This CVPR paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Learning to Highlight Audio by Watching Movies

Chao Huang¹, Ruohan Gao², J. M. F. Tsang³, Jan Kurcius³, Cagdas Bilen³, Chenliang Xu¹, Anurag Kumar³, Sanjeel Parekh³ ¹University of Rochester, ² University of Maryland, College Park, ³Meta Reality Labs Research

Abstract

Recent years have seen a significant increase in video content creation and consumption. Crafting engaging content requires the careful curation of both visual and audio elements. While visual cue curation, through techniques like optimal viewpoint selection or post-editing, has been central to media production, its natural counterpart, audio, has not undergone equivalent advancements. This often results in a disconnect between visual and acoustic saliency. To bridge this gap, we introduce a novel task: visually-guided acoustic highlighting, which aims to transform audio to deliver appropriate highlighting effects guided by the accompanying video, ultimately creating a more harmonious audio-visual experience. We propose a flexible, transformer-based multimodal framework to solve this task. To train our model, we also introduce a new dataset—THE MUDDY MIX DATASET, leveraging the meticulous audio and video crafting found in movies, which provides a form of free supervision. We develop a pseudo-data generation process to simulate poorly mixed audio, mimicking real-world scenarios through a three-step process-separation, adjustment, and remixing. Our approach consistently outperforms several baselines in both quantitative and subjective evaluation. We also systematically study the impact of different types of contextual guidance and difficulty levels of the dataset. Our project page is here: https://wikichao.github.io/VisAH/.

1. Introduction

Be it amateur recordings of memorable moments or professionally created content—telling a story and delivering the best audio-visual experience with a video requires the right balance of audio and visual elements in the scene. Consider for example, the scene in Fig. 1c that depicts a video of a man talking in the sea. The scene is best represented when the focus is on the person and the sea at the appropriate moments. While there are several ways to visually highlight the intended objects during capture or in post-production [50, 62], they remain relatively under-explored and limited for the acoustics. In our case above, the man's speech can



(c) Task Inference

Figure 1. We propose a new task that aims to transform poorly mixed audio into a well-balanced mix using visual guidance. One of our key insights is to use well-curated audio-visual content from a movie database as free supervision to learn the appropriate highlighting effect for audio (L2H).

be obscured by the sound of crashing waves. Is it possible to automatically adjust the levels of the speech and the wave sounds according to the video content to ensure they are both acoustically salient and well-balanced?

A naive approach would involve first demixing the sounds into different source components and then remixing them at their respective intended levels. But doing so has two major drawbacks: (i) imperfect demixing could lead to the highlighting of undesired sources, and (ii) ensuring the right temporal variations and alignment with the video manually is a laborious process. Some research efforts, such as in music remixing [26, 36, 66], focus on adjusting the levels and effects of individual instruments to recompose music tracks. However, this approach is limited to the music domain, neglecting the broader needs of natural audio composition across varied media contexts. In this paper, we aim to bridge this gap by introducing a novel task, **visually-guided acoustic highlighting**. Our approach builds on the hypothesis that the visual stream in media is often curated with intent, implicitly conveying highlighted content. In contrast, due to the limitations of recording devices, such as microphones attached to video cameras that capture all sounds indiscriminately, audio often lacks intentional mixing, resulting in a poorly balanced track. The goal of our task is to use the video as guidance to transform the poorly mixed audio with appropriate highlighting effects, ensuring a better output audio mix.

Dataset is a key requirement for training a learning-based model to perform this task. We observe that movies are inherently well-curated where audio is meticulously crafted alongside video to create intentional highlighting effects. This provides us with free supervision for the acoustic highlighting effect as illustrated in Fig. 1. Consequently, we build a new dataset using movie clips spanning several generes from the Condensed Movie Dataset (CMD) [2]. To simulate real-world scenarios where audio may be poorly mixed or require enhancement, we introduce a pseudo-data generation process that begins with high-quality movie audio and then applies imperfect separation, followed by adjustment and remixing of individual audio sources.

We tackle acoustic highlighting as an audio-to-audio translation problem and propose a transformer-based Visually-guided Acoustic Highlighting (VisAH) model. VisAH uses a U-Net-like audio backbone with a dual encoder [7] that takes both the spectrogram and the waveform as inputs to extract latent representations from the poorly mixed audio. In this latent space, a transformer encoder processes context, such as the video stream or its corresponding caption, and a transformer decoder with cross-attention integrates this video context to guide the transformation. The decoder then converts the poorly mixed audio representation into highlighted audio. This design is flexible, supporting easy adjustments in both the backbone and latent modules. Specifically, our model combines video context encoding and visually-guided audio decoding, enabling it to fully capture temporal and semantic trends in the video and leverage them for effective audio highlighting.

To summarize, our main contributions are threefold:

- We propose to intelligently highlight the audio content in a video guided by visual cues, and we design VisAH, a mutimodal transformer-based model to achieve that goal.
- Leveraging the free supervision from movies, where both audio and video are already meticulously crafted, we introduce THE MUDDY MIX DATASET, a new dataset curated for this task.
- Our method outperforms a series of baselines, effectively highlighting audio across different types of video content.

