Learning to Highlight Audio by Watching Movies

Supplementary Material

1. Project Page

We have created a project page (https://wikichao.github.io/VisAH/) to illustrate our method and showcase our results. **We strongly encourage readers to visit this webpage and use headphones**. 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. On the demo page, we show the following applications:

- **Comparisons to Other Methods.** We present examples from THE MUDDY MIX DATASET, showcasing the following: the input poorly mixed video (which is created through the process described in Sec.4, the highlighting results produced by LCE [1], the outputs from our VisAH model, and the original movie clips for comparison.
- **Video-to-Audio (V2A) Generation Refinement.** Generating audio from video has recently gained popularity due to impressive video generation results and the growing demand for an immersive audio-visual experience. Existing V2A models, such as Seeing-and-Hearing [3] and the more recent MovieGen [2], have demonstrated promising outcomes. However, these methods primarily focus on generating temporally aligned audio for videos, which can sometimes neglect the subtle differences between audio sources. Our approach, inspired by cinematic techniques, serves as a post-processing method to enhance audio quality in these cases.
- **Real Web Video Refinement.** Unlike movies, web videos are often recorded in less controlled environments, which can lead to undesirable effects. For example, viewers may experience an overpowering personal voice in egocentric videos or focus on distracting sound sources due to distance or background noise. In this context, we apply our model to web videos, aiming to deliver an improved cinematic-like audio-visual experience.

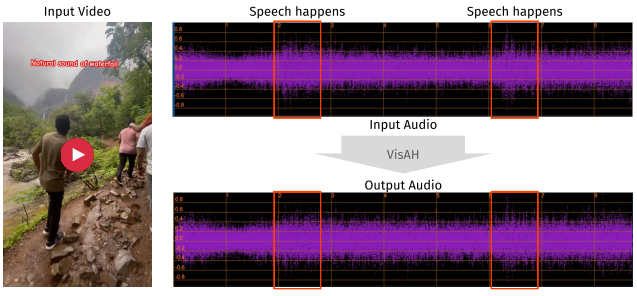

Figure 7. Failure case analysis: the sound effect (waterfall) overwhelms the speech.

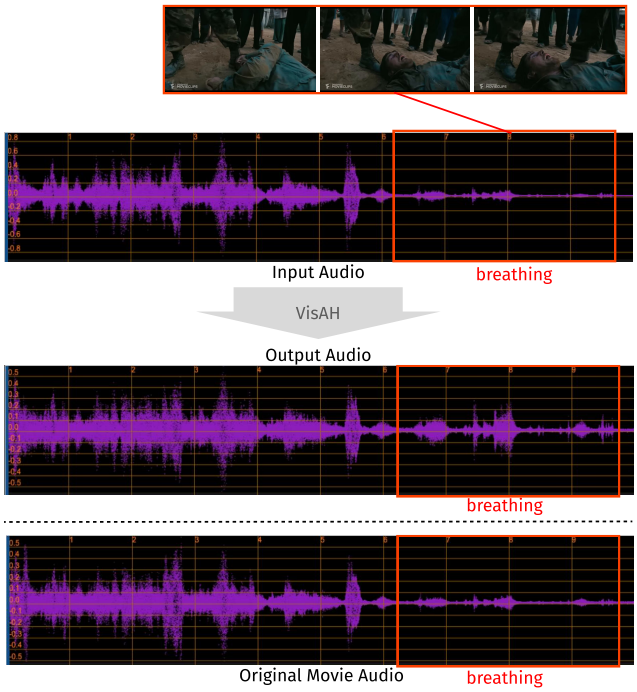

Figure 8. Failure cases analysis: Our method highlights the breathing sound based on the video context but diverges from the movie audio ground truth.

2. Failure Case Analysis

While our VisAH model is effective at highlighting audio guided by video content, there are scenarios where it might fail. Here, we provide case studies to illustrate the conditions under which such failures occur.

In Fig. 7, the video captures a natural waterfall scene with people hiking. The audio stream predominantly features the sound of the waterfall, with occasional moments of speech. Ideally, our VisAH model should balance these two audio sources to enhance the audio-visual experience. However, due to the overwhelming dominance of the waterfall sound, the speech becomes difficult to perceive. This results in the model failing to properly highlight the speech. As shown in Fig. 7, the input and output audio remain similar in this case, highlighting the challenge of separating and emphasizing speech under such conditions.

