TV-AUDIOREMOVER: JOINT TEXT-VISUAL GUIDED SOUND REMOVAL WITH MULTI-TASK HARD-MIXTURE CURRICULUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual object removal can eliminate a target from video frames, yet its acoustic trace persists in the soundtrack, causing obvious audio-visual inconsistency. Existing video inpainting models operate solely on pixels, while audio editing models, especially for the sound removal task, are typically driven by text and therefore rely on limited single-modal control, which is less effective than multimodal guidance that provides stronger semantic grounding and temporal synchronization cues. In this paper, we present Text-Visual Guided Sound Removal (**TV-AudioRemover**), a target sound removal framework that leverages natural-language instructions and optionally visually edited videos to suppress the sound corresponding to the removed visual object from the original audio mixture. To acquire high-quality training data, we devise a pipeline to construct a million-scale dataset of single-object audio-visual aligned samples, from which we synthesize mixture-target pairs customized for model training. To effectively leverage visual context and follow instruction intent, we augment the model architecture with task tokens, generalizable instruction modeling, and modality-specific global guidance. We further adopt multi-task training to strengthen task-role comprehension, and employ a hard-mixture curriculum that leverages semantically similar acoustic mixtures during fine-tuning to enhance fine-grained source discrimination. To support evaluation, we present **AV-Remove-Bench**, a comprehensive audio-visual object removal benchmark, along with dedicated objective metrics and an MLLM-based evaluation protocol. Experiments demonstrate that our method achieves state-of-the-art performance on both subjective and objective metrics across both audiovisual joint removal task and audio-only removal task. Demos are available at `https://tv-audioremover.netlify.app/`.

1 INTRODUCTION

Object removal has become a fundamental operation in modern video editing. Recent video erasers can remove a visible object and synthesize plausible pixels for the uncovered region (Miao et al., 2026; Fu et al., 2026; Liu et al., 2026; Hu et al., 2026). However, videos are not purely visual signals. A removed person may still speak, a deleted dog may still bark, and an erased vehicle may still leave engine noise in the soundtrack. A common workaround is to mute the entire audio track or regenerate a new soundtrack, but both choices are undesirable. Muting destroys useful background ambience, speech, or music, while full regeneration often changes timing, reverberation, and source identity. A practical system should instead perform selective acoustic erasure: remove only the sound associated with the visually removed target and preserve everything else.

There are two main families of models that perform audio editing: audio-visual editing models that jointly modify image and sound under shared control, and audio editing models that edit audio only while taking text, reference audio, or visual queries as multimodal control signals. Although both directions have made rapid progress, they still expose a structural gap for post-removal scenarios. For audio-visual editing models, existing strategies typically either use text instructions to synchronize the editing of visual frames and audio (Lin et al., 2026; Xu et al., 2025b; Liang et al., 2026), or first edit one modality and then use the edited result to guide the other modality (Zheng et al., 2026a; Ishii et al., 2025). However, most existing audio-visual editing models are designed for human-centric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

general editing tasks, making them better suited for human speech-related editing rather than general sound effect removal. For audio-only editing models, the visual signal is either underused—with the model relying mainly on text and the mixed audio (Wang et al., 2023; Ungersböck et al., 2026; Tao et al., 2025)—or the task is formulated as visual-query-based sound separation (Cheng et al., 2025; Takahashi et al., 2026; Shi et al., 2025), which does not transfer well to deletion-oriented settings.

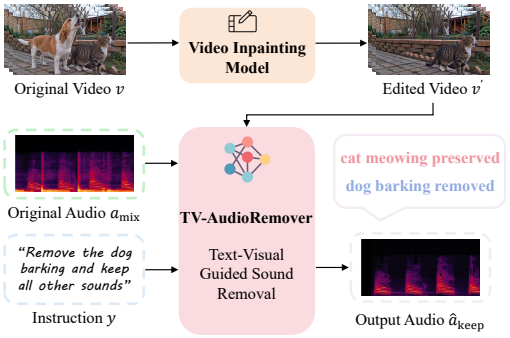

Figure 1: Overview of the Text-Visual Guided Sound Removal framework TV-AudioRemover.

We introduce **TV-AudioRemover**, a sound removal model with multimodal control tailored for audio-visual media requiring sound elimination. As illustrated in Figure 1, given the original mixed audio, a textual instruction, and an optional target-object-removed edited video, our model aims to generate the target audio that suppresses undesirable source sounds while fully preserving edit-irrelevant audio components. Faced with the problems of mapping acoustic components to visual entities and distinguishing mutually exclusive extraction and removal outputs in text-visual-controlled target sound removal, we make improvements in data construction, model architecture and training strategy. **In terms of data construction**, we first develop a dedicated pipeline to construct a million-scale, high-quality dataset of single-object audio-visual aligned samples. It combines MLLM recognition (Xu et al., 2025a), SAM Audio segmentation, and quality filtering based on CLAP (Wu et al., 2023) and SAJ scores (Wang et al., 2026). Afterwards, we perform audio mixing based on the curated dataset to synthesize mixture-target pairs tailored for the target sound removal task. **As for model architecture**, we introduce task tokens together with generalized instruction modeling to improve instruction following. To leverage sufficient visual information without conflicting with textual features, We employ modality-specific global guidance, where dedicated global conditions are applied to each modal branch. With the above design, the model determines preservable audio content corresponding to retained visual elements and identifies acoustic components to be suppressed for visually removed objects. Meanwhile, textual instructions remain focused on clarifying the editing intent. **Regarding the training strategy**, we jointly apply two optimization schemes. On one hand, multi-task supervised learning integrates extraction, removal, and joint editing tasks. It enhances the model's awareness of task boundaries and target sound source localization, enabling the model to differentiate audio preservation and suppression. On the other hand, we adopt a two-stage curriculum learning framework. The pre-training stage utilizes semantically distant mixtures for the model to preliminarily acquire audio elimination capability, while the fine-tuning stage leverages hard mixtures with similar semantic and acoustic features to boost fine-grained source discrimination. **For evaluation**, we construct AV-Remove-Bench, an audio-visual target removal benchmark with rich scene diversity and broad acoustic coverage. This benchmark consists of samples spanning public data, generated data and recorded real-world data, covering speech, music and sound effects. Furthermore, based on this dataset, we propose a comprehensive set of objective metrics as well as an MLLM-based evaluation protocol for the sound removal task. Finally, We compare our method against a variety of audio editing baselines. Experimental results demonstrate that TV-AudioRemover achieves state-of-the-art performance on both subjective and objective metrics across both audiovisual joint removal task and audio-only removal task.

2 Related Work

2.1 Audio Editing

Text-guided audio editing modifies an input waveform according to natural language. Training-free diffusion editing methods include inversion-based approaches such as ZETA and AudioEditor, which invert the input audio into the diffusion trajectory for prompt-based editing (Manor & Michaeli, 2024; Jia et al., 2025), as well as inversion-free approaches such as AudioMorphix and DirectAudioEdit (Liang et al., 2025; Ge et al., 2026). Meanwhile, unified audio models such as Audio-Omni, UNISON and AudioWeave support editing across multiple audio tasks (Tian et al., 2026; Li et al., 2026; Dong et al., 2026). These models take only text and mixed audio as inputs,

without leveraging any visual cues. Several models leverage visual conditions for audio manipulation, such as OmniSep, MMAudioSep and SAM Audio. Nevertheless, these are sound extraction models designed for source separation driven by multimodal queries, and they are not tailored for target removal. Specifically, most source separation systems are optimized for extraction tasks, where the query defines the sound source to retain. In contrast, sound removal tasks aim to exclude specified target sources. Although multimodal extraction models like OmniSep and MMAudioSep can take edited videos and negative text prompts for removal tasks, they face fundamental paradigm conflicts. These models inherently extract dominant sounds corresponding to visible video contents, while negative texts intend to eliminate target sounds and preserve all other audio components. This mismatch confuses model optimization, especially for complex multi-source background audio. Consequently, visual guidance becomes invalid, and these models essentially degenerate into pure text-guided removal paradigms.