2. Related Work

Audio Remixing. Highlighting a mixed audio track is identical to rebalancing its individual sources, *i.e.*, transferring from one mixing style to another. In previous research, music mixing [26, 38, 49, 59] has been extensively studied, including creative manipulations that shape a song's emotive and sonic identity. Reproducing the mixing style of a target song typically involves balancing tracks using audio effects to achieve harmony and aesthetic appeal, often through knowledge-based [44] or learning-based [38, 54, 59] approaches. Related but different from these methods that focus on music, we handle a broad range of sounds, including speech, music, and sound effects. Moreover, we propose to use visual cues in videos to guide the highlighting process.

Video Highlight and Saliency Detection. Web videos, often created by professionals, are typically edited, such as trimming to capture key moments or adjusting camera focus to highlight engaging regions. This has led to the development of video understanding tasks, including video highlight detection [29, 33, 35, 40], which identifies key temporal segments, and video saliency prediction [16, 22–24, 28], which identifies salient regions within a scene. Developing methods to smartly emphasize the right visual content in a video and the the development of corresponding benchmark datasets [15, 29] have become essential areas of study. In our work, we focus on the mirror side of the problem: assuming the video is already well-curated to visually convey highlight information, as in movies, we leverage this visual narrative to highlight the audio stream accordingly.

Audio-Visual Learning. Exploring connections between acoustic and visual signals has been widely studied across various tasks, including audio-visual localization [17, 19, 39, 46, 56–58], representation learning from cross-modal supervision [1, 12, 42], audio-visual learning for 3D scenes [21, 30, 31], audio-spatialization leveraging visual cues [9, 13, 41], and sound generation from videos [3, 5, 32, 43]. Differently, we tackle a new challenging audio-visual learning task, visually-guided acoustic highlighting.

Most closely related to our task is visually-guided audio separation [10, 11, 18, 20, 42, 68], which addresses a specific case of our problem by isolating a target sound and suppressing others to zero, thereby achieving separation. However, these methods lack a well-defined output for a balanced mix and do not address the creation of poorly mixed inputs. Our approach offers a new perspective by focusing on remixing audio to produce a coherent, visually aligned output.

3. Approach

We introduce the task of visually-guided acoustic highlighting, which aims to transform audio using visual guidance to achieve an appropriate highlighting effect. In Sec. 3.1,



(a) Visually-guided Acoustic Highlighting Network

(b) Latent Highlighting Transformer

Figure 2. Overview of VisAH: (a) Our model takes a poorly mixed audio waveform as input and produces the highlighted audio using a dual U-Net architecture. For simplicity, skip connections are omitted in the illustration. (b) The latent highlighting transformer incorporates vision and text encoders to integrate temporal information, guiding the transformer decoder to transform audio features effectively.

we elaborate on the task formulation and Sec. 3.2 discusses design of our proposed multimodal model VisAH.

3.1. Task Formulation

Let $\mathbf{v} \in \mathbb{R}^{T_v \times H \times W \times 3}$ represent the visual stream, a sequence of T_v RGB images, and let $\mathbf{a} \in \mathbb{R}^{T_a}$ be the corresponding audio sequence. Here \mathbf{a} is considered a poorly mixed audio sequence that lacks the intended highlighting effect, as is common in raw daily recordings where audio is directly captured without deliberate curation. Our goal is to predict an audio signal s that preserves the content of \mathbf{a} but conveys the appropriate highlighting effects. In other words, we learn a mapping from the poorly mixed audio signal:

$$\mathbf{a}, \mathbf{v} \mapsto \mathbf{s}.$$
 (1)

Creating a highlighting effect in audio involves rebalancing the sources at different levels to reflect their relative prominence. When describing system design next we will assume access to data tuples of the form $(\mathbf{a}, \mathbf{v}, \mathbf{s})$. How such a dataset is created and utilized in practice will be the subject of our discussion in Sec. 4.

3.2. VisAH: Visually-guided Acoustic Highlighting

To address the problem formulated in Eq. (1), we focus on two primary design choices: (i) how to structure the audio framework to accept poorly mixed audio as input and produce highlighted audio, and (ii) how to effectively incorporate contextual information, such as video streams or other modalities like text, to guide the acoustic highlighting. In this section, we elaborate on our framework design, presenting a flexible approach that includes both an audio backbone and a context-aware module. The overall architecture of VisAH is illustrated in Fig. 2.

3.2.1 Audio Backbone

The audio backbone is designed to produce an output with the same shape as the input, making U-Net [51] architectures particularly suitable. We consider two common types of audio input: time-domain and frequency-domain. Frequency-domain input, often used in audio-visual separation tasks [4, 58, 68], captures distinct frequency patterns of sounds, while time-domain input, frequently employed in audio-only separation [37, 55], provides higher accuracy in reconstructing the final waveform. In this work, we unify the advantages of both input types within our audio backbone, offering flexibility for future studies to utilize either or both. Based on the HybridDemucs architecture [7], we implement a dual U-Net model with two branches: one for spectrogram inputs and the other for waveform inputs.

