SPONGEBOB: SYNC-AWARE HARMONIOUS AUDIO-VISUAL GENERATIVE EDITING

Anonymous authors
Paper under double-blind review



Figure 1: **SpongeBob unifies visual editing and audio synthesis in a single pass, achieving frame-level synchronization while seamlessly preserving unedited context.**

ABSTRACT

Visual and acoustic events in the physical world are inherently coupled, yet existing video editing methods typically adopt decoupled pipelines, lacking bidirectional modality interaction. This results in two key limitations: (i) audio-visual desynchronization and (ii) contextual conflicts between generated audio and preserved content. To address these, we propose **SpongeBob**, an end-to-end audio-visual joint editing framework featuring bidirectional cross-modal interaction. For synchronization, a Sync-Aware Editing Mechanism aligns visual edits with sound events via bidirectional attention, temporal alignment, and spatial constraints. For contextual consistency, a Context-Aware Module leverages acoustic and visual context attention to prevent semantic clashes. Additionally, we introduce Sync-Preserving Training and Guidance (SPTG) to enhance alignment without degrading quality. Due to the scarcity of paired data, we construct a scalable data pipeline and a large-scale subject-level dataset. We also propose SpongeBob-Bench for systematic evaluation. Experiments show SpongeBob significantly outperforms existing baselines, improving Sync-C by 19.0% and Ctx-F1 by 9.5%.

1 INTRODUCTION

Recent advances in diffusion models have revolutionized video editing (Jiang et al., 2025; Liang et al., 2025b; Qin et al., 2024; Yang et al., 2025; Ju et al., 2026; Liao et al., 2025; Zi et al., 2026), enabling

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.

precise subject-level manipulation of visual content with high temporal consistency. However, in the physical world, visual events are inextricably coupled with their acoustic counterparts; any modification to a visual subject (e.g., deleting a speaking person or exchanging a barking dog for a meowing cat) must be naturally reflected in its synchronized audio stream to maintain physical plausibility. Despite the maturity of visual-only editing, without cross-modal synergy, even pixel-level visual fidelity fails to preserve realism. Consequently, developing a framework for synchronized audio-visual editing has emerged as a critical demand for next-generation multimodal content creation.

Existing audio-visual editing methods follow three paradigms. Cascaded pipelines (Jiang et al., 2025; Shan et al., 2025; Ishii et al., 2025) edit the video first and then generate or repair the audio with separate models, and training-free methods (Lin et al., 2026) rely on cross-modal noise inversion or unidirectional condition injection. Both keep the modalities in disjoint stages without feedback during denoising, causing (1) audio-visual desynchronization: edited lip motion and sound events drift apart at the frame level, e.g., speech lags the mouth opening or a door-slam sound arrives after the door closes. End-to-end models SAVE (Xu et al., 2025), InstructAV2AV (Zheng et al., 2026b), and JAVEdit (Chen et al., 2026) denoise both modalities jointly and narrow this gap, but all are trained on curated source-target editing pairs, and InstructAV2AV does not model conflicts between newly generated and preserved audio. (2) Audio-visual context conflict thus remains largely unaddressed: newly generated sounds ignore the unedited context, e.g., when only speaker A of a dialogue is edited, the regenerated voice may overlap speaker B's unedited turn. What is missing is *an end-to-end editor that learns without paired editing data and explicitly resolves conflicts with the preserved audio track.*

Building such an editor raises two fundamental challenges. (1) Data challenge. Conventionally, training an end-to-end editor would require (pre-edit audio-visual content, post-edit audio-visual content, editing instruction) triplets as supervision. However, such triplets do not occur naturally at scale, since no web-scale corpus contains the same scene edited in two different ways, while collecting them by hand is prohibitively expensive. (2) Architecture challenge. Even with data in place, the framework must satisfy two closely related aspects during denoising: (i) Synchronization modeling: the model must continuously maintain temporal correspondence between target visual motion and sound events during denoising, rather than passively aligning them post-generation; cross-modal interaction must also incorporate spatial constraints so that audio-driven visual changes act only on the target subject region without spreading to the background or other instances. (ii) Context preservation: audio editing must preserve the source context, including background sounds, ambient audio, and non-target speakers, so that newly generated subject-specific sounds coexist harmoniously with the original audio-visual scene, rather than rebuilding the audio track from scratch.

In this paper, we present **SpongeBob**, *a dual-stream Diffusion Transformer (DiT) that addresses both challenges within a single unified framework*. At its core, SpongeBob reformulates audio-visual editing as a self-supervised inpainting task: given any ordinary audio-visual clip, we mask the target subject in both modalities and train the model to reconstruct the original signal conditioned on a textual description of what was masked; at inference, the user provides a different textual description for the masked region, so reconstruction becomes targeted editing. This reformulation replaces the impractical need for (pre-edit, post-edit, instruction) triplets with (clip, mask, caption) samples that can be produced from ordinary single-take videos via automated segmentation, audio separation, and multi-stage filtering, thereby resolving the data challenge and unlocking end-to-end training. Under this formulation, SpongeBob addresses the architecture challenge through three tightly coupled components. Sync-Aware Editing Mechanism targets synchronization modeling, aligning target visual motion and sound events during denoising via bidirectional cross-modal attention (interaction), three-way temporal RoPE unification (temporal correspondence), and mask-guided asymmetric routing (spatial constraints). Context-Aware Module targets context preservation by adding two zero-initialized cross-attention layers, Acoustic Context Attention over the base audio track and Visual Context Attention over the unedited video region, so the generated audio perceives what must be preserved rather than resynthesizing from scratch. Sync-Preserving Training and Guidance (SPTG) activates these capabilities through a multi-task alignment training schedule and a two-stage inference guidance scheme. SpongeBob integrates bidirectional cross-modal attention within a unified denoising step for subject-level audio-visual editing *without requiring paired editing data*.

Our main contributions are summarized as follows:

1. Problem reformulation. We **recast subject-level audio-visual editing from a supervised task into a self-supervised inpainting task** that needs only ordinary audio-visual clips

paired with textual descriptions of their content. This reformulation enables end-to-end training without any paired editing data.

2. Architecture. We propose **SpongeBob**, an end-to-end audio-visual joint editing framework that realizes bidirectional cross-modal interaction within a unified denoising process and requires no paired editing data, with three key designs: Sync-Aware Editing Mechanism addresses desynchronization from interaction, temporal, and spatial dimensions; Context-Aware Module addresses context conflict from audio and visual dimensions; SPTG protects cross-modal synchronization and context consistency at both training and inference stages.
3. Data engineering. We build a scalable data pipeline that produces a large-scale dataset from unlabeled video for training and benchmarking on subject-level audio-visual editing.