2.2 Audio-Visual Editing

The first category of audio-visual editing methods modifies audio and visual content under shared instruction control. Language-guided AVEdit (Liang et al., 2024) uses text instructions to adapt paired image and audio events; Object-AVEdit (Fu et al., 2025) moves this idea toward object-level manipulation, while InstructAV2AV and JAVEdit further study instruction-following joint audio-video editing at broader task levels (Zheng et al., 2026b; Chen et al., 2026). These methods do not fully exploit multimodal interactions and perform less well on dedicated tasks such as audio-visual removal. Instead, they are more suited to target replacement — for example, swapping a cat for a dog while preserving the original object region and audio timing, where only the object category changes within a fixed spatial extent. Furthermore, most of these models are designed for human-centric editing tasks and lack training and understanding of general sound effects. They tend to fail when sound elements cannot be aligned with visual content. The second category edits one modality first and then uses the edited result to guide the other modality. AVI-Edit uses an audio agent and audio-synchronized cues to guide instance-level video editing with mask refinement. Although such cascaded architectures perform audio and video editing separately, each module is tightly coupled and cannot be replaced. Errors introduced by the first-stage audio agent propagate to subsequent stages, and this audio agent is not optimizable. In another case, CAVE leverages edited videos to guide the audio generation process, yet it merely supports limited audio style transfer and lacks native capability for visual-object-based sound removal. Unlike the aforementioned approaches, TV-AudioRemover is designed as a plug-and-play downstream audio remover that accepts outputs from arbitrary visual object removal methods. This decoupled formulation facilitates reusing powerful visual removal models, isolating failure sources, and optimizing the audio module alone. It is also agnostic to visual-only conditioning and supports arbitrary input combinations: text only, vision only, and joint text-vision. In addition, Open-source benchmarks for audio-visual removal remain scarce in terms of both data volume and category diversity. For instance, JAVEditBench proposed in JAVEdit only contains 25 audio-visual removal test samples, all of which are human-centric.

3 Method

As shown in Figure 1, given the original mixed audio $\mathbf{a}_{\mathrm{mix}}$, an instruction $\mathbf{y}$, and optionally edited video $\mathbf{v}'$ obtained via a video inpainting model, TV-AudioRemover generates $\hat{\mathbf{a}}_{\mathrm{keep}}$ that removes the target sound while preserving the remaining audio.

3.1 Data Construction

To acquire high-quality task-specific training data, we construct a million-scale dataset of single-object audio-visual aligned samples. The data processing pipeline is shown in Figure 2. The raw corpus combines audio-visual samples from video datasets, including VGGSound (Chen et al., 2020), AudioSet (Gemmeke et al., 2017), and ACAVCaps(Niu et al., 2026), with audio-only samples from WavCaps (Mei et al., 2024) and self-collected high-quality audio resources. Qwen3-Omni (Xu et al., 2025a) is used to annotate each candidate with structured information, including scene type, source multiplicity, sound category, and the main sounding-object label; for video candidates, it further checks whether the dominant sound source is visually present as the main object. Candidates that fail this audio-visual correspondence check are rejected, and the remaining aligned samples

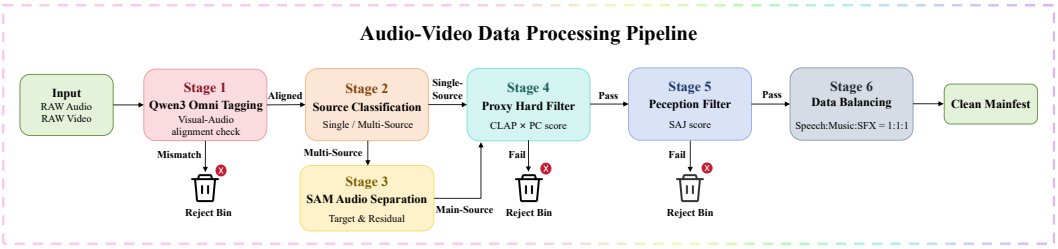

Figure 2: Data processing pipeline for constructing single-object audio-visual aligned samples.

are divided into single-source and multi-source cases. Single-source samples are directly sent to quality filtering, while multi-source samples are processed by SAM Audio to separate the source corresponding to the main visual object for audio-visual data, or the dominant sound source for audio-only data. Finally, we apply a proxy hard filter based on CLAP and PC scores (Tjandra et al., 2025), a perceptual filter based on SAJ, and category balancing over speech, music, and sound effects to obtain the clean manifest. The Qwen3-Omni tagging protocol and filtering thresholds are detailed in Appendix A.

Based on the curated single-object aligned samples, we synthesize task-specific mixture-target pairs for model training. For each clean sample A, we sample semantically different interference clips B and optionally C, apply random temporal cropping and signal-to-noise-ratio scaling to the interference sources, and mix them with A to form both two-source mixtures for basic role supervision and three-source mixtures for more complex multi-source scenes. This gives rise to three edit modes:

$$\begin{aligned}
&\text{extract:} \quad \mathbf{a}_{\text{mix}} = \mathbf{A} + \mathbf{B}(+\mathbf{C}), \quad \mathbf{a}_{\text{tar}} = \mathbf{A}, \\
&\text{delete:} \quad \mathbf{a}_{\text{mix}} = \mathbf{A} + \mathbf{B}(+\mathbf{C}), \quad \mathbf{a}_{\text{tar}} = \mathbf{A}(+\mathbf{C}), \\
&\text{both:} \quad \mathbf{a}_{\text{mix}} = \mathbf{A} + \mathbf{B}, \quad \mathbf{a}_{\text{tar}} = \mathbf{A}.
\end{aligned} \tag{1}$$

For audio-visual samples, the visual condition fed into the model is usually taken from the source sample A; in the three-source delete case, where the target output becomes $A + C$, the implementation masks the visual condition to avoid providing a cue that only corresponds to part of the preserved output. Audio-only samples use empty visual features by default. As illustrated in Figure 3, the *same* mixture serves as input for all three tasks; only the designated $\hat{a}_{\text{keep}}$ differs, allowing one synthesized sample to provide supervision for Extract, Delete, and Both simultaneously.

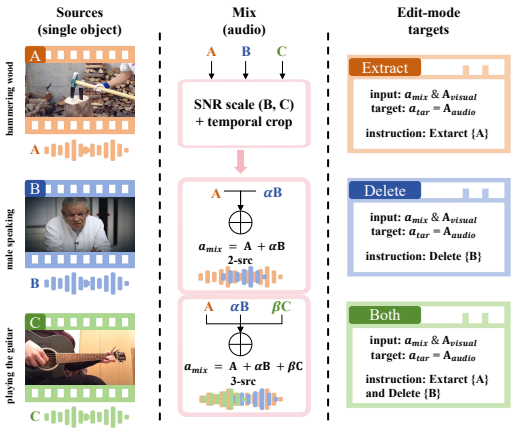

Figure 3: An example of synthesized training data, where the same mixture a_{mix} provides supervision for all three edit modes.

3.2 Model Architecture

TV-AudioRemover follows a latent flow-matching formulation, with the internal architecture illustrated in Figure 4. During training, we sample the target latent as $\mathbf{x}_1$ and the mixture latent as $\mathbf{x}_{\text{cond}}$, normalize both latents, and construct the noisy state $\mathbf{x}_t$ between Gaussian noise $\mathbf{x}_0$ and $\mathbf{x}_1$. The current noisy latent and the mixture latent are then concatenated along the channel dimension before audio input projection:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad \mathbf{h}_t = \text{Proj}_{\text{audio}}\big([\mathbf{x}_t, \mathbf{x}_{\text{cond}}]\big), \tag{2}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation. The denoising network takes $\mathbf{h}_t$ as the audio input and predicts the flow toward $\mathbf{x}_1$ conditioned on edited video, instruction, and $\mathbf{x}_{\text{cond}}$; at inference, $\mathbf{x}_{\text{cond}}$ is concatenated with the current denoising latent at each step.

In parallel, visual and textual features are projected into the Transformer hidden space and interact with the audio branch via MM-DiT blocks, where visual cues mark the sound sources to retain,

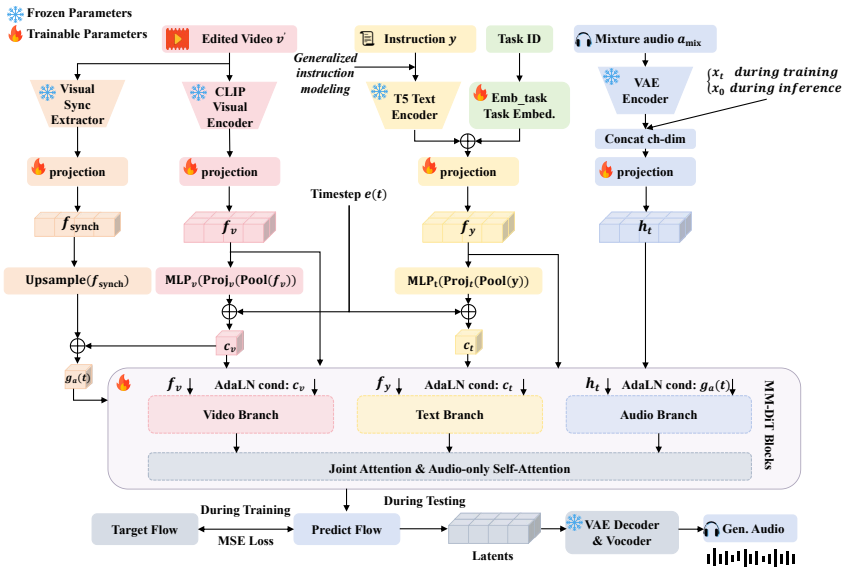

Figure 4: Architecture of the TV-AudioRemover multimodal diffusion Transformer.