Spectrogram U-Net Encoder. Given an input audio signal $\mathbf{a} \in \mathbb{R}^{T_a}$, we apply a Short-Time Fourier Transform (STFT) to a with a window size of 4096 and hop length of 1024 to obtain its spectrogram. Specifically, we use the magnitude spectrogram as input, denoted by A. In the original Hybrid-Demucs [7], the spectrogram is normalized using its mean and standard deviation. However, we omit this normalization, as we found that mean normalization can significantly suppress ambient sounds, reducing the model's sensitivity to sound effects. The magnitude encoder consists of 5 layers, with each layer reducing the number of frequency bins by a factor of 4, except for the final layer, which reduces it by a factor of 8. After passing through the magnitude encoder, the frequency dimension is reduced to 1, aligning it with the output shape from the waveform branch. Details of the encoder design can be found in the supplementary materials. Waveform U-Net Encoder. In this framework, the waveform branch acts as a residual path to capture fine-grained temporal details. To facilitate processing, we normalize the

waveform input **a**. The encoder design mirrors that of the spectrogram U-Net encoder, with the main difference being the use of 1D convolutions instead of 2D convolutions.

Latent Highlighting Module. With both the magnitude and waveform embeddings in the same shape, we add them element-wise to create a unified audio embedding. An additional encoder layer then reduces the temporal dimension by half, producing $\mathbf{f_a} \in \mathbb{R}^{C_a \times L}$, where *L* represents the temporal dimension of the latent audio features. To transform $\mathbf{f_a}$ into highlighted audio representations, we design a latent highlighting module that incorporates both the audio features and contextual information **c** (such as the encoded features of video streams or other multi-modal input) to output the features representing the highlighted audio, denoted as:

$$\hat{\mathbf{f}}_{\mathbf{a}} = \mathcal{H}_{\theta}(\mathbf{f}_{\mathbf{a}}, \mathbf{c}),$$
 (2)

where θ is the model parameters. Since both the latent audio features and the contextual input are temporal signals, we utilize a transformer-based framework to process them effectively, as depicted in Fig. 2(b).

Waveform/Spectrogram U-Net Decoder. The refined features \hat{f}_a will first pass through an additional decoder layer to double the temporal length. Next, the output features serve as the input to both the waveform and spectrogram U-Net decoders, each mirroring the structure of the corresponding encoder. The spectrogram decoder outputs a predicted ratio mask, denoted as M, which represents the highlighting information. We multiply the mask with the original magnitude spectrogram element-wise to obtain a refined magnitude, using the phase information from the input to reconstruct the output waveform. In addition, the audio output from the waveform decoder is then combined with the spectrogram-based waveform output, producing the final prediction \hat{s} .

3.2.2 Latent Highlighting Transformer

We now discuss the design of latent highlighting module \mathcal{H}_{θ} introduced in Eq. (2).

Latent audio features $\mathbf{f_a}$ from the audio backbone capture temporal and semantic characteristics of the poorly mixed audio. To transform these features into representations that convey appropriate highlighting effects, we consider two key insights: (i) Audio captures information from the entire surrounding environment, while the visual field-of-view is narrower, focusing on salient regions and content. This necessitates leveraging the temporal dynamics of the visual context as guidance for acoustic highlighting. (ii) In movies, complex interactions often occur between different sources (such as speech, music, and sound effects), with music saliency, in particular, driven by emotional cues. Relying solely on visual signals may not fully convey these nuanced relationships. This prompts the question: can additional modality enhance this process? To study this, we design a transformer-based latent highlighting module \mathcal{H}_{θ} that can flexibly incorporate various types of temporal context, such as video streams or text captions.

Context Encoding. Given the video sequence \mathbf{v} , we use CLIP ViT-L/14 [47] to transform each frame into a feature vector, denoted as $\mathbf{f}_{\text{vid}} \in \mathbb{R}^{C_{\text{vid}} \times T_v}$. To address the second insight mentioned above, we incorporate text captions as an additional modality. Vision Language Models (VLMs) have demonstrated impressive capabilities in summarizing images and reasoning about text. We leverage text captions as a bridge to convey deeper sentiment and context beyond raw visual features. To generate captions automatically for each frame, we use InternVL2-8B [6]. Each caption is then embedded using T5-XXL encoder [48], resulting in textual embeddings denoted as $\mathbf{f}_{text} \in \mathbb{R}^{C_{text} \times T_v}$. Since raw frame and text features are extracted at a per-frame level and lack temporal interaction, we apply a transformer encoder for each modality to capture the temporal context, denoted as \mathcal{E}_{vid} and \mathcal{E}_{text} . To preserve temporal order, we add sinusoidal positional encoding [60] to the input of each transformer encoder layer. We can concisely define contextual information encoding as a sequence of the following operations:

$$\hat{\mathbf{f}}_i = \mathcal{E}_i(\mathcal{G}_i(\mathbf{f}_i)),\tag{3}$$

where $i \in \{\text{video, text}\}, \mathcal{G}_i(\cdot)$ is a linear projection layer to project C_i to C, the same channel dimension as \mathbf{f}_a .