2 Related Work

Audio-Visual Editing. Existing methods either decouple the two modalities or depend on paired editing data. Zero-shot methods (Lin et al., 2026) lack instance-level control; cascaded methods (Jiang et al., 2025; Shan et al., 2025; Ishii et al., 2025) edit video before generating or editing audio, so the video stage cannot perceive audio; Object-AVEdit (Fu et al., 2025) edits the two modalities with independent models that exchange no state; and AVI-Edit (Zheng et al., 2026a) uses audio only as a unidirectional condition. The end-to-end models closest to ours, SAVE (Xu et al., 2025), the contemporaneous InstructAV2AV (Zheng et al., 2026b), and JAVEdit (Chen et al., 2026), edit both modalities jointly but require paired source-target supervision; SAVE supports only object removal, and InstructAV2AV does not model conflicts with the preserved base audio. SpongeBob performs *bidirectional cross-modal interaction within a unified diffusion process*, learned without paired editing data. Appendix A gives a detailed discussion and reviews video-only and audio-only editing.

3 The SpongeBob Framework

As shown in Fig. 2, SpongeBob employs a dual-stream Diffusion Transformer (DiT) architecture based on Wan2.2-TI2V-5B (Wan et al., 2025) that simultaneously edits video and audio within a unified denoising process. Specifically, for the video stream, the DiT takes a composite latent of a reference image, the masked video (the context), and visual noise as input, while the visual description is injected via cross-attention to guide the reconstruction of the original video clip. For the audio stream, the model reconstructs the target audio (i.e., the isolated sound of the target subject) from audio noise. This process is conditioned on the audio description, speech text, and the base audio (i.e., the ambient audio remaining after target audio separation) via dedicated cross-attention layers. To ensure effective audio editing, we categorize the target audio into speech and non-speech streams: for speech, the audio description is fixed to a generic prompt (e.g., "*a person is speaking*") while the speech text contains the specific linguistic content; for non-speech events, the audio description provides a semantic depiction of the sound (e.g., "*a dog is barking*") while the speech text remains null. Denoising operates entirely on the target-audio latent. The predicted latent is decoded by the Audio VAE and only then combined with the preserved base audio *in the waveform domain* to recover the complete acoustic signal.

3.1 Sync-Aware Editing Mechanism

Bidirectional cross-modal attention between the two streams (Appendix B) enables interaction, but achieving frame-level audio-visual alignment still requires explicit synchronization mechanisms. SpongeBob further establishes alignment along two dimensions via the sync-aware editing mechanism: **temporal correspondence** that ensures frame-level synchronization between lip motion, action onset, and sound events; and **spatial awareness** that confines audio-driven visual changes to the target editing region without disturbing the background or other instances.

RoPE alignment for temporal correspondence. A fundamental challenge in synchronized audio-visual editing is the lack of explicit temporal correspondence between heterogeneous token streams. A sequential assignment of Rotary Positional Embeddings (RoPE) would assign distinct indices to different modalities, treating them as logically separate events even if they occur simultaneously. Video-to-audio generators such as MMAudio (Cheng et al., 2025) already place audio and video in a shared RoPE coordinate system. Editing adds a third stream, the masked condition video, which must share one timeline with both the target video and the audio. To resolve this and enforce cross-modal

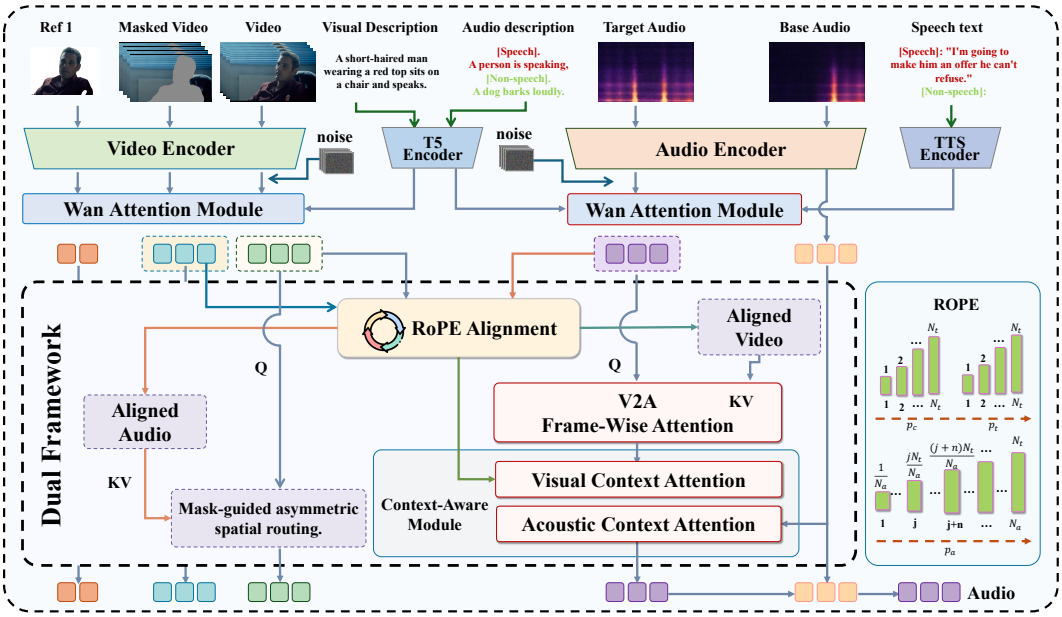

Figure 2: **Overview of SpongeBob.** Given a source video with an object mask, text instructions, and a reference image, SpongeBob jointly edits the visual content and synthesizes synchronized audio through a dual-stream DiT with sync-aware editing mechanism and context-aware module.

temporal equivalence, we propose a three-way temporal alignment strategy:

$$\underbrace{p_{\text{ref}}{=}0,\ p_{\text{c}}^{(i)}{=}i,\ p_{\text{t}}^{(i)}{=}N_t{+}i,\ p_{\text{a}}^{(j)}{=}j}_{\text{Naïve}} \longrightarrow \underbrace{p_{\text{ref}}{=}0,\ p_{\text{c}}^{(i)}{=}i,\ p_{\text{t}}^{(i)}{=}i,\ p_{\text{a}}^{(j)}{=}j{\cdot}\frac{N_t}{N_a}}_{\text{Ours}}, \tag{1}$$

where $p_{\text{ref}}, p_{\text{c}}, p_{\text{t}}, p_{\text{a}}$ denote the temporal indices for the reference image, condition video, target video, and audio streams, respectively. Specifically, we set $p_{\text{ref}} = 0$ to anchor the static reference image outside the dynamic timeline while maintaining its global accessibility. The condition and target video tokens share identical temporal indices $[1, N_t]$ to ensure the temporal alignment, and are distinguished by different diffusion timesteps ($t_{\text{cond}}{=}0$ vs. $t_{\text{target}}{=}t$). Most crucially, to align the audio with the video despite the difference in token counts ($N_a \neq N_t$), we map each audio token j to a continuous virtual position $p_{\text{a}}^{(j)} = j \cdot (N_t/N_a)$, achieving sub-frame temporal synchronization.

