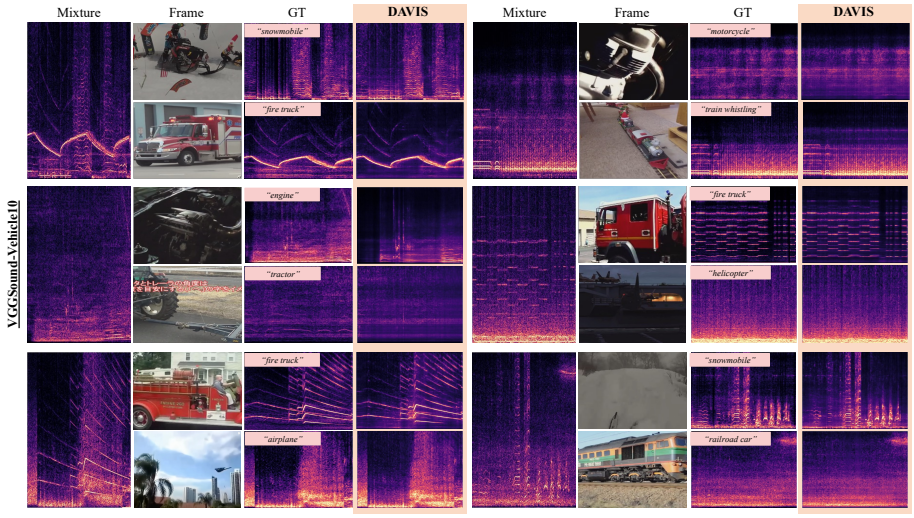

Fig. 4: Qualitative visualizations on VGGSound-Vehicle10. There is a lot of background noise in this dataset, which makes it difficult to separate and poses significant challenges in achieving high-quality results.

is trained with the Adam optimizer, with a learning rate of 10^{-4}. The training is conducted on two 4090 GPUs for 150 epochs with a batch size of 8.

5 Training and Inference Pseudo Code

The complete training procedure for our DAVIS framework is shown in Algorithm 1. Given the sampled audio-visual pairs from the dataset, we first use the "mix and separate" strategy to create the mixture, and compute the magnitudes $x^{(1)}, x^{(2)}, x^{mix}$ using STFT. We then apply a logarithmic transformation to the magnitude spectrogram to convert it to a log-frequency scale. Finally, we ensure consistent scaling of the log-frequency magnitudes by multiplying by a scale factor σ and clipping to the range [0, 1].

The visual frames are encoded to embeddings with the pre-trained visual backbone and aggregated by a trainable temporal transformer followed by an averaging operation. This gives us the visual conditions $\boldsymbol{v^{(1)}}$,$\boldsymbol{v^{(2)}}$. For the training process, 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. In practice, we use both video (1) and (2) for optimization.

As illustrated in Algorithm 2, our inference process starts from a sampled latent variable x_T, and takes the mixture x^{mix} and visual condition $\boldsymbol{v}$ to produce the separated magnitude x_0 through T iterations. At each iteration, we adopt the silence mask-guided sampling strategy to refine the output. In the end, the output is rescaled to the original range.

Algorithm 1 Training

Input: A dataset D that contains audio-visual pairs $\{(a^{(k)}, v^{(k)})\}_{k=1}^{K}$, total diffusion step T
Initialize: randomly initialize Separation U-Net ϵ_θ and temporal transformer $\phi(\cdot)$, and load the pre-trained visual encoder $\mathbf{Enc_v}$
repeat
  Sample $(a^{(1)}, v^{(1)})$ and $(a^{(2)}, v^{(2)}) \sim D$
  Mix and compute $x^{mix}, x^{(1)}$
  Scale $x = \log_e(1+x) \cdot \sigma$ and clip $x^{mix}, x^{(1)}$ to [0,1]
  Encode visual frames $v^{(1)}$ as $\boldsymbol{v^{(1)}} := \phi(\mathbf{Enc_v}(v^{(1)}))$
  Sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $t \sim \text{Uniform}(1, ..., T)$
  Take gradient step on
    $\nabla_\theta ||\epsilon - \epsilon_\theta(x_t^{(1)}, x^{mix}, \boldsymbol{v}^{(1)}, t)||, x_t^{(1)} = \sqrt{\bar{\alpha}_t} x^{(1)} + \sqrt{1-\bar{\alpha}_t}\epsilon$
until converged