Context-aware Acoustic Highlighting. We use a transformer decoder, \mathcal{D} , to generate highlighted acoustic representations. Sinusoidal positional encoding is added to $\mathbf{f_a}$. The transformer decoder \mathcal{D} consists of multiple layers, each containing a self-attention layer, a cross-attention layer, and a feed-forward layer. Rather than directly treating the decoder's output as the final prediction, we interpret it as an offset to the original features, adding it back to $\mathbf{f_a}$ to preserve the audio's semantic content while adjusting inter-source differences. Additionally, we incorporate a zero-initialized convolution layer [67], denoted $\mathcal{Z}(\cdot)$. This layer is a 1×1 convolution with both weights and biases initialized to zero. The overall process is described as:

$$\hat{\mathbf{f}}_{\mathbf{a}} = \mathbf{f}_{\mathbf{a}} + \mathcal{Z}(\mathcal{D}(\mathbf{f}_{\mathbf{a}}, \hat{\mathbf{f}}_i)), \tag{4}$$

where the context $\hat{\mathbf{f}}_i$ can include visual, textual, or both types of contextual information. In relation to Eq. (2), \mathcal{H}_{θ} acts as the integrative component that connects \mathcal{D} , \mathcal{E}_{vid} , and potentially \mathcal{E}_{text} .

3.2.3 Training and Inference

Our VisAH framework takes the input audio a along with visual context v or its textual captions to predict the highlighted audio \hat{s} . The loss is computed at the waveform level and backpropagated through the network. In this work, we use a multiscale STFT (MR-STFT) loss [65] between the



Figure 3. We generate poorly mixed audio from the well-mixed movie audio through the following steps: 1) **Separation**: We separate the ground truth movie audio into individual tracks for speech, music, and sound effects, allowing for some imperfections in the separation process; 2) **Adjustment**: For each separated track, we apply either suppression or emphasis, with the intensity selected from three levels: [high, moderate, low]; 3) **Remixing**: Finally, we combine the adjusted tracks through simple addition to create the poorly mixed input audio.

predicted audio \hat{s} and the ground truth audio s, which is implemented by computing the ℓ_1 distance between their amplitude spectrograms:

$$\mathcal{L} = \text{MR-STFT}(\hat{\mathbf{s}}, \mathbf{s}). \tag{5}$$

The window sizes are set to 2048, 1024, and 512. It is worth noting that the training loss is intentionally simple, and any arbitrary waveform or spectrogram loss could be applied. We demonstrate that even a standard loss can effectively drive training, leaving further exploration of loss design to future work.

At test-time, given badly mixed input audio and the associated video frames as context, VisAH outputs wellhighlighted audio that is coherent both temporally and semantically with the provided visual guidance.

4. THE MUDDY MIX DATASET

It is worth reiterating that training our model requires access to badly mixed input audio, well-highlighted output audio and the associated visual frames, as shown in Eq. (1). Our key observation is that movies serve as a reliable source of well-mixed data, implicitly conveying what good highlighting and its synchronization with video sounds like. Our final requirement of having access to the corresponding badly mixed input audio is satisfied through the data modification process described in this section. It essentially involves demixing, adjusting and then remixing movie audio so as to disturb its original highlighting effect. We select the CMD [2] as our data source, which includes 33,976 clips from 3,605 diverse movies spanning various genres¹, countries, and decades, covering salient parts of each film. Each clip is approximately two minutes long. We concentrate on films tagged in the "Action" category, ensuring good presence of multiple acoustic sources beyond speech and music. As a movie in CMD may belong to multiple categories, our selection still remains diverse and covers a range of genres. We leave further expansion of the dataset to future work. To prepare the data for training and evaluation, we segment each movie clip into 10-second segments and extract the video stream at 1 fps using ffmpeg, while filtering out segments that lack an audio stream.

Separate, Adjust, and Remix. Given a high-quality movie audio s, we prepare the poorly mixed input a through a three-step process as shown in Fig. 3:

- 1. Separation. In practice, audio may consist of an infinite variety of sources, making it impractical to separate and remix every possible source individually. We follow the Cinematic Sound Demixing Challenge [8], which segments audio into three broad categories: *speech, music*, and *sound effects*. Accordingly, we apply a three-stem separation model trained for cinematic audio source separation on the DnR v3 dataset [63], to decompose s into three substreams: \hat{s}_h (speech), \hat{s}_m (music), and \hat{s}_e (sound effects). Additionally, we calculate any residual component, \hat{s}_r , to ensure that $s = \hat{s}_h + \hat{s}_m + \hat{s}_e + \hat{s}_r$. This formulation guarantees that even if the separation is imperfect, the sum of all components matches the original audio track.
- 2. Adjustment. Using these imperfect separations, we alter their original relative levels, creating an input audio signal that intentionally mismatches the video's highlighting effect. Specifically, we adjust the relative loudness of each stream. We first measure the original loudness of each separated source using the *pyloudnorm* library [53]. For the source with the highest loudness, we apply a "Suppress" action, reducing its loudness by a randomly selected strength from the categories [high, moderate, low]. For the other two sources, we apply a "Highlight" action, increasing their loudness by a value chosen from [high, moderate, low]. To retain the original mixture's content, we use the combined track $\mathbf{\hat{s}_e} + \mathbf{\hat{s}_r}$ for the sound effects input signal. We implement the loudness adjustments as follows: [high, moderate, low] for highlighting corresponds to increases of $\{12, 9, 6\}$ dB, while for suppressing, we apply decreases of $\{-12, -9, -6\}$ dB.
- 3. **Remixing.** After adjusting loudness, we remix the three sources linearly to create a poorly mixed input that contrasts with the ground truth highlighting effect.