textual instructions specify the intended operation, and the source mixture supplies acoustic materials for reconstruction. On top of this design, the architecture introduces three key modifications: (1) **Task tokens.** For the three edit modes $m \in \{\text{ext}, \text{del}, \text{both}\}$, we encode their task ids as $q_m \in \{0, 1, 2\}$ and map them to task embeddings $\mathbf{e}_m = \text{Emb}_{\text{task}}(q_m)$. The embedding is added to the encoded instruction features, i.e., $\mathbf{f}_y = E_{\text{txt}}(\mathbf{y}) + \mathbf{e}_m$, so that the instruction condition explicitly carries the designated task role. (2) **Generalized instruction modeling.** We diversify instructions across task types and language forms: extraction, deletion, and combined preserve-remove commands are each described by multiple natural-language templates. This helps the model learn complementary acoustic roles, improves robustness to linguistic variation, and reduces prompt-pattern overfitting. Representative examples and the complete template set are provided in Appendix B. (3) **Modality-specific global guidance.** Visual and textual features are projected into separate global conditions, and the audio branch is modulated by the visual condition together with diffusion-time and synchronization cues:

$$\begin{aligned} \mathbf{c}_v &= \mathbf{e}(t) + \text{MLP}_v\big(\text{Proj}_v\big(\text{Pool}(\mathbf{f}_v)\big)\big), \\ \mathbf{c}_t &= \mathbf{e}(t) + \text{MLP}_t\big(\text{Proj}_t\big(\text{Pool}(\mathbf{f}_y)\big)\big), \\ \mathbf{g}_a(t) &= \mathbf{c}_v + \text{Up}(\mathbf{f}_{\text{synch}}). \end{aligned} \tag{3}$$

where $\mathbf{e}(t)$ is the diffusion-time embedding, $\mathbf{f}_{\text{synch}}$ denotes synchronization-aware visual features, and $\text{Up}(\cdot)$ aligns them to the audio-latent length. The visual branch is conditioned on $\mathbf{c}_v$, the audio denoising branch is guided by $\mathbf{g}_a(t)$, and the instruction branch receives $\mathbf{c}_t$. Notably, the audio branch uses visual rather than textual global modulation, since the edited video provides a positive global cue about the desired visual-acoustic state, whereas text may specify either preservation or removal.

3.3 Training Strategy

The training strategy consists of two components: multi-task learning across complementary edit modes, and a two-stage hard-mixture curriculum built upon source pairs of two distinct difficulty levels. For multi-task learning, each batch randomly samples from source extraction, source deletion, and joint editing modes. The extraction task is not introduced mainly for standalone extraction, but to strengthen target-source localization, condition alignment between visual cues and acoustic sources, and the distinction between positive preservation and negative suppression. Training only with deletion samples and instructions can both overfit to specific prompt patterns and cause over-removal. For the two-stage hard-mixture curriculum, training proceeds from easy to hard mixtures. In the pre-training stage, the interference sources B and C mixed with source A are selected as

semantic-far samples, requiring different labels from A and T5-based label similarity below a preset threshold. During fine-tuning, we instead sample hard pairs, such as male versus female voices or timbre-similar instruments, to enhance fine-grained and attribute-level discrimination. Signal-to-noise ratios are also varied so that the model handles both dominant and subtle targets. The advantages of multi-task training and sample construction details are discussed in Appendix C.

We further use modality-wise classifier-free training with mutually exclusive dropout masks: both visual and text conditions are dropped with probability 0.05, only visual conditions with probability 0.15, and only text conditions with probability 0.05; when text is dropped, the task id is also masked. This enables a unified model to operate under visual-only, text-only, and joint visual-text conditioning. Given conditions $\mathcal{C} = \{\mathbf{f}_v, \mathbf{f}_{\text{synch}}, \mathbf{f}_y, \mathbf{x}_{\text{cond}}\}$, the network F_θ predicts the latent velocity field from noise $\mathbf{x}_0$ to target audio latent $\mathbf{x}_1$. The flow-matching objective is

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|F_\theta(\mathbf{x}_t, t, \mathcal{C}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2\right]. \tag{4}$$

At inference, classifier-free guidance compares conditional predictions with an unconditional branch that masks visual-text cues but keeps $\mathbf{x}_{\text{cond}}$, so guidance adjusts the editing intent without removing access to the source-audio reconstruction prior.

3.4 AV-REMOVE-BENCH

To support evaluation, we present AV-Remove-Bench, a comprehensive audio-visual object removal benchmark with 77 samples. All samples are audio-visual clips of about six seconds, covering speech, music, and sound effects from three sources: public datasets, including MUSIC-AVQA (Liu et al., 2024) for music, AVSpeech (Ephrat et al., 2018) for speech, and Condensed Movies (Bain et al., 2021), AVSBench (Zhou et al., 2022), and VGGSound-test for sound effects; Seedance 2.0 (Seedance et al., 2026) generated videos with prompts spanning the same three audio categories; and real recordings of sound-effect-centric daily scenes. Each sample falls into either a multi-sounding-object setting or a foreground-target-plus-stable-background setting, and the goal is to remove a specified sounding object from both the visual and audio streams. The benchmark design is detailed in Appendix D, and the data distribution is shown in Figure 5.

To thoroughly evaluate sound removal performance, we combine two proposed objective metrics, Target Source Suppression Ratio (TSSR) and Preserved Source Fidelity (PSF), with established evaluation metrics described in the experimental section, assessing both target-source suppression and non-target sound preservation.

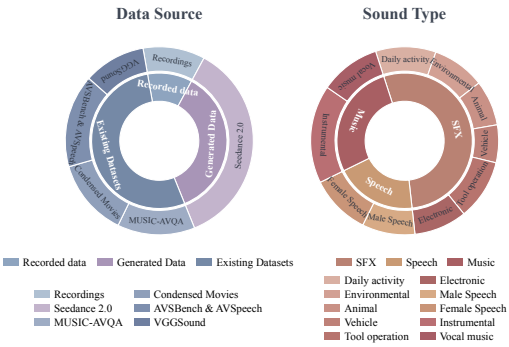

Figure 5: Data distribution of AV-Remove-Bench.

TSSR and PSF are computed using an AST audio classifier (Gong et al., 2021), which predicts event-level confidence scores for the queried source categories from the original and edited audio. Let $P_{A,\text{r}}$ and $P_{B,\text{r}}$ denote the confidence of the target-to-remove in the original audio A and edited audio B, and let $P_{A,\text{k}}$ and $P_{B,\text{k}}$ denote the confidence of the source to be preserved. TSSR measures the relative confidence drop of the removed source:

$$\text{TSSR} = 10 * \frac{P_{A,\text{r}} - P_{B,\text{r}}}{P_{A,\text{r}}}, \quad P_{A,\text{r}} > \epsilon. \tag{5}$$

When $P_{A,\text{r}} \leq \epsilon$, the original audio barely contains the target source, so TSSR is undefined. Larger TSSR indicates stronger suppression. For preservation, PSF penalizes confidence drops but allows a bounded bonus when denoising improves the preserved source:

$$\text{PSF} = \begin{cases} 10 - 10 * (P_{A,\text{k}} - P_{B,\text{k}}), & P_{B,\text{k}} < P_{A,\text{k}}, \\ 10 + 10 * \min(P_{B,\text{k}} - P_{A,\text{k}}, \gamma), & P_{B,\text{k}} \geq P_{A,\text{k}}, \end{cases} \tag{6}$$

where $\gamma = 0.5$ limits score drift. Higher PSF indicates better preservation of the non-target source.

In addition, we use Gemini 2.5 Pro (Comanici et al., 2025) as a multimodal judge to approximate human subjective perception. The detailed Gemini scoring criteria are provided in Appendix E.

Table 1: Objective evaluation on the AV-Remove Bench. Best results are shown in bold and second-best results are underlined. For ZETA, we adopt AudioLDM2-Large as its backbone.