Mask-Guided Asymmetric Spatial Routing. To achieve precise subject-level control while preventing cross-modal contamination, we implement an asymmetric spatial routing mechanism based on the visual mask $\mathbf{M}$. Specifically, in the audio-to-video direction, acoustic features are injected strictly into visual tokens within $\mathbf{M}$. This localized routing acts as a spatial gate, ensuring that audio-driven updates are confined to the target subject and do not leak into the immutable background. Conversely, the video-to-audio direction maintains a global receptive field. This asymmetry is essential because, while the sound source is localized, the acoustic signature is intrinsically shaped by the global context.

3.2 Context-Aware Module

The synchronization-aware attention in Section 3.1 ensures alignment between modalities but lacks awareness of the unedited environment. Without such context, generated subject-specific sounds may conflict with the background (e.g., overlapping with non-target speakers or completely losing the ambient audio). To enforce harmonious coexistence, we introduce the Context-Aware Module, which integrates the target audio with the unedited visual and acoustic surroundings through two zero-initialized cross-attention layers.

First, to ensure the synthesized audio is physically consistent with the visual scene, we enable the target audio stream to perceive the unedited visual background. Formally, we define the Visual Context Attention as:

$$\mathbf{h}_{\text{vis}} = \text{CrossAttn}(\mathbf{Q} = \mathbf{x}_a^{\text{target}},\ \mathbf{K} = \mathbf{W}_k^v \mathbf{z}_{\text{cond}},\ \mathbf{V} = \mathbf{W}_v^v \mathbf{z}_{\text{cond}}), \tag{2}$$

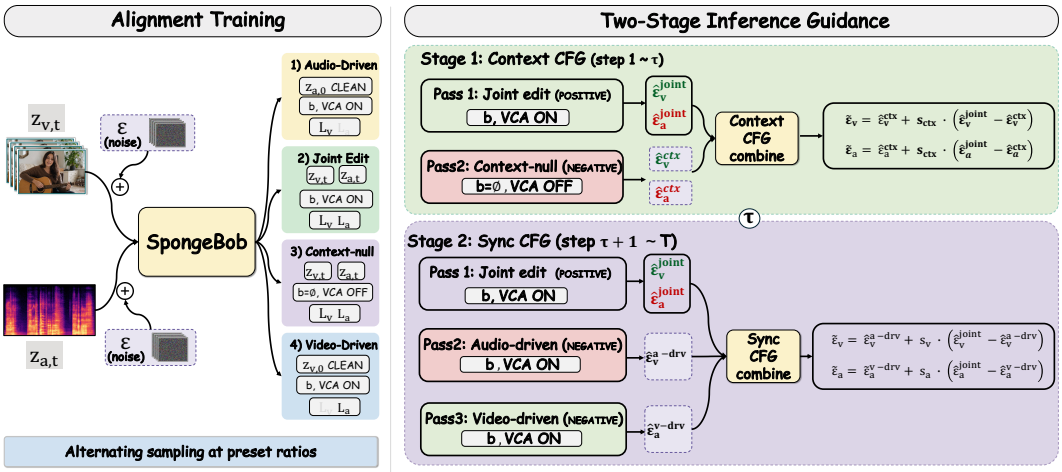

Figure 3: **Overview of SPTG. Left**: Multi-task alignment training co-trains four modes (joint edit, audio-driven, video-driven, and context-null) to teach stable cross-modal alignment under varied conditioning. **Right**: Two-stage inference guidance first resolves context conflicts via context CFG (Stage 1), then enhances temporal synchronization via sync CFG (Stage 2).

where $\mathbf{z}_{\text{cond}}$ is the masked video feature. The query originates from the target audio latent, and keys/values are derived from the masked video, enabling audio generation to perceive the visual context beyond the localized editing region.

Second, while visual grounding provides spatial context, avoiding conflicts with existing sounds requires direct perception of the base audio. Therefore, we formulate the Acoustic Context Attention as:

$$\mathbf{h}_{\text{base}} = \text{CrossAttn}(\mathbf{Q} = \mathbf{x}_a^{\text{target}},\ \mathbf{K} = \mathbf{W}_k^b \mathbf{b},\ \mathbf{V} = \mathbf{W}_v^b \mathbf{b}), \tag{3}$$

where $\mathbf{b} \in \mathbb{R}^{N_a \times D_{\text{audio}}}$ is the base audio encoded by the Audio VAE after source separation. By attending to $\mathbf{b}$, the target audio stream can perceive the presence of non-target speakers and ambient noise in real-time, allowing it to adapt its energy and timing to avoid overlaps.

3.3 Sync-Preserving Training and Guidance (SPTG)

The architecture in the previous sections integrates cross-modal synchronization and context awareness into the denoising process. However, relying solely on joint denoising yields suboptimal results: cross-modal attention falters in high-noise regimes, and standard inference-time CFG enhances text fidelity without explicitly enforcing synchronization or context consistency. As illustrated in Fig. 3, SPTG addresses these limitations through targeted training and inference strategies.

3.3.1 Multi-Task Alignment Training

To mitigate correspondence drift, SpongeBob incorporates three auxiliary objectives alongside the primary editing task. During training, each sample is dynamically routed to one of four modes based on preset probabilities, with the loss computed exclusively for the assigned mode.