¹Some videos are unavailable due to inactive YouTube links

Table 1. Main comparison: The best results are highlighted in **bold**, while the second best are highlighted with <u>underline</u>. We report metrics on waveform distance, semantic alignment, and time alignment. All results are multiplied by 100.

Method	$\mathrm{MAG}\downarrow$	ENV↓	KLD↓	$\Delta \mathrm{IB} \downarrow$	W-dis↓
Poorly Mixed Input	22.69	6.30	20.61	1.52	1.94
DnRv3 [63]+CDX [8]	26.32(-16%)	7.62(-21%)	15.87 (+23%)	1.78(-17%)	2.84(-46%)
Learn2Remix [66]	19.07 (+16%)	4.16(+34%)	$\overline{61.76(-199\%)}$	$\overline{8.27(-444\%)}$	1.20(+38%)
LCE-SepReformer [25]	17.18(+24%)	$\overline{4.28(+32\%)}$	30.99(-50%)	1.88(-24%)	$\overline{1.28(+34\%)}$
VisAH (Ours)	$\overline{10.08~(+\mathbf{56\%})}$	$3.43\ (+46\%)$	$11.01\;(+\mathbf{47\%})$	0.80~(+47%)	0.79~(+59%)

Following this procedure, we generate input audio for each video clip, resulting in 15,078/1,927/1,789 clips for train/validation/test sets, respectively.

5. Experiments

5.1. Experimental Setting

Implementation Details. In our experimental setup, the audio waveform is sampled at 44 kHz in stereo. We convert the input to mono by averaging the two stereo channels. Within the encoders, we set the dimensionality of the audio latent representation f_a to $C_a = 768$, with the original channel dimensions for visual and text features set to $C_{vid} = 768$ and $C_{text} = 4096$, respectively. During training, we use a batch size of 12 per GPU and the Adam optimizer with a learning rate of 0.0001. The model is trained for 200 epochs. All experiments are conducted on two RTX 4090 GPUs, with training taking approximately 18 hours to complete.

Evaluation Metrics. We employ the following groups of objective metrics to evaluate output quality:

- Waveform distance: The simplest way to assess the closeness of the prediction to the target is through waveform distance. We use magnitude distance (MAG) [64] to evaluate audio quality in the time-frequency domain and envelope distance (ENV) [31] to assess quality in the time domain.
- Semantic alignment: Since our goal is to adjust the relative distribution of audio across three categories: human speech, music, and sound effects. We apply KL divergence (KLD) [34, 61] using the pre-trained PaSST [27] model to compare the label distributions of the target and generated audio. Additionally, given that the video provides guidance, we assess audio-to-video semantic relevance using the ImageBind [14] model, calculated as the cosine similarity between audio and video embeddings, denoted as IB score. Since we have the target movie audio, we use the difference between the target and predicted IB scores:

$$\Delta IB = IB(\mathbf{v}, \mathbf{s}) - IB(\mathbf{v}, \mathbf{\hat{s}}).$$
(6)

• **Time alignment**: The relative variation between underlying sources (speech, music, and sound effects) can lead to significant timing differences, as each track follows its own temporal pattern. To test how well the model highlights all sources, we measure the minimum cost to align

Table 2. Ablation study on different context types. We compare a no-context baseline with models using semantic (single frame or text caption) and temporal context (multiple frames or captions).

Context	$\mathrm{MAG}\downarrow$	KLD↓	$\Delta \mathrm{IB} \downarrow$
No Context	10.35	11.95	0.99
+Semantic Vision +Semantic Text	$\begin{array}{c} 10.35\\ 10.32 \end{array}$	$11.67 \\ 11.83$	$\begin{array}{c} 0.91 \\ 0.84 \end{array}$
+Temporal Vision +Temporal Text	$\begin{array}{c} 10.24 \\ 10.08 \end{array}$	$\begin{array}{c} 11.18\\ 11.01 \end{array}$	$\begin{array}{c} 0.88\\ 0.80 \end{array}$

the predicted audio distribution with the target distribution. This is quantified using Wasserstein Distance $(W-dis)^2$.

5.2. Baselines

This is a novel task with no prior works. We adapt several methods for relevant and fair comparison with VisAH:

- **Poorly Mixed Input**: This is the manually created poorly mixed input according to our dataset creation strategy, serving as a reference point for comparison.
- **DnRv3** [63]+CDX [8]: To remix speech, music, and sound effects from the input and generate highlighted audio, we include an empirical baseline that adheres to the loudness distribution of these sources as specified by the CDX [8] challenge. We first apply the DnRv3 [63] separator to split the input audio into three tracks: speech, music, and sound effects. Next, we sample loudness values for each track according to their respective distributions. Finally, we adjust the loudness of each source and remix them to create the output audio.
- Learn2Remix [66]: Learn2Remix (L2R) utilizes ConvTasNet [37] as its backbone model to predict and remix different audio sources within feature spaces, making it well-suited for our task of adjusting and rebalancing the underlying speech, music, and sound effects. In our implementation, we adopt the more advanced SepReformer [52] model to replace the ConvTasNet backbone. We use the official code and train it on our dataset.
- Listen, Chat and Edit (LCE) [25]: LCE is a text-guided sound mixture editor capable of performing various audio editing tasks, such as adjusting the volume of specific

 $^{^{2}\}ensuremath{\mathsf{https:}}\xspace$ / /en.wikipedia.org/wiki/Wasserstein_metric.