Group	Method	Audio metrics						Audio-visual metrics		Efficiency	
		IS ↑	SAJ Overall ↑	TSSR ↑	PSF ↑	Instr. Comp.$_a$↑	Fidelity$_a$↑	IB-AV ↑	DeSync ↓	Params	Times (s) ↓
T-AV	AVI-Edit	3.16	2.24	2.60	9.58	3.56	2.57	<u>19.22</u>	0.77	7.24B	255.89
	InstructAV2AV	3.45	2.60	-2.88	9.86	2.65	3.04	14.81	0.76	15.63B	299.55
T-A	ZETA	3.27	<u>2.95</u>	<u>5.71</u>	<u>10.54</u>	3.69	3.22	18.71	<u>0.75</u>	1.5B	86.98
	Audio-Omni	2.15	2.20	4.96	9.01	<u>4.39</u>	1.68	8.23	1.13	3.05B	6.18
	UNISON	**3.63**	2.77	-2.39	9.99	4.27	3.44	18.00	0.78	731.71M	5.82
VT-A	SAM Audio	3.23	2.42	3.54	10.07	3.70	3.90	17.89	0.93	3.72B	1.94
	MMAudioSep	3.48	2.67	2.98	10.15	3.41	<u>3.94</u>	23.49	0.97	1.03B	2.03
VT-A (Ours)	**TV-AudioRemover**	<u>3.55</u>	**3.46**	**6.35**	**10.79**	**4.79**	**3.96**	**27.32**	**0.60**	1.04B	2.07

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details. Training uses the curated single-object aligned corpus described in the Data Construction section. Each training clip is 8 seconds long. The pre-training split contains about 1.11M general clips, including 602.9K audio-visual clips and 511.2K audio-only clips; the fine-tuning split contains 5.65K hard clips with an approximately 1:1 ratio of audio-visual to audio-only data. We synthesize mixture-target pairs online rather than storing fixed mixtures, and train extraction, deletion, and preserve-remove modes with probabilities 0.4, 0.4, and 0.2, respectively. We use a 44 kHz audio sampling rate and perform flow-matching inference with Euler sampling for 25 denoising steps. Optimization uses AdamW with weight decay 10^{-6} and gradient clipping at 1.0; pre-training runs for 350k iterations with learning rate 10^{-4} and step decay, and fine-tuning runs for 80k iterations from the pre-trained checkpoint with learning rate 2×10^{-6}.

Evaluation datasets and Baselines. The evaluation datasets consist of two parts: the removal subset (50 samples) from MMAE (Ma et al., 2026), an open-source text-guided audio editing benchmark, and our previously proposed text-visual guided sound removal benchmark AV-Remove-Bench. We compare against three groups of systems. The first group consists of the joint audio-visual edit models AVI-Edit and InstructAV2AV, used to evaluate sound removal under an end-to-end audio-visual setting. The second group consists of pure text-guided audio editing models that do not use visual conditions, including ZETA, Audio-Omni, and UNISON; together, they cover inversion-based, instruction-guided, and LLM-guided audio editing paradigms. The third group of baselines includes SAM Audio and MMAudioSep, both video-and-text-controlled audio extraction models that are closely related to our task. For evaluations of SAM Audio, we reformulate removal as extracting the sources to keep for multi-sounding-object cases; for foreground-target-plus-background cases, we reformulate it as extracting the source to remove and adopt the residual branch for output. For fair comparison, we set reranking-candidates to 1 during SAM-Audio inference. For MMAudioSep, we use negative prompts to guide sound removal. Furthermore, we perform comprehensive comparisons with the commercial audio-video editing model, including Seedance 2.0, Seedance 2.5 and MiniMax H3, and demonstrate the superiority of our method. Complete experimental results are available in Appendix F.

Metrics. For the removal subset of MMAE, we adopt its official evaluation script and report three metrics: Instruction Following, Consistency, and Exact Match. All metrics are reported on a 0–100 percentage scale. For AV-Remove-Bench, reference-based separation metrics like SDR are not applicable because no target-removed reference audio is available for evaluation samples. We therefore evaluate sound removal with both model-based metrics and MLLM-based judging. For model-based metrics, IS measures audio quality and diversity using a pre-trained audio classifier PANNs Kong et al. (2020), SAJ Overall measures perceptual separation quality, while IB-AV and DeSync measure audio-visual consistency, where IB-AV evaluates semantic audio-visual alignment with ImageBind Girdhar et al. (2023) and DeSync estimates temporal offset with Synchformer. TSSR and PSF are the two removal-specific metrics, together with MLLM-based metrics Instruction Compliance$_a$ and Fidelity$_a$, are proposed in the AV-Remove-Bench section. Detailed computation protocols are provided in Appendix G. We further conduct human listening evaluation on the AV-Remove-Bench. The scores are provided by 10 raters with professional audio-editing experience, and the model identities are blinded during scoring.

4.2 Objective Results

Table 2: Objective results on the MMAE removal subset.

Method	Instruction Following↑	Consistency↑	Exact Match↑
ZETA	58.99	42.70	0.00
MMEdit	51.70	**82.47**	10.42
Audio-Omni	48.51	56.49	2.08
UNISON	<u>72.19</u>	73.77	<u>20.83</u>
TV-AudioRemover	**76.77**	<u>82.25</u>	**22.92**

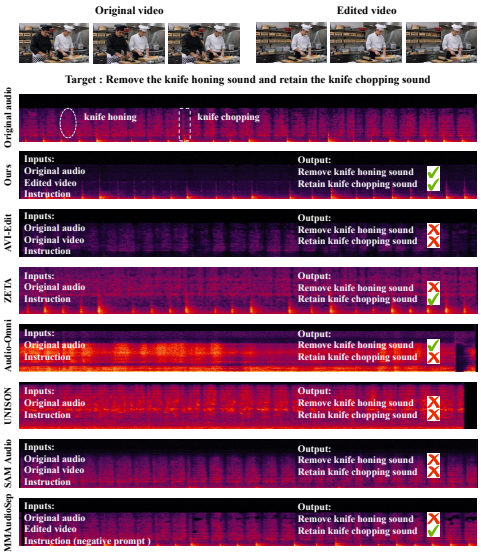

Figure 6: Qualitative comparison.

Table 2 demonstrates that our model achieves SOTA performance on the natural-language-controlled sound removal task. We do not report results of MMAudioSep and SAM-Audio on this dataset. MMAE tasks rely on natural-instruction guidance, where the residual components after sound removal are unknown or unspecifiable. This precludes negative prompts and conversion to an extraction task. The results in Table 1, our core table, are analyzed in detail below. For audio-editing-only methods, IB-AV and DeSync are computed by pairing their sound-removal outputs with SVOR's edited videos. SVOR, validated as the best-performing visual remover in Appendix H, provides visual input to our model and MMAudioSep. For PSF, we only compute the cases where the original audio contains two sounding objects, so that the source to be preserved can be described by a single label. Pure audio editing models show different strengths but remain unbalanced for removal. Although UNISON achieves the highest IS metric, which reflects general audio quality, it yields a low TSSR and PSF, indicating limited suppression of target sound sources and limited preservation of non-target sound. This phenomenon arises because the model retains a large portion of the original input mixture, leading to an inflated IS score. Audio-Omni obtains moderate instruction compliance, but its low Fidelity$_a$ reveals that the model incurs substantial damage to the preserved audio when performing target sound removal. SAM Audio achieves high audio fidelity and relatively competitive overall performance. However, comprehensive comparisons against our model reveal that an extraction-oriented model is not directly optimized for sound deletion and residual source preservation. Lastly, Joint audio-visual editing models including AVI-Edit and InstructAV2AV achieve poor performance on the sound removal task. TV-AudioRemover achieves the best TSSR, PSF, Instruction Compliance$_a$, and Fidelity$_a$, demonstrating stronger target suppression while preserving non-target sources. It also obtains the best IB-AV and DeSync scores, indicating that the edited audio is more semantically and temporally consistent with the edited video and demonstrating the advantages introduced by visual input for audio-visual consistency. The visualization results are shown in Figure 6.

4.3 Subjective Results

Table 3 presents human listening scores. Here, TRC, BP, TN, and OQ denote target removal completeness, background preservation, temporal naturalness, and overall quality, respectively. Raters score each output on a three-level scale $\{0, 0.5, 1\}$ along these dimensions. AC denotes the audio composite score, computed as $(0.35\times\text{TRC} + 0.35\times\text{BP} + 0.15\times\text{TN} + 0.15\times\text{OQ}) \times 100$, and ASR denotes the audio success rate, i.e., the fraction of samples whose four human scores are all non-zero. In the human evaluation, TV-AudioRemover ranks first on all reported metrics. Against SAM Audio, a strong video-text baseline, TV-AudioRemover gains 13.8% on TRC, 7.3% on BP, 6.0% on TN, 7.5% on OQ, 9.4% on AC, and 19.4% on ASR, highlighting the benefit of task-specific tuning for removal. Compared with strong text-guided audio editing baselines such as UNISON, TV-AudioRemover further improves all metrics, showing that visual conditioning enhances source identification and preserves scene-consistent non-target sounds.