Joint edit. Both video and audio targets are denoised at timestep t with all conditions and context:

$$\mathcal{L}_{\text{joint}} = \|\boldsymbol{\epsilon}_v - \hat{\boldsymbol{\epsilon}}_v(\mathbf{z}_{v,t}, \mathbf{z}_{a,t}, \mathbf{c}, \mathbf{b})\|^2 + \|\boldsymbol{\epsilon}_a - \hat{\boldsymbol{\epsilon}}_a(\mathbf{z}_{a,t}, \mathbf{z}_{v,t}, \mathbf{c}, \mathbf{b})\|^2. \tag{4}$$

Audio-driven. Audio timestep is set to t_a=0 (clean audio), video denoises at t_v=t. The clean audio provides a deterministic anchor for cross-modal attention, mitigating alignment drift. Only the video loss is computed:

$$\mathcal{L}_{\text{a-drv}} = \|\boldsymbol{\epsilon}_v - \hat{\boldsymbol{\epsilon}}_v(\mathbf{z}_{v,t}, \mathbf{z}_{a,0}, \mathbf{c}, \mathbf{b})\|^2. \tag{5}$$

Video-driven. Symmetrically, video timestep is set to t_v=0, audio denoises at t_a=t:

$$\mathcal{L}_{\text{v-drv}} = \|\boldsymbol{\epsilon}_a - \hat{\boldsymbol{\epsilon}}_a(\mathbf{z}_{a,t}, \mathbf{z}_{v,0}, \mathbf{c}, \mathbf{b})\|^2. \tag{6}$$

Context-null. Both modalities denoise normally with cross-modal attention active, but base audio is nulled ($\mathbf{b}\to\varnothing$) and Visual Context Attention is skipped. The model learns a baseline prediction without context awareness:

$$\mathcal{L}_{\text{ctx}} = \|\boldsymbol{\epsilon}_v - \hat{\boldsymbol{\epsilon}}_v(\mathbf{z}_{v,t}, \mathbf{z}_{a,t}, \mathbf{c}, \varnothing)\|^2 + \|\boldsymbol{\epsilon}_a - \hat{\boldsymbol{\epsilon}}_a(\mathbf{z}_{a,t}, \mathbf{z}_{v,t}, \mathbf{c}, \varnothing)\|^2. \tag{7}$$

The sampling probabilities for the four modes are p_{joint}, $p_{\text{a-drv}}$, $p_{\text{v-drv}}$, and p_{ctx} respectively. Text conditions and base audio are independently dropped at preset probabilities. All audio losses above are defined on the target-audio latent. The base audio enters solely as a condition through Acoustic Context Attention, and gradients reach that module through the conditioning pathway rather than through any composed full-audio latent.

3.3.2 Two-Stage Inference Guidance

Unlike standard CFG, which prioritizes textual fidelity without explicitly strengthening context awareness or synchronization, SpongeBob leverages its four trained modes to devise a two-stage guidance strategy. This scheme addresses context conflicts before enhancing cross-modal alignment.

Stage 1: Context conflict resolution (steps $1\sim\tau$). Full-conditional and context-null predictions construct a context CFG. Let $\hat{\boldsymbol{\epsilon}}^{\text{joint}}$ denote the full-conditional prediction and $\hat{\boldsymbol{\epsilon}}^{\text{ctx}}$ the context-null prediction ($\mathbf{b}\to\varnothing$, Visual Context Attention skipped):

$$\tilde{\boldsymbol{\epsilon}}_v = \hat{\boldsymbol{\epsilon}}_v^{\text{ctx}} + s_{\text{ctx}} \cdot (\hat{\boldsymbol{\epsilon}}_v^{\text{joint}} - \hat{\boldsymbol{\epsilon}}_v^{\text{ctx}}), \qquad \tilde{\boldsymbol{\epsilon}}_a = \hat{\boldsymbol{\epsilon}}_a^{\text{ctx}} + s_{\text{ctx}} \cdot (\hat{\boldsymbol{\epsilon}}_a^{\text{joint}} - \hat{\boldsymbol{\epsilon}}_a^{\text{ctx}}). \tag{8}$$

The guidance direction isolates the contribution of the Context-Aware Module, ensuring the generated audio respects unedited content and avoids context conflicts. This stage requires 2 forward passes.

Stage 2: Temporal synchronization enhancement (steps $\tau{+}1\sim T$). Since clean target audio/video is unavailable at inference, muted audio $\mathbf{z}_{a,0}^{\varnothing}$ (an all-zero waveform encoded by the Audio VAE) and static video $\mathbf{z}_{v,0}^{\varnothing}$ (a white frame repeated over all frames and encoded by the Video VAE) serve as negative anchors via the audio-driven and video-driven pathways:

$$\tilde{\boldsymbol{\epsilon}}_v = \hat{\boldsymbol{\epsilon}}_v^{\text{a-drv}} + s_v \cdot (\hat{\boldsymbol{\epsilon}}_v^{\text{joint}} - \hat{\boldsymbol{\epsilon}}_v^{\text{a-drv}}), \qquad \hat{\boldsymbol{\epsilon}}^{\text{a-drv}} = \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_{v,t}, \mathbf{z}_{a,0}^{\varnothing}, \mathbf{c}, \mathbf{b}), \tag{9}$$

$$\tilde{\boldsymbol{\epsilon}}_a = \hat{\boldsymbol{\epsilon}}_a^{\text{v-drv}} + s_a \cdot (\hat{\boldsymbol{\epsilon}}_a^{\text{joint}} - \hat{\boldsymbol{\epsilon}}_a^{\text{v-drv}}), \qquad \hat{\boldsymbol{\epsilon}}^{\text{v-drv}} = \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_{v,0}^{\varnothing}, \mathbf{z}_{a,t}, \mathbf{c}, \mathbf{b}). \tag{10}$$

The guidance directions isolate audio-driven visual changes (lip motion, sound-source actions) and vision-driven audio changes (speech rhythm, action sound effects). This stage requires 3 forward passes ($\hat{\boldsymbol{\epsilon}}^{\text{joint}}$ shared). The two stages are complementary: Stage 1 establishes context consistency in early denoising steps, while Stage 2 refines frame-level temporal correspondence in later steps.

3.4 Data Pipeline

We construct a **scalable data pipeline** (Fig. 4) that automatically converts raw videos into subject-level (clip, mask, caption) reconstruction samples without manual annotation.

The pipeline consists of six stages: (1) **Video collection and classification**: Videos are sourced from public video datasets, including Koala-36M (Wang et al., 2025) and OpenHumanVid (Li et al., 2025), filtered by 50+ fine-grained acoustic categories to ensure a single dominant target subject. (2) **Multimodal source identification**: Gemini (Comanici et al., 2025) jointly analyzes audio-visual cues to classify sound sources as foreground (target) or background. (3) **Instance segmentation**: Grounded-SAM-2 (Ren et al., 2024) produces per-frame masks of the target subject from its text description. (4) **Text-guided separation**: SAM-Audio (Shi et al., 2025) separates the mixture into target and base audio conditioned on source descriptions, so any non-target speakers and ambient sounds in the clip stay in the base audio. (5) **Multi-dimensional verification**: Gemini assesses separation quality (matching, completeness, leakage, and fidelity); only qualified samples proceed. (6) **Joint filtering**: Samples are exported upon passing strict criteria for audio quality, mask validity, ASR correctness, and base-audio cleanliness. Each sample comprises the original video, editing mask, target/base audio, reference image, and text description. The original video serves as the reconstruction ground truth, while the mask defines the editing region and base audio provides context for the Context-Aware Module. The resulting dataset contains 400K samples (≈390 hours).