Table 3. Ablation study on transformer encoders \mathcal{E}_{vid} and \mathcal{E}_{text} . V and T represent vision frames and text captions, respectively. Note that the text captions are obtained automatically from the video.

#layers	#params	context	$\mathrm{MAG}\downarrow$	KLD↓	W-dis↓
0	55.3M	V	10.36	10.91	0.83
3	61.6M	V	10.24	11.18	0.81
6	67.9M	V	10.69	12.42	0.83
0	55.4M	Т	10.66	12.53	0.85
3	61.7M	Т	10.34	11.75	0.81
6	68.0M	Т	10.08	11.01	0.79

sources based on text instructions. However, in our setup, we assume that explicit instructions on which sounds to highlight are unavailable and instead should be inferred from visual cues. To ensure a fair comparison, we provide text captions as guidance for LCE. Originally, LCE uses ConvTasNet [37] and Sepformer [55] as the SoundEditor models. For a fair comparison, we also replace the backbone with the SepReformer [52] model.

5.3. Quantitative Results

5.3.1 Comparison with Baselines

We compare our method with the baselines, and the results are presented in Tab. 1. The empirical baseline DnRv3+CDX performs worse than the input in waveform distance and time alignment metrics because it relies on a non-specific statistical distribution rather than data-dependent remixing. However, it outperforms the other two baselines in KLD, which we hypothesize is due to the versatile loudness distributions of the movies in CDX; remixing the speech, music, and sound effects at those levels shifts the distribution toward real movies. On the other hand, Learn2Remix, an audioonly baseline, struggles to enhance audio without guidance. It improves the poorly mixed input in terms of waveform distance, which occurs because it learns the global loudness distribution of our dataset. However, it fails to achieve the necessary semantic alignment, reinforcing the need for contextual input. For LCE, we utilize text captions as guidance for acoustic highlighting, resulting in better outcomes compared to Learn2Remix. However, it still underperforms significantly when compared to our approach. This is because it is not a method designed for acoustic highlighting and lacks the ability to capture the global trends required for the task. In contrast, VisAH demonstrates strong performance across all metrics, showcasing the effectiveness of our proposed framework.

5.3.2 Ablations

In this section, we review the design of context encoding and context choice, providing further insights into the task setup. **Does Contextual Information Matter?** In Tab. 2, we present a naive baseline that does not utilize any contex-

Table 4. Ablation study on dataset difficulty. We report the performance of Input (-I) and our Predictions (-P) at the three different levels of dataset difficulty.

Level	$\mathrm{MAG}\downarrow$	ENV↓	KLD↓	$\Delta \mathrm{IB} \downarrow$	W-dis↓
High-I High-P	$27.70 \\ 12.03$	$7.64 \\ 3.93$	$32.52 \\ 16.11$	$2.35 \\ 1.25$	$\begin{array}{c} 2.38 \\ 0.97 \end{array}$
Moderate-I Moderate-P	$22.70 \\ 8.73$	$6.25 \\ 3.23$	$20.59 \\ 11.08$	$\begin{array}{c} 1.50 \\ 0.81 \end{array}$	$\begin{array}{c} 1.92 \\ 0.65 \end{array}$
Low-I Low-P	$\begin{array}{c} 16.40\\ 9.55\end{array}$	$4.48 \\ 3.20$	9.89 7.16	$0.75 \\ 0.35$	$1.37 \\ 0.80$

tual guidance. This means that the model relies solely on the input audio to learn how to highlight relevant information. However, when we incorporate a single frame or its corresponding text caption, the semantic alignment metrics, KLD and Δ IB, show improvement. This indicates that context plays a crucial role in enhancing the model's performance. Since audio is a time sequence, the highlighting effects should ideally capture certain temporal patterns. We further conduct ablation experiments using the full length of video frames and captions, referring to this as temporal vision or text. The performance shows a significant boost, underscoring the importance of temporal context.

Number of Transformer Encoder Layers. We evaluate the impact of encoding temporal context by varying the number of transformer encoder layers. Without transformer encoders, video and text features are encoded frame by frame, which results in a lack of interaction across time steps. As shown in Table Tab. 3, we find that increasing the number of layers generally improves performance, suggesting that temporal context reasoning is essential for effectively understanding video-level content. Specifically, we observe continuous improvement when using text context, while performance with vision context initially improves but then deteriorates when the number of layers further increases, which we hypothesize that the CLIP vision features are already compact.

Analysis of Dataset Difficulty. In Sec. 4, we outline three levels of adjustments that can be made during the creation of the dataset, which are randomly selected. In Tab. 4, we provide an ablation of the impact of these difficulty levels. We create three test sets that consist solely of low, moderate, or high levels of adjustments. Our observations show continuous improvements in metrics as the difficulty of the dataset decreases. This supports both the design of our dataset and metrics, as well as the generalization capability of training on randomly selected levels.