Table 3: Human subjective scores.

Human subjective scores Method	TRC ↑	BP ↑	TN ↑	OQ ↑	AC ↑	ASR ↑
AVI-Edit	0.51	0.32	0.35	0.35	39.37	21.43%
InstructAV2AV	0.38	0.43	0.28	0.31	37.20	18.86%
ZETA	0.42	0.63	0.55	0.55	53.36	36.69%
Audio-Omni	0.48	0.24	0.20	0.19	30.75	11.69%
UNISON	0.65	0.60	0.60	0.62	62.00	49.35%
SAM Audio	0.86	0.83	0.89	0.88	85.72	81.58%
MMAudioSep	0.75	0.88	0.92	0.81	83.00	80.74%
TV-AudioRemover	**0.98**	**0.89**	**0.94**	**0.95**	**93.80**	**97.40%**

4.4 Ablation

Table 4: Ablation studies of TV-AudioRemover on the AV-Remove-Bench.

Setting	SAJ ↑	TSSR ↑	PSF ↑	IB-AV ↑	DeSync ↓
Training strategy					
Single-stage	3.22	0.48	1.05	25.17	0.62
Two-stage	**3.46**	**0.64**	**1.08**	**27.32**	**0.60**
Training task					
Single-task	3.11	0.52	1.04	26.90	0.69
Multi-task	**3.46**	**0.64**	**1.08**	**27.32**	**0.60**
Input modalities					
Text only	**3.46**	0.63	1.07	25.02	0.63
Visual only	2.97	-0.36	1.03	26.72	0.67
Visual + Text	**3.46**	**0.64**	**1.08**	**27.32**	**0.60**

Tables 4 summarizes the ablation results on three aspects. First, the training-strategy ablation shows that two-stage training is crucial for removal quality. Compared with single-stage training, the two-stage variant improves all metrics and is especially beneficial for target suppression. This is likely because pre-training builds general sound removal ability, while fine-tuning sharpens fine-grained source discrimination. Second, the multi-task training strategy outperforms single-task training across all evaluation metrics. Multi-task training achieves higher SAJ, TSSR, PSF and IB-AV scores while yielding a lower DeSync value, which demonstrates that joint optimization of multiple related tasks effectively boosts all performance metrics of TV-AudioRemover for sound removal tasks. Third, the input-modality ablation shows that text remains a strong control signal, while visual information provides complementary grounding. Text-only conditioning yields the same SAJ Overall score as the full model and only slightly lower TSSR and PSF, suggesting that linguistic instructions already specify most removal targets. However, when the model struggles to interpret the instructions, the visual stream supplies additional semantic and temporal cues, as evidenced by the higher IB-AV and lower DeSync under joint visual-text conditioning. Under visual-only conditioning, TSSR decreases substantially. The model still succeeds on a considerable subset of samples, indicating that visual information can guide source removal but cannot deliver precise editing without textual task intent. Overall, joint visual-text conditioning achieves the best performance across all metrics. All experiments above demonstrate that our multimodal-guided sound removal model, on the one hand, is capable of performing matched sound removal for edited videos to remedy modality deficiency. On the other hand, compared with existing audio editing models, our method achieves SOTA performance on text-only tasks. Incorporating the visual modality further boosts model performance and improves audio-video consistency, a capability that many existing audio editing models do not possess.

5 Conclusion

In this paper, we propose TV-AudioRemover, a target sound removal system tailored for audio-visual media requiring sound elimination. To support this task, we construct a million-scale corpus of single-object audio-visual aligned samples, and further introduce AV-Remove-Bench for evaluation. Built on a MM-DiT-based audio editing architecture with task tokens and global modal guidance, and trained with a multi-task two-stage curriculum, TV-AudioRemover better distinguishes preservable from suppressible audio under joint visual-text conditions. Experiments show that the proposed system achieves state-of-the-art performance on both objective and subjective metrics across audio-visual joint editing models and audio-editing models, demonstrating the effectiveness of our data construction, model design, and training strategy in target sound removal.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript polishing and to assist in coding our designed algorithms. We have not used them to formulate claims or proofs, nor to produce any experimental measurements or human evaluation data. We reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

Audio source removal technology enables the deletion of selected sound components from a mixed audio track. Misuse of this technique could erase speech or event sounds from recordings, fabricate misleading multimedia evidence, distort the facts of an event, and facilitate disinformation. Tampered audio may be difficult for humans to distinguish from authentic recordings. Therefore, any audio edited by such methods should be explicitly labeled as manipulated and cannot be presented as original authentic recordings. Our research only focuses on legitimate research scenarios, and we strictly follow the license agreements of all adopted datasets. We do not release raw media files that may involve personal identity information. Human evaluation in this work adopts a blind assessment paradigm conducted by independent raters. All evaluators are external participants who are not authors of this paper and have not been involved in any stage of system design, training, or experiments. No personal private information of the raters was collected throughout the study. To ensure research integrity and objectivity, all primary evaluation metrics were predefined before experiments. All experimental deviations and inherent limitations of human evaluation are fully reported in the appendix. This work uses the following datasets: VGGSound (CC BY 4.0), AudioSet (CC BY 4.0), AVSpeech (CC BY 4.0), Condensed Movies (CC BY 4.0), WavCaps (research-only), ACAVCaps (CC BY-NC 4.0), MUSIC-AVQA (CC BY-NC 4.0), and AVSBench (CC BY-NC 4.0). All datasets are utilized solely for academic research and non-commercial demonstration.

REPRODUCIBILITY STATEMENT

We will release part of the inference code, the evaluation metric computation scripts, and a subset of the AV-Remove-Bench evaluation dataset. These resources are available for research use under the license of the backbone model and the terms of the source datasets.

REFERENCES

M Bain, A Nagrani, A Brown, and A Zisserman. Condensed movies: story based retrieval with contextual embeddings. In *15th Asian Conference on Computer Vision, 2020*, Lecture Notes in Computer Science, pp. 460–479. Springer, 2021.

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

Yinan Chen, Chuming Lin, Zhennan Chen, Yuxiang Zeng, Junwei Zhu, Yali Bi, Xijie Huang, Chengming Xu, Donghao Luo, Zhucun Xue, et al. Javedit: Joint audio-visual instruction-guided video editing with agentic data curation. *arXiv preprint arXiv:2606.03168*, 2026.

Xize Cheng, Siqi Zheng, Minghui Fang, Ziang Zhang, Rongjie Huang, Shengpeng Ji, Jialong Zuo, Tao Jin, Zhou Zhao, et al. Omnisep: Unified omni-modality sound separation with query-mixup. In *International Conference on Learning Representations*, volume 2025, pp. 55196–55213, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Haocheng Dong, Yuheng Lu, Cheng Gong, Shansong Liu, Xiao-Lei Zhang, and Xuelong Li. Unified audio generation and editing via joint condition modeling and progressive training. *arXiv preprint arXiv:2606.16435*, 2026.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, 2018. ISSN 1557-7368. doi: 10.1145/3197517.3201357. URL http://dx.doi.org/10.1145/3197517.3201357.

Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. EffectErase: Joint video object removal and insertion for high-quality effect erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Youquan Fu, Ruiyang Si, Hongfa Wang, Dongzhan Zhou, Jiacheng Sun, Ping Luo, Di Hu, Hongyuan Zhang, and Xuelong Li. Object-avedit: An object-level audio-visual editing model. *arXiv preprint arXiv:2510.00050*, 2025.

Zhengkun Ge, Xiaoqian Liu, Haoran Zhang, Yuan Ge, Junxiang Zhang, Zhengtao Yu, Jingbo Zhu, and Tong Xiao. Directaudioedit: Inversion-free text-guided audio editing via diffusion prediction contrast. *arXiv preprint arXiv:2606.07356*, 2026.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audioset: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

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*, pp. 15180–15190, 2023.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021. URL https://arxiv.org/abs/2104.01778.

Jiagao Hu, Yuxuan Chen, Fuhao Li, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. From ideal to real: Stable video object removal under imperfect conditions. *arXiv preprint arXiv:2603.09283*, 2026.

Masato Ishii, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. Coherent audio-visual editing via conditional audio generation following video edits. *arXiv preprint arXiv:2512.07209*, 2025.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Zhaoqing Li, Haoning Xu, Jingran Su, Yaofang Liu, Zhefan Rao, Huimeng Wang, Jiajun Deng, Tianzi Wang, Zengrui Jin, Rui Liu, et al. Unison: A unified sound generation and editing framework via deep llm fusion. *arXiv preprint arXiv:2605.31530*, 2026.