In Fig. 8, we present an example where our method fails to align perfectly with the original movie ground truth. Specifically, the breathing sound between 7 and 10 seconds is not emphasized in the movie's ground truth audio. However, the corresponding video frames during this period show close-up

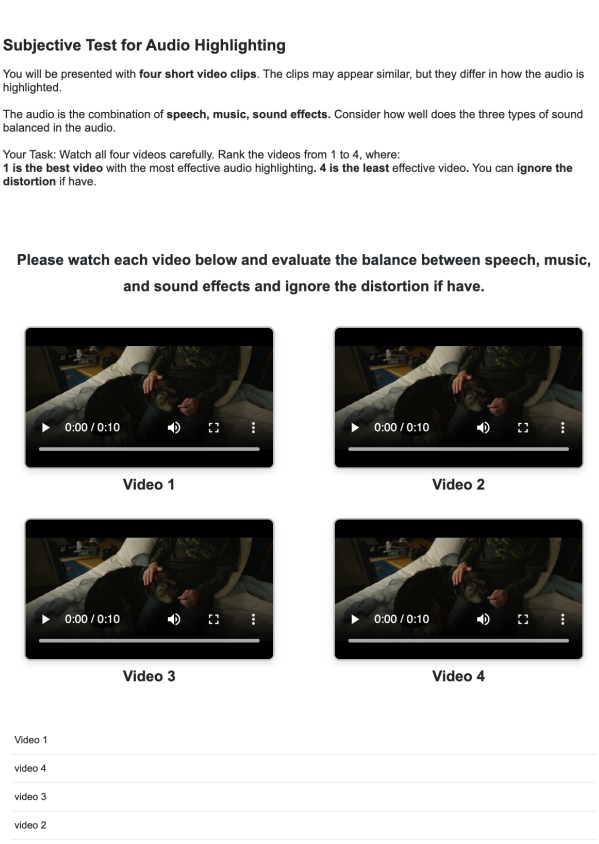

Figure 9. Screenshot of subjective test interface.

shots of a man's face, visually depicting the breathing action. Given these video conditions, our method predicts output audio that highlights the breathing sound, aligning with the visual context but diverging from the original movie audio. This failure highlights the need for a deeper understanding of movie content to achieve better alignment with the intended audio design.

3. Subjective Test Design

We illustrate the interface design of our subjective test in Fig. 9. The instructions emphasize that users should evaluate whether the speech, music, and sound effects in the videos are well-balanced and acoustically pleasing, and whether the audio aligns effectively with the video content.

Participants are shown four videos: the poorly mixed input, the best-performing baseline (LCE), our method, and the movie ground truth. After watching all the videos, users are asked to rank them from 1 to 4, with 1 being the most effective in audio highlighting and 4 being the least effective. The analysis of the ranking results is presented in Fig. 5.

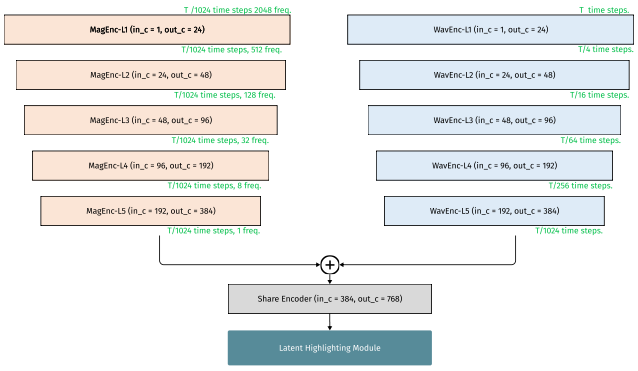

Figure 10. Design of magnitude and waveform encoders. Each encoder consists of five layers. The features from the waveform and magnitude encoders are combined through element-wise addition after the fifth layer, followed by an additional layer to encode the fused features.

4. Network Details

We detail the design of the magnitude and waveform encoders, along with their input and output dimensions. As illustrated in Fig. 10, each encoder consists of five layers, and the output shapes for both branches after the fifth layer are identical. At each layer, the output features are used for skip connections (not shown in the figure). This design facilitates straightforward element-wise addition of the two branches. The fused feature is then processed through a shared encoder layer before being passed to the latent highlighting module. Similarly, the magnitude and waveform decoders mirror the architectures of the encoders in reverse order.

5. Loss Function Details

Here, we give a more detailed illustration on the MR-STFT (Multi-Resolution Short-Time Fourier Transform) loss function used for training the model. The MR-STFT loss is implemented by computing the ℓ_1 distance between the amplitude spectrograms of the predicted signal $\hat{s}$ and the ground truth signal s. Mathematically, the loss function can be expressed as:

$$L_{\text{MR-STFT}}(\hat{s}, s) = \sum_{k=1}^{K} \left\| |STFT_k(\hat{s})| - |STFT_k(s)| \right\|_1 ,$$

where $STFT_k(\cdot)$ denotes the Short-Time Fourier Transform with the k-th window size, and $|\cdot|$ represents the magnitude of the spectrogram. The window sizes are set to 2048, 1024, and 512, corresponding to different resolutions of the spectrogram. This multi-resolution approach allows the loss function to capture both fine-grained and coarse-grained spectral details of the signals. 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, such as the MR-STFT loss, can effectively drive training and lead to high-quality results.