5.4. Qualitative Analysis

Qualitative Visualizations. We display the magnitude spectrograms and waveforms of the highlighted audio produced by various methods, along with the input and ground truth in Fig. 4. These visualizations illustrate that our method ef-



Figure 4. We perform a qualitative comparison by visualizing the waveform and magnitude spectrograms of the highlighted audio results from different methods, along with the input and ground truth. Our method produces results that are closest to the movie GT. The orange box denotes suppressed snippets, and green box indicates highlighted snippets.



Figure 5. Subjective test: We ask users to rank the four methods based on audio-visual balance to evaluate acoustic highlighting.

fectively captures temporal variations and performs acoustic highlighting across speech, music, and sound effect sources.

Subjective Test. We conduct a subjective test to compare the highlighting results of our model with those of the LCE baseline, as well as the input and ground truth. Nine participants evaluated ten videos, each featuring four different audio tracks generated by various methods. They ranked the four methods based on perception of the balance between audio and visual quality. Our method achieves a top-2 ranking rate of 77%, outperforming the LCE baseline and the input by 63% and 62%, respectively, as shown in Fig. 5. Interestingly, our method even surpasses the GT for 34% of the videos, indicating strong highlighting performance, comparable to actual films at times.

Application: Refinement of Video-to-Audio Generation. Our VisAH has several potential downstream applications, one of which is refining video-to-audio generation. In Fig. 6, we demonstrate that by using the audio from MovieGen [45] as input, along with the video as guidance, our VisAH pro-



Figure 6. We demonstrate that our VisAH method enhances the quality of video-to-audio generation results.

duces audio that achieves a better IB score. This indicates enhanced audio-visual alignment, and human preferences confirm these improvements. We encourage the readers to see and listen to examples on our demo webpage in the attached supplementary materials.

6. Conclusion

We presented a new task—visually-guided acoustic highlighting—to bridge the gap between visual and acoustic saliency in video content. To address this task, we have proposed VisAH, a transformer-based multimodal framework that uses visual information to guide audio highlighting. By leveraging movies for free supervision, we develop a pseudo-data generation process that simulates real-world video quality, allowing for a labor-free training setup. Our evaluations show that our approach outperforms several baselines in both objective and human perceptual assessments. This framework enhances the alignment of audio-visual cues, offering a more cohesive viewing experience.

References

- Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *European Conference on Computer Vision* (ECCV), 2020. 2
- [2] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 2020. 2, 5
- [3] Changan Chen, Puyuan Peng, Ami Baid, Sherry Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. Action2sound: Ambient-aware generation of action sounds from egocentric videos. In ECCV, 2024. 2
- [4] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023. 3
- [5] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, 2017. 2
- [6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 4
- [7] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Work*shop on Music Source Separation, 2021. 2, 3
- [8] Giorgio Fabbro, Stefan Uhlich, Chieh-Hsin Lai, Woosung Choi, Marco A. Mart'inez-Ram'irez, Wei-Hsiang Liao, Igor Gadelha, Geraldo Ramos, Eddie Hsu, Hugo Rodrigues, Fabian-Robert Stoter, Alexandre D'efossez, Yi Luo, Jianwei Yu, Dipam Chakraborty, Sharada Prasanna Mohanty, Roman A. Solovyev, Alexander L. Stempkovskiy, Tatiana Habruseva, Nabarun Goswami, Tatsuya Harada, Minseok Kim, Jun Hyung Lee, Yuanliang Dong, Xinran Zhang, Jiafeng Liu, and Yuki Mitsufuji. The sound demixing challenge 2023 - music demixing track. *Trans. Int. Soc. Music. Inf. Retr.*, 7:63–84, 2023. 5, 6
- [9] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 324–333, 2019. 2
- [10] Ruohan Gao and Kristen Grauman. Visualvoice: Audiovisual speech separation with cross-modal consistency. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15490–15500. IEEE, 2021. 2
- [11] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision* (ECCV), pages 35–53, 2018. 2
- [12] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial visual representation learning through echolocation. In *European Conference* on Computer Vision (ECCV), 2020. 2
- [13] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometryaware multi-task learning for binaural audio generation from