Jinhua Liang, Yuanzhe Chen, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhuo Chen, Yuping Wang, and Yuxuan Wang. Audiomorphix: Training-free audio editing with diffusion probabilistic models. *arXiv preprint arXiv:2505.16076*, 2025.

Sen Liang, Cong Wang, Fengbin Guan, Zhentao Yu, Yiting Lu, Yuanzhi Wang, Yuan Zhou, Xin Li, and Zhibo Chen. Spongebob: Sync-aware harmonious audio-visual generative editing. *arXiv preprint arXiv:2605.25193*, 2026.

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*, pp. 1011–1027, 2024.

Yan-Bo Lin, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, Xiaofei Wang, Gedas Bertasius, and Lijuan Wang. Zero-shot audio-visual editing via cross-modal delta denoising. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7344–7354. IEEE, 2026.

Dingming Liu, Wenjing Wang, Chen Li, and Jing Lyu. From understanding to erasing: Towards complete and stable video object removal. *arXiv preprint arXiv:2604.01693*, 2026.

Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4478–4487, January 2024.

Ziyang Ma, Ruiqi Yan, Ruiyang Xu, Jie Fang, Zhikang Niu, Yi-Wen Chao, Wenming Tu, Tianrui Wang, Qi Chen, Wenxi Chen, et al. Mmae: A massive multitask audio editing benchmark. *arXiv preprint arXiv:2606.07229*, 2026.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Hantang Liu, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. Rose: Remove objects with side effects in videos. *Advances in Neural Information Processing Systems*, 38:149140–149162, 2026.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Junbo Zhang, and Jian Luan. Acavcaps: Enabling large-scale training for fine-grained and diverse audio understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15347–15351. IEEE, 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Akira Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. Mmaudiosep: Taming video-to-audio generative model towards video/text-queried sound separation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15667–15671. IEEE, 2026.

Ye Tao, Wen Wu, Chao Zhang, Mengyue Wu, Shuai Wang, and Xuenan Xu. Mmedit: A unified framework for multi-type audio editing via audio language model. *arXiv preprint arXiv:2512.20339*, 2025.

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lyu, Wei Xue, et al. Audio-omni: Extending multi-modal understanding to versatile audio generation and editing. *arXiv preprint arXiv:2604.10708*, 2026.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Michael Ungersböck, Florian Grötschla, Luca Lanzendörfer, June Young Yi, Changho Choi, and Roger Wattenhofer. Sao-instruct: Free-form audio editing using natural language instructions. *Advances in Neural Information Processing Systems*, 38:83411–83437, 2026.

Helin Wang, Bowen Shi, Andros Tjandra, John Hoffman, Yi-Chiao Wu, Apoorv Vyas, Najim Dehak, Ann Lee, and Wei-Ning Hsu. Sam audio judge: A unified multimodal framework for perceptual evaluation of audio separation, 2026.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025a.

Weihan Xu, Kan Jen Cheng, Koichi Saito, Muhammad Jehanzeb Mirza, Tingle Li, Yisi Liu, Alexander H Liu, Liming Wang, Masato Ishii, Takashi Shibuya, et al. Schrodinger audio-visual editor: Object-level audiovisual removal. *arXiv preprint arXiv:2512.12875*, 2025b.

Haojie Zheng, Shuchen Weng, Jingqi Liu, Siqi Yang, Boxin Shi, and Xinlong Wang. Avi-edit: Audio-sync video instance editing with granularity-aware mask refiner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23150–23160, 2026a.

Haojie Zheng, Yixin Yang, Siqi Yang, Shuchen Weng, and Boxin Shi. Instructav2av: Instruction-guided audio-video joint editing. *arXiv preprint arXiv:2605.18467*, 2026b.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2022.

A QWEN3-OMNI TAGGING PROTOCOL

We use Qwen3-Omni to obtain structured labels for raw audio-only and audio-visual candidates. The protocol is audio-first: the main sounding object is selected according to the audio evidence rather than the visible subject. Before producing the final label, the model is instructed to internally compare candidate sources by loudness dominance and clarity/intelligibility, and to choose a single dominant source.

For audio-only samples, Qwen3-Omni outputs six fields: scene type, whether multiple similar sounding objects exist, whether the clip is effectively single-source for the selected main object, the coarse sound category, the fine-grained category, and a concise main sounding-object prompt. The scene label is selected from {Indoor, Urban, Nature}. The multiple-source and single-source judgments are binary Yes/No labels. The coarse category is selected from {Human Voice, Music, SFX}, while the fine-grained category is selected from {Speech, Non-verbal, Crowd, Instrument, Score, Animal, Ambience, Mechanical, Foley}. The prompt is written as a short NP/VP-style phrase, such as "dog barking", "single adult male speaking", or "piano playing".

For audio-visual samples, Qwen3-Omni outputs the same audio-centric fields and additionally predicts whether the selected audio source is the main visible object and what the main visual object is. The visual correspondence label is set to Yes only when the dominant sounding object is clearly visible, visually salient, and plausibly responsible for the dominant sound. Otherwise, the sample is regarded as audio-visual mismatched and rejected from the aligned data pool. This rule prevents visually salient but acoustically irrelevant objects from being used as positive audio-visual pairs.

The output space is therefore fixed and auditable: binary fields use Yes/No labels, scene and sound types are selected from predefined category sets, and the sounding-object prompt is constrained to a concise separable NP/VP phrase.

After data annotation, the pipeline first filters out audio-video samples where the sounding object does not match the visual content. All audio-video and audio-only samples are then split into single-source and multi-source subsets according to the binary single-source label. Multi-source samples containing multiple similar sounding objects are discarded. For remaining multi-source samples, SAM Audio extracts target audio guided by the prompt of the main sounding object; these extracted audio samples are fed into the filtering module together with raw single-source samples. The first filtering stage removes silent segments via VAD detection. Next, CLAP and PC scores are jointly adopted to evaluate candidate samples in terms of semantic alignment and structural complexity. For native single-source samples, we retain samples satisfying either (PC ¡ 2.5 and CLAP ¿ 0) or (PC ¡ 4 and CLAP ¿ 0.35). For samples derived from multi-source data after target audio extraction by SAM Audio, we keep cases where the CLAP score between SAM-extracted target audio and the target label exceeds 0.25, while the CLAP score between residual audio and the target label is below 0. The second filtering stage leverages SAJ scores to screen samples based on audio quality. Specifically, only samples with an SAJ score higher than 3.5 are preserved. Through the above pipeline, around 50% of candidate samples are eliminated, yielding approximately one million high-quality single-target audio-video aligned samples.

B INSTRUCTION TEMPLATES

Training uses generalized instructions to cover both task-level and language-level diversity. At the task level, templates are organized into extraction, deletion, and combined preserve-remove modes, so that the model learns positive preservation, negative suppression, and their joint use in a single command. At the language level, each mode is expressed by multiple surface forms to reduce prompt overfitting and improve robustness to natural user instructions.

For extraction, the target source is denoted as x, and we use the following templates: "extract x", "isolate x", "separate x", "keep x", "keep only x", "only keep x", "I want to hear x", "I want only x", "give me x", "give me only x", "leave only x", "preserve x", "retain x", "the sound of x", "focus on x", "highlight x", "amplify x", "let me hear x", "please isolate x", and "please keep x".

For deletion, the removed source is denoted as y, and we use the following templates: "delete y and keep all other sounds", "remove y and keep all other sounds", "eliminate y and preserve the remaining audio", "exclude y while preserving everything else", "get rid of y but keep the other

sounds", "suppress y and keep the background intact", "mute y without changing the other sounds", "silence y and preserve all other audio", "drop y while keeping the remaining sounds", "discard y without removing anything else", "remove the sound of y and keep everything else", "I do not want y, but keep the rest", "do not include y, preserve all other sounds", "without y, keep the remaining audio unchanged", "no y, keep all other sounds", "everything except y", "anything but y", "all sounds except y", "please remove only y", and "kindly remove y and keep the rest".