4 Experiments

4.1 Experimental Setup

SpongeBob-Bench and evaluation metrics. We propose SpongeBob-Bench for systematic evaluation of joint audio-visual editing, comprising 700 test samples across three subsets: Speech-Video (400), Sound-Video (100), and Complex Scene (200). Metrics cover video quality (Huang et al., 2024) (FVD (Unterthiner et al., 2019), MS, DD, BG), audio quality (PQ, CLAP (Elizalde et al., 2023)),

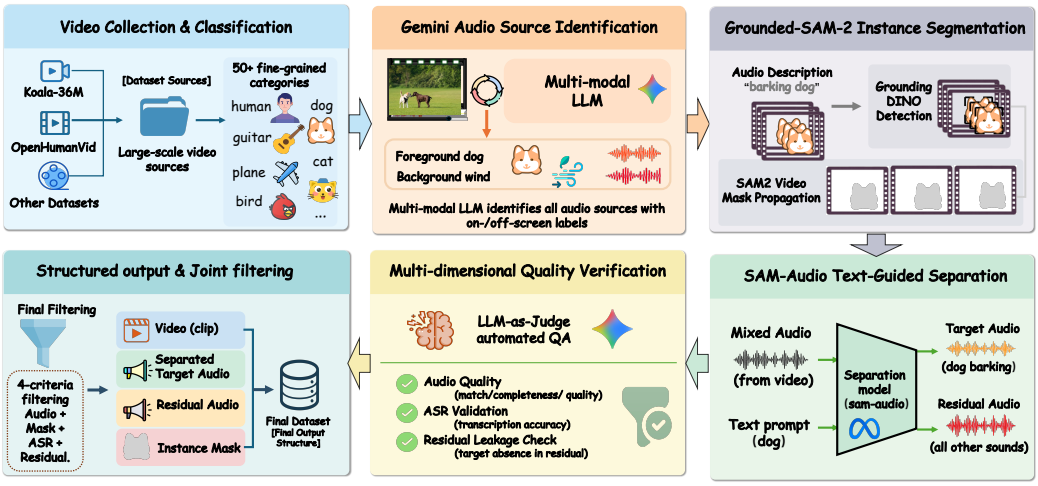

Figure 4: **Data pipeline.** The pipeline performs six automated stages to convert raw videos into subject-level (clip, mask, caption) reconstruction samples without manual annotation.

Table 1: **SpongeBob-Bench comprehensive evaluation.** Evaluation across video quality, audio quality, AV synchronization, and context consistency. Best in **bold**, second best underlined.

	Video Quality				Audio Quality		AV Sync				Context	
Method	FVD↓	MS↑	DD↑	BG↑	PQ↑	CLAP↑	Sync-C↑	Sync-D↓	IB↑	DeSync↓	Ctx-F1↑	G-Score↑
AvED	548.37	0.952	0.18	0.862	4.85	0.215	1.15	12.85	0.150	1.28	0.52	3.6
VACE-Foley	372.15	0.982	0.32	0.918	5.85	0.208	1.85	11.42	0.190	0.96	0.62	5.3
VACE-Coh.	372.15	0.982	0.32	0.918	6.16	0.222	3.39	10.33	0.212	0.80	0.74	6.3
AVI-Edit	318.56	0.985	0.35	0.932	6.12	0.225	3.45	10.28	0.210	0.82	0.72	6.2
InstructAV2AV	300.50	0.987	0.34	0.937	6.21	0.231	3.78	9.96	0.216	0.74	0.73	6.5
Ours	**285.93**	**0.990**	**0.36**	**0.951**	**6.45**	**0.238**	**4.50**	**8.73**	**0.240**	**0.58**	**0.81**	**7.6**

AV synchronization (Sync-C/Sync-D (Chung & Zisserman, 2016), IB (Girdhar et al., 2023), and Synchformer (Iashin et al., 2024) DeSync for non-speech events), and context consistency (Ctx-F1 based on pyannote (Bredin et al., 2020) and the Gemini 2.5 Pro G-Score). We additionally evaluate on the external AvED-Bench (Lin et al., 2026) with its original metric suite (FVD, IS, FC, TC, AC) to verify generalization. Benchmark construction and metric definitions are given in Appendix C.

Baselines. We compare with five methods representing different paradigms: the zero-shot AvED (Lin et al., 2026); the cascades VACE-Foley (Jiang et al., 2025; Shan et al., 2025), VACE-Coherent (Jiang et al., 2025; Ishii et al., 2025), and AVI-Edit (Zheng et al., 2026a); and the end-to-end InstructAV2AV (Zheng et al., 2026b). All methods use their original pretrained weights and are evaluated on the same test set; their configurations are detailed in Appendix D.

4.2 Main Results

Qualitative comparison. Fig. 5 compares results on multi-speaker dialogue, single-speaker, and non-speech sound scenarios. SpongeBob achieves synchronized editing across all cases, whereas the baselines exhibit audio-visual misalignment or context disruption. Even InstructAV2AV falls short despite visually clean frames: lacking a reference-image input, it cannot reproduce the reference identity; in the non-speech case it produces a brown-coated animal instead of the fluffy, light-colored cat in the prompt; and in the dialogue it likewise does not fully meet the editing instruction.