video. In British Machine Vision Conference (BMVC), 2021. 2

- [14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180–15190, 2023. 6
- [15] Yury Gitman, Mikhail Erofeev, Dmitriy Vatolin, Bolshakov Andrey, and Fedorov Alexey. Semiautomatic visual-attention modeling and its application to video compression. In 2014 IEEE international conference on image processing (ICIP), pages 1105–1109. IEEE, 2014. 2
- [16] Jonathan Harel, Christof Koch, and Pietro Perona. Graphbased visual saliency. Advances in neural information processing systems, 19, 2006. 2
- [17] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33, 2020.
- [18] Chao Huang, Susan Liang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Davis: High-quality audio-visual separation with generative diffusion models. *arXiv preprint arXiv:2308.00122*, 2023. 2
- [19] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. arXiv preprint arXiv:2303.13471, 2023. 2
- [20] Chao Huang, Susan Liang, Yunlong Tang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Scaling concept with textguided diffusion models. arXiv preprint arXiv:2410.24151, 2024. 2
- [21] Chao Huang, Dejan Markovic, Chenliang Xu, and Alexander Richard. Modeling and driving human body soundfields through acoustic primitives. arXiv preprint arXiv:2407.13083, 2024. 2
- [22] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 2
- [23] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 3520– 3527. IEEE, 2021.
- [24] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *The European Conference on Computer Vision* (ECCV), 2018. 2
- [25] Xilin Jiang, Cong Han, Yinghao Aaron Li, and Nima Mesgarani. Listen, chat, and edit: Text-guided soundscape modification for enhanced auditory experience. *arXiv preprint arXiv:2402.03710*, 2024. 6
- [26] Junghyun Koo, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Stefan Uhlich, Kyogu Lee, and Yuki Mitsufuji. Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In *ICASSP 2023-2023 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023. 1, 2

- [27] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022, pages 2753–2757. ISCA, 2022.
- [28] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 2
- [29] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2
- [30] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. arXiv preprint arXiv:2309.15977, 2023. 2
- [31] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for realworld audio-visual scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 6
- [32] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Language-guided joint audio-visual editing via one-shot adaptation. In *Proceedings of the Asian Conference* on Computer Vision, pages 1011–1027, 2024. 2
- [33] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 2
- [34] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503, 2023. 6
- [35] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3042–3051, 2022. 2
- [36] Chien-Yu Lu, Min-Xin Xue, Chia-Che Chang, Che-Rung Lee, and Li Su. Play as you like: Timbre-enhanced multi-modal music style transfer. In *Proceedings of the aaai conference* on artificial intelligence, pages 1061–1068, 2019. 1
- [37] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019. 3, 6, 7
- [38] Marco A Martínez Ramírez, Emmanouil Benetos, and Joshua D Reiss. Deep learning for black-box modeling of audio effects. *Applied Sciences*, 10(2):638, 2020. 2
- [39] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 2
- [40] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for

moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 2

- [41] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. Advances in neural information processing systems, 2018. 2
- [42] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648, 2018. 2
- [43] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 2
- [44] Julian D Parker, Sebastian J Schlecht, Rudolf Rabenstein, and Maximilian Schäfer. Physical modeling using recurrent neural networks with fast convolutional layers. arXiv preprint arXiv:2204.10125, 2022. 2
- [45] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720, 2024. 8
- [46] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, pages 292–308. Springer, 2020. 2
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4
- [48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 4
- [49] Marco A Martínez Ramírez and Joshua D Reiss. Modeling nonlinear audio effects with end-to-end deep neural networks. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 171–175. IEEE, 2019. 2
- [50] Anyi Rao, Xuekun Jiang, Yuwei Guo, Linning Xu, Lei Yang, Libiao Jin, Dahua Lin, and Bo Dai. Dynamic storyboard generation in an engine-based virtual environment for video production. In ACM SIGGRAPH 2023 Posters, pages 1–2. 2023. 1
- [51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015. 3
- [52] Ui-Hyeop Shin, Sangyoun Lee, Taehan Kim, and Hyung-Min Park. Separate and reconstruct: Asymmetric encoderdecoder for speech separation. In *The Thirty-eighth Annual*

Conference on Neural Information Processing Systems, 2024. 6, 7

- [53] Christian J. Steinmetz and Joshua D. Reiss. pyloudnorm: A simple yet flexible loudness meter in python. In *150th AES Convention*, 2021. 5
- [54] Christian J Steinmetz, Nicholas J Bryan, and Joshua D Reiss. Style transfer of audio effects with differentiable signal processing. arXiv preprint arXiv:2207.08759, 2022. 2
- [55] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 21–25. IEEE, 2021. 3, 7
- [56] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2
- [57] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 436–454. Springer, 2020.
- [58] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2745–2754, 2021. 2, 3
- [59] Soumya Sai Vanka, Christian Steinmetz, Jean-Baptiste Rolland, Joshua Reiss, and George Fazekas. Diff-mst: Differentiable mixing style transfer. arXiv preprint arXiv:2407.08889, 2024. 2
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 4
- [61] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. arXiv preprint arXiv:2312.15821, 2023. 6
- [62] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the* 29th International Conference on Intelligent User Interfaces, pages 699–714, 2024. 1
- [63] Karn N Warcharasupat, Chih-Wei Wu, and Iroro Orife. Remastering divide and remaster: A cinematic audio source separation dataset with multilingual support. In 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2), pages 1–10. IEEE, 2024. 5, 6
- [64] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 6
- [65] Ryuichi Yamamoto, Eunwoo Song, Min-Jae Hwang, and Jae-Min Kim. Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators. In ICASSP 2021-2021 IEEE International Conference

on Acoustics, Speech and Signal Processing (ICASSP), pages 6039–6043. IEEE, 2021. 4

- [66] Haici Yang, Shivani Firodiya, Nicholas J Bryan, and Minje Kim. Don't separate, learn to remix: End-to-end neural remixing with joint optimization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 116–120. IEEE, 2022. 1, 6
- [67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023. 4
- [68] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2, 3