For combined preserve-remove instructions, we first sample an extraction phrase $\mathrm{ext}(x)$ from the extraction templates and a removal phrase $\mathrm{rem}(y)$ from a deletion-oriented phrase pool, including "delete y", "remove y", "eliminate y", "exclude y", "get rid of y", "suppress y", "mute y", "silence y", "drop y", "discard y", "remove the sound of y", "I do not want y", "do not include y", "without y", "no y", "please remove y", and "kindly remove y". The two phrases are then composed with connective templates: "$\mathrm{ext}(x)$ and $\mathrm{rem}(y)$", "$\mathrm{ext}(x)$, and $\mathrm{rem}(y)$", "$\mathrm{ext}(x)$; $\mathrm{rem}(y)$", "$\mathrm{ext}(x)$ but $\mathrm{rem}(y)$", "$\mathrm{ext}(x)$ while $\mathrm{rem}(y)$", "$\mathrm{rem}(y)$ and $\mathrm{ext}(x)$", "$\mathrm{rem}(y)$, then $\mathrm{ext}(x)$", "$\mathrm{rem}(y)$ so I can hear x", and "$\mathrm{ext}(x)$ and at the same time $\mathrm{rem}(y)$". During training, surface perturbations such as capitalization, punctuation, and polite suffixes are randomly applied.

C Multi-Task Training and Mixture Construction

Multi-task training complements the deletion objective with extraction and joint editing samples. Extraction improves target-source localization and multimodal condition alignment, while joint editing exposes the model to positive preservation and negative suppression in the same instruction. This design reduces over-dependence on deletion-only prompts and helps prevent over-removal of acoustically similar non-target sources. Specifically, multi-task training yields several distinct benefits, as summarized below.

- **Enhanced target understanding.** The extraction task forces the model to learn the explicit correspondence between visual/textual conditions and sounding sources. With such awareness, the model can accurately determine which acoustic components should be removed during deletion-oriented editing.
- **Positive-negative contrast learning.** Joint optimization of two opposite task polarities enables the model to clearly distinguish between preservable and suppressible acoustic targets.
- **Improved visual condition utilization.** Since visual cues primarily describe objects to be retained, the extraction task strengthens the binding between visual objects and their corresponding audio components. This avoids the model over-reliance on textual conditions in pure deletion scenarios.
- **Alleviated over-removal artifacts.** Models trained solely on deletion tasks tend to indiscriminately suppress background sounds and acoustically similar interfering sources. Incorporating extraction supervision guides the model to produce structurally meaningful and source-aware outputs.
- **Stronger generalization capability.** Introducing extraction and composite preserve-remove tasks enriches the semantic coverage of training instructions, leading to better generalization toward diverse real-world user prompts.

Training samples for the two-stage curriculum are divided into easy and hard samples according to different training phases. In the pre-training stage, simple mixtures are sampled from general-mix datasets using semantic-far negative sampling: For an anchor source A, interference sources B and optional C must have labels different from A, and their T5-based label similarity to the selected sources must be lower than a preset threshold (e.g., 0.5). This generates semantically distant source pairs and provides unambiguous supervision for source discrimination. In the fine-tuning stage, difficult mixtures are sampled from hard-mix datasets. These hard-mix datasets contain groups of samples with fine-grained acoustically confusing sounds. The defined groups include human voice groups (*voice_age_emotion*, *voice_gender_speech*, *voice_gender_singing*) covering male/female speech, male/female singing, children crying, laughter, screams, crowd sounds and more; instrument groups (*instruments_strings*, *instruments_winds*, *instruments_percussion*) including strings, winds, and percussion; animal sound groups (*animals_mammals_pets*, *animals_birds*,

animals_insects_amphibians) consisting of mammals such as cats, dogs, cattle, sheep, horses, donkeys and wolves, birds like ducks, geese, crows and parrots, as well as insects and frogs; crowd and action groups; mechanical, traffic and tool groups; foley material groups; and natural ambience groups. During the fine-tuning stage, each training mixture is constructed by mixing distinct samples drawn from the same hard group, and mixtures are uniformly sampled across all hard groups.

D AV-REMOVE-BENCH DESIGN

AV-Remove-Bench is designed to evaluate whether an audio editing model can make the soundtrack consistent with a visually edited video. Each sample is an audio-visual clip of about six seconds and specifies one sounding object to be removed from both the visual and audio streams. The benchmark contains two scenario types: multi-sounding-object scenes, where a single clip contains two or more clearly distinguishable sounding objects and corresponding sources, and foreground-target-plus-stable-background scenes, where a foreground target sound is mixed with a stable background ambience.

The benchmark draws from three data sources. Public-dataset samples cover music, speech, and sound effects: Music-AVQA and Music-Duets are used for music cases, AVSpeech for speech cases, and Condensed Movies, AVSBench, and VGGSound-test for sound-effect cases. Generated samples are produced by Seedance 2.0 from manually designed prompts. These prompts are organized to cover speech, music, and sound effects, with sound effects further spanning tool actions, natural ambience, electronic and mechanical sounds, vehicles, and animals. The prompt set also explicitly balances the two scenario types above, so that generated clips include both multiple foreground sounding objects and foreground objects accompanied by stable background sound. Real recordings are collected from daily real-world scenes and mainly focus on sound-effect-centric object removal, such as vehicles and environmental sounds.

For public datasets, we manually select 42 high-quality samples from the aforementioned sources. For synthetic samples, we adopt the top 27 generated instances. For real recordings, we carefully collect samples covering eight distinct real-life scenarios with clear documentation. Furthermore, an example prompt for Seedance 2.0 video generation is provided below:

```
Two people working on a backyard fence --- one hammering nails with a
    hammer, the other sawing a board with a hand saw --- both working
    simultaneously. Landscape orientation (16:9 horizontal frame). Fixed
     camera, side view capturing both workers beside the fence. AUDIO
    REQUIREMENTS: Two clearly distinguishable tool sounds emitted by two
     different visible people working concurrently: (1) sharp metallic
    clang produced by the hammer striking nails into wood; (2) rhythmic
    back-and-forth rasping sound as the saw cuts through the board. The
    two sounds overlap continuously, and each sound is synchronized with
     the visible motions of the corresponding worker. Fixed camera, clip
     duration of 6 seconds. No voiceover, narration, or off-screen audio
    . High signal-to-noise ratio, clear and crisp audio, high-definition
     video.
```

Finally, Figure 7 shows several data examples from AV-Remove-Bench.

E GEMINI-BASED SCORING CRITERIA

We use Gemini as a multimodal judge to provide subjective-style scores for the edited results. Each evaluation item contains the original video, the edited video, the removal instruction, and, when available, the edit mask. The judge outputs four 1–5 scores: visual target removal (Instruction Compliance$_v$), video fidelity (Fidelity$_v$), audio target removal (Instruction Compliance$_a$), and preserved-audio fidelity (Fidelity$_a$).

For video evaluation, Gemini is instructed to watch both the original and edited videos and to focus only on visual content. Instruction Compliance$_v$ measures whether the target object or action specified by the instruction has disappeared from the edited video. The mask, when provided, is used to localize the edited region. The score is based on semantic removal: changing the target's



Figure 7: Data examples from AV-Remove-Bench.

appearance, color, texture, or blur level does not count as removal if the target identity remains recognizable. Fidelity$_v$ measures both the realism of the generated content inside the mask and the preservation of non-target content outside the mask; severe inpainting artifacts, boundary seams, blur, or damage to preserved objects lower this score.

For audio evaluation, Gemini compares the original and edited audio tracks. Instruction Compliance$_a$ only measures whether the target sound is removed, with the target status categorized as absent, faint, or clear. Fidelity$_a$ measures whether non-target sounds remain audible, natural, and close to the original, while also considering overall audio quality. Near-silent or severely distorted edited audio receives a low fidelity score even if the target sound is absent. The two audio scores are intentionally independent, so a result can achieve high target-removal compliance but low fidelity if it removes all sounds or introduces strong artifacts.

Table 5: Consistency Analysis Between MLLM Scores and Human Subjective Ratings

metric	Kendall ↑
Instruction Compliance$_v$	0.7494
Fidelity$_v$	0.7113
Instruction Compliance$_a$	0.7063
Fidelity$_a$	0.7092
Overall	0.7191

To verify the accuracy of scores generated by the MLLM, we conduct a consistency analysis between model predictions and human subjective ratings on AV-Remove-Bench, quantified via the Kendall rank correlation coefficient. The results are presented in Table 5. The Kendall rank correlation coefficient measures the agreement of sample ordering between two sets of scores and ranges from -1 to 1. Higher values indicate stronger consistency in ranking; zero signifies no correlation, while negative values imply opposite ordering. Our results yield Kendall coefficients of 0.7494, 0.7113, 0.7063 and 0.7092 for each metric, with an overall average of 0.7191. This demonstrates strong agreement between model and human evaluations, validating the reliability of the MLLM.

F Comparison with Commercial Models

Seedance 2.0, Seedance 2.5, and MiniMax H3 are commercial closed-source and open-source generative models that supporting joint audio-video editing. These models enable collaborative modification, regional inpainting and content completion for input videos and their audio tracks conditioned on text or multimodal prompts. We compare our method with these commercial models in terms of target sound removal on AV-Remove-Bench, and the corresponding results are summarized in Table 6.