SpongeBob-Bench results. Table 1 reports averaged results across the three subsets. SpongeBob achieves the best performance across all four evaluation dimensions. In video quality and audio quality, SpongeBob leads comprehensively, demonstrating that end-to-end joint denoising does not sacrifice single-modality quality. More importantly, the core advantages lie in AV synchronization and context consistency: SpongeBob improves Sync-C over the strongest baseline InstructAV2AV by **19.0%** and Ctx-F1 over the best baseline, VACE-Coherent, by **9.5%**. Notably, VACE-Foley and VACE-Coherent share identical video metrics (both use VACE-14B, producing identical video output). VACE-Coherent surpasses VACE-Foley on every audio-side metric, indicating that conditioning audio generation on the source track preserves acoustic structure and event timing far better than

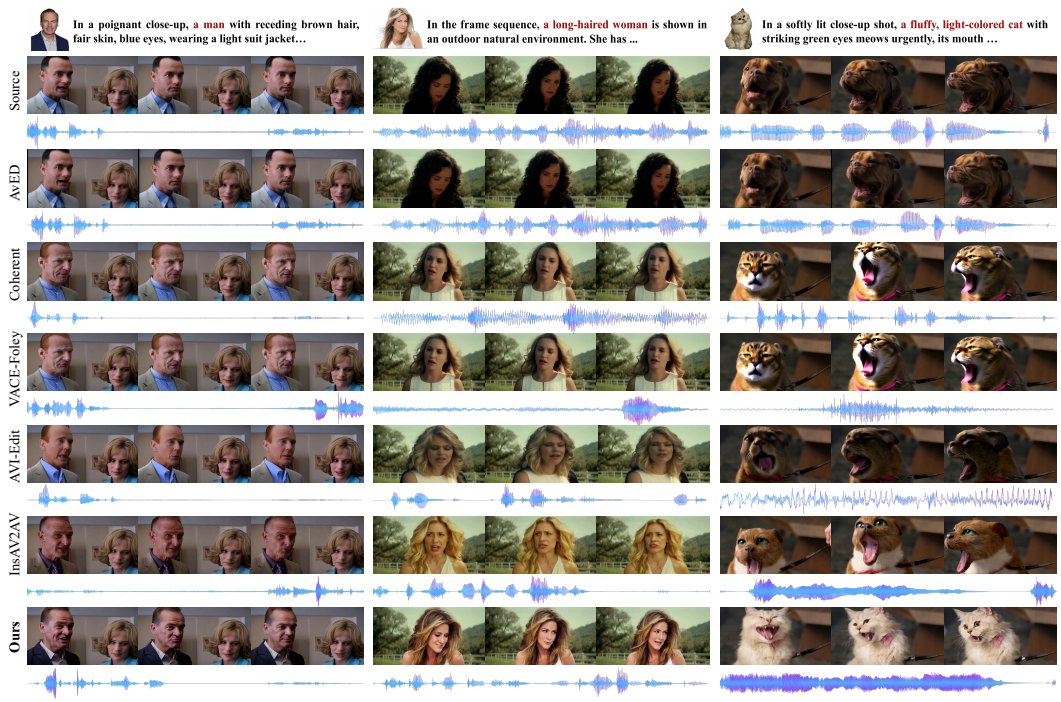

Figure 5: **Multi-scenario qualitative comparison.** SpongeBob achieves faithful visual editing with synchronized audio across all scenarios. "Coherent" and "InsAV2AV" denote VACE-Coherent and InstructAV2AV.

Table 2: **Overall component ablation.**

Variant	PQ↑	BG↑	Sync-C↑	Sync-D↓	IB↑	Ctx-F1↑
w/o T	6.40	0.943	3.72	9.79	0.21	0.80
w/o M	6.36	0.922	3.85	9.61	0.22	0.79
w/o M & T	6.28	0.915	3.18	10.52	0.20	0.78
w/o Ctx Mod	6.15	0.924	4.25	9.45	0.22	0.75
w/o SPTG	6.08	0.935	3.65	10.12	0.21	0.76
Full	**6.45**	**0.951**	**4.50**	**8.73**	**0.24**	**0.81**

Table 3: **AvED-Bench generalization.**

Method	FVD↓	IS↑	FC↑	TC↑	AC↑
AvED	435.2	1.110	94.52	24.35	20.18
VACE-Foley	418.2	1.105	95.48	25.02	21.12
VACE-Coh.	418.2	1.105	95.48	25.02	21.47
AVI-Edit	358.5	1.120	95.68	25.18	21.52
InstructAV2AV	347.9	1.116	95.61	25.29	21.79
Ours	**338.6**	**1.130**	**96.05**	**25.45**	**22.15**

Table 4: **Context-aware module ablation.**

Variant	PQ↑	BG↑	Sync-C↑	Sync-D↓	Ctx-F1↑
No context	6.15	0.924	4.25	9.45	0.75
Acoustic only	6.25	0.925	4.30	9.38	0.72
Visual only	6.18	0.932	4.35	9.32	0.80
Full (A+V)	**6.45**	**0.951**	**4.50**	**8.73**	**0.81**

Table 5: **SPTG ablation.**

Variant	CLAP↑	PQ↑	Sync-C↑	Sync-D↓	BG↑	Ctx-F1↑
No CFG	0.218	5.82	3.85	9.82	0.942	0.73
Std. 2-pass	0.232	6.25	3.90	9.78	0.940	0.73
S1 (ctx)	0.228	6.15	3.95	9.68	0.948	0.78
S2 (sync)	0.230	6.18	4.32	9.35	0.940	0.72
Full	**0.238**	**6.45**	**4.50**	**8.73**	**0.951**	**0.81**

synthesizing the track from scratch. However, both remain far below SpongeBob on synchronization, confirming that source audio conditioning within a cascade cannot compensate for the lack of bidirectional interaction. InstructAV2AV, the only other end-to-end model, is the strongest baseline on most metrics, including all synchronization metrics, which confirms the value of denoising the two modalities jointly. Its Ctx-F1, however, stays below that of VACE-Coherent: joint denoising by itself does not model conflicts between the newly generated target and the preserved base audio, which is precisely what the Context-Aware Module adds.

AvED-Bench generalization. As shown in Table 3, SpongeBob leads across all metrics, demonstrating that our framework generalizes beyond speech scenarios (details in Appendix F).

4.3 Ablation Study

Overall Ablations. As shown in Table 2, removing any component causes notable degradation. Temporal unification (T) and mask-guided routing (M) are ablated individually as well as jointly,

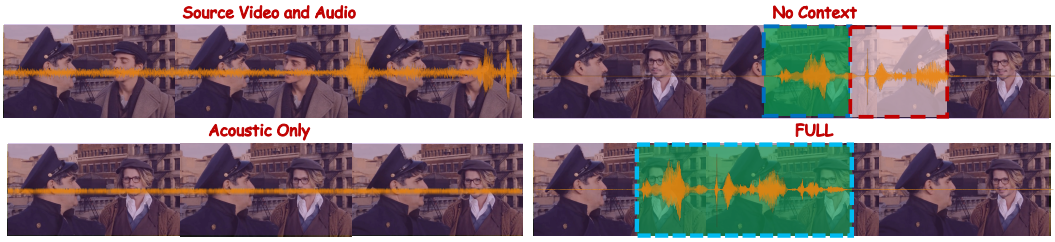

Figure 6: **Context conflict visualization.** Red indicates overlap with non-target speech; green indicates correct generation. The full module precisely avoids non-target speech while generating normally during ambient sounds.

with bidirectional cross-modal attention held fixed so that each configuration isolates a single factor. Removing T mainly degrades synchronization while leaving background preservation nearly unchanged, showing that T establishes frame-level temporal correspondence. Removing M degrades both background preservation and synchronization: without spatial gating, audio-driven updates are no longer confined to the target subject, which in two-speaker scenes surfaces as audio-driven mouth motion on the non-target speaker. Removing both yields the largest synchronization drop, confirming that the two supply complementary temporal and spatial constraints. Removing the Context-Aware Module yields the lowest Ctx-F1 among all variants, confirming its necessity for resolving context conflicts. Removing SPTG degrades every metric, validating that SPTG is indispensable for video quality, audio quality, synchronization, and context consistency alike.