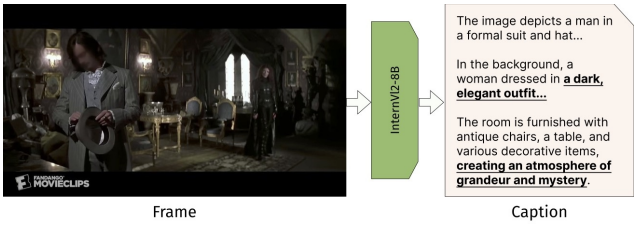

Figure 11. An example of a video frame and its generated caption.

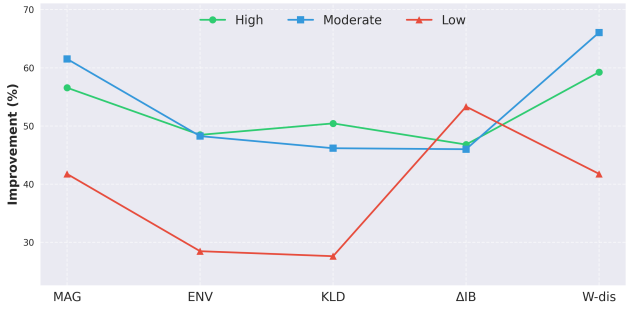

Figure 12. The improvement trend across the three difficulty levels is evaluated over five metrics.

6. Motivation for Text Condition.

Textual captions supplement video frames by leveraging strong reasoning capabilities of MLLMs. In Fig. 11, the caption generated by InternVL2-8B captures not only visual content, such as the appearance of individuals and room decorations, but also the scene's atmosphere, demonstrating the added semantic richness that textual information can provide. Moreover, it provides information more *explicitly* (e.g. "a dark, elegant outfit") than the visual encoder may extract. This supports the observation in **??** of why text conditioning outperforms visual signals. Regarding the performance metrics of the visual encoder in **??**, we hypothesize that CLIP vision features are more compact, and the 1fps video sampling rate drops motion information. Consequently, vision features are easier to overfit, as observed with the peak performance when the number of vision encoder layers is 3, and more encoder layers cause smoothing. To address this, we can try adopting a higher framerate (*e.g.*, 8fps) or exploring motion-aware architectures such as temporal transformers or 3D convolutions, which better model temporal dynamics while minimizing computational overhead. Learned downsampling can be another potential solution.

7. Inference Time Comparison

The inference times for VisAH, LCE, and L2R audio backbone are 0.028s, 0.017s, and 0.018s, respectively. While our method requires more time, it remains efficient for practical applications.

8. Analysis of Dataset Difficulty

We visualize the improvement trends in Fig. 12 across different levels of dataset difficulty, as discussed in Sec 5.3.2 and shown in Tab. 4. The magnitude of improvement is similar for the high and moderate difficulty levels, demonstrating that our method is robust in highlighting audio sources, even when they are highly suppressed. In contrast, the lower improvement observed for the low-difficulty level is attributed to the fact that the input audio is already relatively close to the ground truth and thus inherently conveys the ground truth highlighting effects to some extent. Consequently, the potential for improvement is reduced in this group.

9. Limitations and Future Works

Our method leverages versatile temporal conditions as guidance for audio highlighting, outperforming baseline methods and demonstrating applicability to real-world scenarios, including transferring knowledge from movies to daily and generated videos. However, there are areas where improvements can be made:

(i) Multimodal Condition Fusion. In our approach, we use either the video or its corresponding frame captions as guidance, achieving effective highlighting results. However, integrating these two modalities remains an open challenge. Text captions can infer the sentiment of the movie, complementing the video stream. Designing a more sophisticated strategy to fuse these modalities could enhance performance and remains an interesting direction for future research.

(ii) Dataset Generation Strategy. This paper introduces a three-step process for generating pseudo data through separation, adjustment, and remixing. While effective, each step can be further improved. For instance, employing multiple separators with varying granularity levels could offer greater flexibility and control. Additionally, replacing discrete loudness categories with continuous sampling could introduce more variability and challenge the model. Temporal loudness adjustments, such as varying the loudness at one-second intervals within a 10-second audio clip, could further enrich the dataset and present more complex training scenarios.

In summary, this work presents a novel task—visually guided acoustic highlighting—along with a versatile dataset generation process and a universal network. While our method demonstrates strong potential, many avenues for improvement remain, paving the way for future advancements in this area.

References

[1] 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. 1

[2] 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. 1

[3] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024. 1