Table 6: Objective Evaluation: TV-AudioRemover versus commercial models.

Method	Audio metrics						Audio-visual metrics	
	IS ↑	SAJ Overall ↑	TSSR ↑	PSF ↑	Instr. Comp.$_a$ ↑	Fidelity$_a$ ↑	IB-AV ↑	DeSync ↓
Seedance 2.0	3.36	2.50	-5.63	9.70	3.17	3.83	23.71	0.58
Seedance 2.5	2.85	2.93	3.63	10.59	4.36	4.11	**30.66**	**0.49**
MiniMax H3	3.30	2.25	0.76	9.97	3.41	**4.30**	23.52	0.64
TV-AudioRemover	**3.55**	**3.46**	**6.35**	**10.79**	**4.79**	3.96	27.32	0.60

The results demonstrate that our model achieves overall superior performance on the target sound removal task compared with these commercial models. Seedance 2.0 and MiniMax H3 exhibit limited instruction-following capability for audio editing, while Seedance 2.5 achieves improved audio-editing performance over Seedance 2.0, yet it tends to regenerate both video frames and audio tracks, which accounts for its strong audiovisual synchronization. An example prompt for audiovisual editing with Seedance 2.0, Seedance 2.5 and MiniMax H3 is provided below:

```
Perform integrated editing for both video frames and audio tracks
    synchronously:
Visual editing: Remove all subjects that produce meowing sounds (the
    meowing cat) from the frames. The barking dog, dog barks, all other
    sound sources and visual content should be fully preserved without
    any deletion, replacement or modification.
Audio editing: Completely eliminate all cat meows. The dog barks, the
    barking dog and all other sound sources are retained. The generated
    result requires natural audiovisual temporal alignment, free of
    audiovisual desynchronization and unnatural acoustic artifacts.
```

G Objective Metric Computation

We use two groups of objective metrics. The first group is model-based and measures acoustic quality, audio-visual consistency, target suppression, and preservation fidelity. The second group is MLLM-based and measures whether the edited audio follows the removal instruction while preserving non-target content.

For audio quality, IS is computed with a pre-trained PANNs audio classifier and summarizes both confidence and diversity of predicted audio classes. The classifier outputs probability distributions across audio categories for every edited audio, and the average category distribution over the full test set is estimated. We calculate the KL divergence between each sample's predicted distribution and the global distribution. The final IS score is acquired by averaging all KL divergence values and performing exponential transformation. SAJ Overall is produced by SAM Audio Judge; it aggregates faithfulness, recall, and precision, where faithfulness measures whether the preserved sound is undistorted, recall measures whether preserved sources remain complete, and precision measures whether unwanted source leakage is suppressed.

For audio-visual consistency, IB-AV is the cosine similarity between ImageBind audio and video embeddings, with higher values indicating stronger cross-modal alignment. DeSync is estimated by Synchformer and measures the predicted temporal offset between the edited audio and video; lower values indicate better synchronization. In Table 1, for audio-editing-only models, IB-AV and DeSync are computed after pairing each audio editing result with the best tested visual remover, SVOR, so that these metrics reflect audio-visual consistency under the same edited-video condition. The video-side evaluation used to choose SVOR is provided in Appendix H.

For removal-specific source behavior, Rel-TSSR and PSF are computed with an AST audio classifier. The corresponding calculation formulas are presented in the AV-Remove-Bench subsection of the main text.

For MLLM-based evaluation, Gemini receives the original video, edited video, removal instruction, and optional mask. Detailed definitions are provided in Appendix E.

H VISUAL REMOVAL EVALUATION

Table 7: Gemini-Based Evaluation of Visual Removal Quality on AV-Remove-Bench.

Method	Instr. Comp.$_v$ ↑	Fidelity$_v$ ↑
ROSE	4.86	3.30
AVI-Edit	2.68	2.48
EffectErase	4.74	3.33
UnderEraser	4.91	3.49
SVOR	**5.00**	**4.03**

Table 8: Subjective video-quality evaluation on AV-Remove-Bench.

Method	TRC ↑	BP ↑	BN ↑	CR ↑	TC ↑	AC ↑	VSR ↑
ROSE	0.53	0.49	0.38	0.44	0.42	46.95	42.21%
AVI-Edit	0.01	0.42	0.21	0.19	0.34	22.15	0.65%
EffectErase	0.64	0.47	0.50	0.46	0.50	52.72	41.56%
UnderEraser	0.78	0.69	0.68	0.64	0.59	70.03	68.83%
SVOR	**0.97**	**0.91**	**0.94**	**0.94**	**0.98**	**94.43**	**91.56%**

We select SVOR as the visual-removal precursor for our sound removal model according to the video-side evaluations in Tables 7 and Tables 8. These evaluations compare the open-source models AVI-Edit, ROSE, EffectErase, UnderEraser, and SVOR on the 77-sample AV-Remove-Bench. Gemini-based scores evaluate whether the edited video follows the removal instruction and preserves non-target visual content, while subjective scores measure target removal completeness (TRC), background preservation (BP), boundary naturalness (BN), generated-content realism (CR), and temporal consistency (TC). AC denotes the video composite score, computed as $(0.3\times\text{TRC} +0.3\times\text{BP} +0.15\times\text{BN} +0.15\times\text{CR}+0.1\times\text{TC}) \times 100$, and VSR denotes the video success rate, i.e., the fraction of samples whose four human scores are all non-zero. SVOR obtains the best results in both evaluations, so it is used as the fixed edited-video input when computing IB-AV and DeSync for audio-editing-only methods.

I SUBJECTIVE EVALUATION STATISTICS

For human listening, 10 raters with professional audio-editing experience independently score anonymized model outputs under the same sample, with model identities blinded during annotation. Each sample is rated on four dimensions using scores in $\{0, 0.5, 1\}$: target removal completeness (TRC), background preservation (BP), temporal naturalness (TN), and overall quality (OQ). TRC measures whether the target sound has been removed, BP measures whether non-target sounds are preserved, TN measures whether the edited audio sounds temporally smooth and natural, and OQ summarizes the overall perceptual quality of the result. A score of 1 indicates a clear success on the corresponding aspect, 0.5 indicates a partial or moderate success, and 0 indicates a clear failure.

For the GSB evaluation, annotators perform pairwise comparisons between TV-AudioRemover and all baseline methods across all samples on each evaluation dimension, and vote for three grades: Better, Same, or Worse. As illustrated in Table 9, the GSB pairwise comparison results reveal that TV-AudioRemover obtains considerably higher good rates than bad rates across all evaluation dimensions. Annotators tend to favor the outputs generated by TV-AudioRemover over baseline methods comprehensively, demonstrating the superiority of our approach from human subjective perspectives.

Table 9: GSB scores for TV-AudioRemover versus baselines.

GSB scores for TV-AudioRemover versus baselines			
Metric	Good rate	Same rate	Bad rate
TRC	0.92	0.06	0.03
BP	0.71	0.14	0.15
TN	0.86	0.10	0.04
OQ	0.67	0.21	0.13

Table 10: Friedman significance tests for subjective metrics on the dashboard. All reported metrics show statistically significant differences among models ($p < 0.05$).

Metric	p-value ↓	Kendall's W ↑	Significant
TRC	1.31×10^{-9}	0.154	Yes
BP	4.43×10^{-26}	0.406	Yes
TN	1.74×10^{-26}	0.412	Yes
OQ	6.86×10^{-26}	0.403	Yes

To verify whether the observed advantage that Model A achieves higher average subjective scores than Model B reflects genuine performance gaps instead of random sampling noise, we conduct statistical significance tests. The Friedman test is adopted to identify overall significant differences among multiple models for each metric, and Kendall's W is utilized to measure the magnitude of such differences. Interpretation of Friedman p-values: $p < 0.05$ rejects the null hypothesis, confirming significant differences among models; $p \geq 0.05$ fails to reject the null hypothesis, indicating no significant difference. Interpretation of Kendall's W: Values near 0 correspond to weak ranking consistency and minor differences between models; values close to 1 reflect stable rankings and stronger performance disparities. The corresponding results are presented in Table 10. The Friedman test yields $p < 0.05$ for all four subjective evaluation metrics, confirming statistically significant differences among compared models. Overall statistical results verify that the performance gaps observed in subjective scores are not caused by random noise.

J LIMITATIONS

The existing benchmark suffers from limited scale, especially for high-quality sound-effect samples. Larger public benchmarks paired with reference audios after target sound removal will facilitate standardized evaluation. In addition, highly entangled sound sources, such as overlapping speech from visually similar speakers, remain challenging. Such complicated cases demand more powerful modeling of speaker and object identity.