Necessity of Context-Aware Module. As shown in Table 4, without any context module (No context), Ctx-F1 is 0.75 due to heavy temporal overlap between target and non-target speech. With Visual Context Attention alone, the model leverages lip movements in the masked video to localize non-target speech intervals, achieving selective avoidance and raising Ctx-F1 to 0.80. With Acoustic Context Attention alone, although PQ improves, Ctx-F1 drops to 0.72 (below No context). As illustrated in Fig. 6, since base audio after source separation remains a mixed signal, the model cannot distinguish non-target speech from ambient sound and conservatively avoids all acoustically active intervals, causing excessive target silence. The full configuration combines both: Visual Context Attention provides cross-modal disambiguation to localize non-target speech, while Acoustic Context Attention provides the original acoustic environment as reference, yielding the best Ctx-F1 of 0.81.

Efficacy of SPTG. Table 5 compares five inference guidance strategies on the same trained model. Without any guidance (No CFG), single-modality quality is low (PQ 5.82, CLAP 0.218). Standard 2-pass CFG raises PQ and CLAP but barely improves synchronization and leaves context unchanged (Sync-C 3.85→3.90, Ctx-F1 0.73→0.73), confirming that standard CFG primarily enhances single-modality adherence with limited cross-modal benefit. Stage 1 alone (context CFG) raises Ctx-F1 to 0.78 and BG to 0.948. Stage 2 alone (sync CFG) raises Sync-C to 4.32 and lowers Sync-D to 9.35, but Ctx-F1 slightly decreases (0.72). Full SPTG combining both stages achieves the best across all metrics (Ctx-F1 0.81, Sync-C 4.50), demonstrating complementary enhancement.

5 Conclusion

We present **SpongeBob**, an end-to-end audio-visual joint editing framework based on bidirectional cross-modal interaction that simultaneously edits visual content and synthesizes synchronized audio. By recasting subject-level editing as self-supervised audio-visual inpainting, SpongeBob learns from (clip, mask, caption) samples that our data pipeline extracts from ordinary videos, without paired editing data. It targets two failures of existing editors: audio-visual desynchronization in decoupled pipelines and, across paradigms, conflicts between newly generated sound and the preserved audio track. The Sync-Aware Editing Mechanism aligns visual motion and sound events from interaction, temporal, and spatial dimensions; the Context-Aware Module perceives unedited audio-visual context to avoid such conflicts; and SPTG protects cross-modal synchronization and context consistency at both training and inference stages. Together with **SpongeBob-Bench**, these contributions demonstrate that joint audio-visual editing requires carefully designed cross-modal information flow across spatial, temporal, and acoustic dimensions, which can be learned without paired editing data.

AI USE STATEMENT

We used generative AI tools for editing only: grammar, spelling, and word choice in the manuscript. All technical claims, derivations, and experimental descriptions were written and verified by the authors.

Separately, large multimodal models appear as components of the proposed method and of the evaluation protocol rather than as authoring aids. Gemini performs multimodal sound-source identification, LLM-as-Judge separation-quality verification, and mask-quality checking inside the data pipeline (Section 3.4 and Appendix I), and Gemini 2.5 Pro is the multimodal evaluator behind the G-Score metric (Appendix C). Both uses are documented in the corresponding sections, and the automated filtering was audited against manual review, as reported in Appendix I.

ETHICS STATEMENT

Positive impact. Improved efficiency in film post-production dubbing and sound replacement; accessible content creation; cross-lingual localization of educational videos; creative content production tools.

Potential risks and mitigation. Audio-visual editing technology carries risks of deepfake generation, unauthorized content tampering, and erosion of public trust in video authenticity. To mitigate them, we will (a) embed imperceptible digital watermarks in generated content for provenance verification; (b) develop and open-source detection models; (c) release models under usage terms prohibiting identity forgery and disinformation; and (d) apply tiered access control to high-risk functionalities.

Data ethics. Training data is drawn from public video datasets, including Koala-36M (Wang et al., 2025) and OpenHumanVid (Li et al., 2025), which contain the faces and voices of real people. These data are used solely for academic research, and any released data, model weights, or benchmark will comply with the licenses of the source datasets and be restricted to non-commercial research use; commercial use is strictly prohibited. Data construction is automated; human effort is limited to spot-checking 200 training samples and reviewing the benchmark (Appendices C and I), plus a user study in which 20 evaluators, compensated for their time, rated results presented anonymously in randomized order (Appendix G).

REPRODUCIBILITY STATEMENT

Appendix B specifies the model architecture and initialization, the complete training configuration, and the inference schedule including guidance scales, stage boundaries, and forward-pass counts. The data construction pipeline is summarized in Section 3.4 and detailed in Appendix I, covering the acoustic category taxonomy, the quality-verification criteria together with their measured pass rates, and the segmentation thresholds. Benchmark construction, subset partitioning, and the exact computation of every metric are given in Appendix C. We will release the data construction pipeline, the training dataset, model weights, evaluation code, and SpongeBob-Bench.

REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. pp. 37971–37981, 2026.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 7124–7128. 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.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911. IEEE, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

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.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Bing Han, Junyu Dai, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, Yanmin Qian, and Xuchen Song. Instructme: An instruction guided music edit and remix framework with latent diffusion models. *arXiv preprint arXiv:2308.14360*, 2023.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing supplementary material. pp. 607–623, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

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.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Cherry Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, et al. Editverse: Unifying image and video editing and generation with in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 137234–137255, 2026.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Zitong Lan, Yiduo Hao, and Mingmin Zhao. Guiding audio editing with audio language model. *arXiv preprint arXiv:2509.21625*, 2025.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762. IEEE, 2025.

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*, 2025a.