Algorithm 2 Inference

Input: Audio mixture a^{mix} and the query visual frame v, total diffusion step T
Sample $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Compute $x^{mix} := \mathbf{STFT}(a^{mix})$
Encode visual frames v as $\boldsymbol{v^{(1)}} := \phi(\mathbf{Enc_v}(v))$
for $t = T, ..., 1$ **do**
  Sample $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $z = 0$
  Compute x_{t-1}: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, x^{mix}, \boldsymbol{v}, t)) + \sqrt{\tilde{\beta}_t} z$
  $x_{t-1} = \text{silence\_mask\_guided\_sampling}(x_{t-1})$
end for
return $e^{x_0/\sigma} - 1$

6 Ablation on $\mathcal{L}_1$ Training Loss

Tab. 2 shows that an $\mathcal{L}_1$ loss in training the diffusion model performs better than an $\mathcal{L}_2$ loss for audio-visual separation on both MUSIC and AVE datasets. It's because of the presence 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.

7 Limitation

Visual Embeddings. Our proposed DAVIS framework incorporates the extraction of global visual embedding as a condition for visually-guided source separation. This technique, which utilizes global visual features, has been widely adopted in audio-visual learning [3, 7]. Unlike methods that rely on pre-trained

Table 2: Ablation for the choice of loss function on the MUSIC and AVE datasets.

	MUSIC [7]			AVE [6]		
loss function	SDR↑	SIR↑	SAR↑	SDR↑	SIR↑	SAR↑
$\mathcal{L}_2$	10.84	17.52	14.55	4.53	8.26	10.14
$\mathcal{L}_1$	**11.61**	**18.36**	**14.77**	**4.86**	**9.13**	**9.92**

object detectors for extracting visual features, our framework does not have such a dependency. However, it may encounter limitations when trained on unconstrained video datasets. Intuitively, successful results can be achieved when the video contains a distinct sounding object, such as solo videos in the MUSIC dataset or videos capturing a sounding object performing a specific event in the AVE dataset. Nonetheless, this training assumption may not hold in more challenging scenarios, where multiple objects are likely producing sounds, rendering the global visual embedding inadequate for accurately describing the content of sounding objects. To address this issue, one possible approach is to adapt our framework to leverage more fine-grained visual features and jointly learn sounding object localization and visually-guided sound separation. This adaptation enables the model to utilize localized sounding object information to enhance the audio-visual association.

Evaluation Metrics. We observe from the examples on the AVE and VGGSound datasets that many video clips contain off-screen sound or background noise, rendering the notion of ground truth unsuitable for evaluation. Consequently, comparing separation results with the ground truth audio clip and reporting SDR/SIR/SAR values may be insufficient to assess the method's effectiveness. Therefore, a new metric is needed to evaluate sound separation quality in more noisy or challenging scenarios. One possible metric is to leverage pretrained audio-text models and perform zero-shot classification or measure the cosine similarity between separated audio and the text label (analogous to zero-shot scenarios using the CLIP model).

8 Future Work

Our approach initiates the utilization of generative models for audio-visual scene understanding, paving the way for potential extensions to other multi-modal perception tasks like audio-visual object localization. Humans demonstrate the ability to imagine a "dog" upon hearing a "barking" sound, highlighting the potential of cross-modal generation in advancing audio-visual association learning. This implies that localization and separation tasks can be integrated into a single generative framework. In the future, we plan to explore the application of generative models to jointly address audio-visual localization and separation tasks.

More Types of Conditions. We have investigated the use of visual frame features and text prompts as conditions for sound separation in our work. These conditions are effective for separating sounds from different categories. However, for a more challenging scenario such as separating sounds from the same category, we need a different type of condition that provides discriminative cues to guide separation. Examples of such conditions are optical flow and trajectory. In our future work, we plan to incorporate more conditions in our framework.

References

1. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 2
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 3
3. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. arXiv preprint arXiv:2303.13471 (2023) 5
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3
5. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 1, 2, 3
6. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018) 2, 3, 6
7. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018) 2, 5, 6