Sen Liang, Zhentao Yu, Zhengguang Zhou, Teng Hu, Hongmei Wang, Yi Chen, Qin Lin, Yuan Zhou, Xin Li, Qinglin Lu, et al. Omniv2v: Versatile video generation and editing via dynamic content manipulation. *arXiv preprint arXiv:2506.01801*, 2025b.

Xinyao Liao, Xianfang Zeng, Ziye Song, Zhoujie Fu, Gang Yu, and Guosheng Lin. In-context learning with unpaired clips for instruction-based video editing. *arXiv preprint arXiv:2510.14648*, 2025.

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.

Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stabilizing shape consistency in video-to-video editing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

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.

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.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8428–8437. IEEE, 2025.

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.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16692–16701. IEEE, 2025.

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*, 2025.

Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yan Huang, Min Xu, and Qiang Wu. Unified video editing with temporal reasoner. *arXiv e-prints*, pp. arXiv–2512, 2025.

Haojie Zheng, Shuchen Weng, Jingqi Liu, Siqi Yang, Boxin Shi, and Xinlong Wang. 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.

Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *Advances in Neural Information Processing Systems*, 38, 2026.

A Extended Related Work

Audio-Visual Editing. Existing audio-visual editing methods fall into three categories. Zero-shot methods (Lin et al., 2026) suffer from low frame rates and lack instance-level control. Cascaded methods (Jiang et al., 2025; Shan et al., 2025; Ishii et al., 2025) first edit video then generate or edit audio, where the video editing stage cannot perceive audio. Object-AVEdit (Fu et al., 2025) follows an inversion-regeneration paradigm with two independent models for audio and video, so no state is exchanged between the two denoising processes and cross-modal consistency is inherited from the source structure rather than established explicitly. AVI-Edit (Zheng et al., 2026a) drives video editing with an audio agent, but audio remains a unidirectional condition. Three end-to-end models are closest to ours. SAVE (Xu et al., 2025) denoises audio and video jointly, but supports only object removal and requires paired source-target supervision; because it synthesizes no new appearance or sound, it never confronts synchronization between newly generated streams. InstructAV2AV (Zheng et al., 2026b), contemporaneous with this work, performs instruction-guided joint editing but does not explicitly model conflicts between newly generated target audio and preserved base audio, and its supervision relies on 79K paired source-target samples. JAVEdit (Chen et al., 2026) also targets instruction-guided joint editing, contributing the agent-curated JAVEdit-100k triplet corpus and the JAVEditBench evaluation suite; its supervision is likewise paired. All the above paradigms either decouple the two modalities or depend on paired editing data; the fundamental difference of SpongeBob lies in *bidirectional cross-modal interaction within a unified diffusion process*, learned without any paired editing data: video motion, target sound, and acoustic context continuously exchange information during denoising, jointly constraining the editing result.

Video Editing. Diffusion-based video editing has developed along two main lines. Mask-guided methods (Jiang et al., 2025; Liang et al., 2025b) achieve precise region-level editing through explicit spatial conditions; they are technically mature but rely on user-provided masks, limiting flexibility. Instruction-based methods (Qin et al., 2024; Yang et al., 2025; Ju et al., 2026; Liao et al., 2025; Zi et al., 2026; Bai et al., 2026; He et al., 2026; Wu et al., 2025; Ku et al., 2024; Liu et al., 2025) infer editing intent directly from text instructions without additional spatial annotations, offering broader applicability. Although both lines have achieved significant progress in visual editing quality, their task definition remains confined to the visual modality: when the edited object is itself a sound source, the correspondence between visual content and the original audio is broken, with no mechanism to determine how the corresponding sound should change in synchrony.

Audio Editing. A parallel line of work edits audio alone. AUDIT (Wang et al., 2023) trains a latent diffusion model on (instruction, input audio, output audio) triplets for text-guided addition, removal, and replacement of sound events, and InstructME (Han et al., 2023) extends instruction-based editing and remixing to music under harmony-preserving conditions. AudioMorphix (Liang et al., 2025a) shows that comparable edits are achievable training-free, steering the diffusion process with energy functions over localized time-frequency regions. More recent work broadens the interaction style, with SmartDJ (Lan et al., 2025) decomposing a declarative instruction into atomic steps executed sequentially by a diffusion editor, and Audio-Omni (Tian et al., 2026) coupling a frozen MLLM with a Diffusion Transformer to handle sound, music, and speech in one model. In all of these methods both the edited object and the output are audio. SpongeBob instead edits the two modalities jointly: the region is delineated by a per-frame object mask in the video, and the generated sound must remain frame-level synchronized with the edited motion while avoiding conflicts with the preserved base audio and non-target speakers.

B Implementation Details

B.1 Model Architecture

The video branch is built on Wan2.2-TI2V-5B: 30-layer DiT blocks, hidden dimension 3072, 24 attention heads, patch size $1\times2\times2$ (temporal $\times$ spatial). Video VAE (Wan2.2-VAE) compression ratio is temporal $4\times$, spatial 16×16, with 48 latent channels. The audio branch adopts the VAE from MMAudio (Cheng et al., 2025) for 16 kHz audio, a 1D convolutional VAE that compresses 80-bin mel spectrograms into 20-channel latents at 31.25 fps. The audio-side DiT reuses the attention block structure of Wan2.2-TI2V-5B and is initialized from its weights; the "Wan Attention Module" in Fig. 2 refers to this reuse.

Video Key/Value in V→A are detached from the computation graph to prevent audio loss from backpropagating into the video stream.

Each audio block contains Visual Context Attention followed by Acoustic Context Attention (both cross-attention layers with zero-initialized output projections; Acoustic Context Attention local window size w=8). The condition patch embedding $\mathcal{E}_{\text{cond}}$ is initialized from $\mathcal{E}_{\text{target}}$; masked video, reference image, and target latent are concatenated along the temporal dimension into a unified input sequence.

B.2 Training Configuration

We train SpongeBob on 240 GPUs (96 GB each) with a total batch size of 240, learning rate 1×10^{-5} with cosine decay. Each training sample contains up to 121 frames (approximately 5 s at 24 FPS) at 540p resolution. Training runs for 10K steps. The four training modes are sampled with probabilities: Joint edit 0.4, Audio-driven 0.2, Video-driven 0.2, Context-null 0.2. Condition drop probabilities are 0.1 for both text and base audio (forced to 1.0 in Context-null mode). Mask augmentation applies random dilation up to 20 px per side with 30% probability of replacing the precise mask with its bounding box.

B.3 Inference Configuration

We use 50 total denoising steps with Flow Matching (linear schedule). The objectives in Section 3.3.1 and the guidance in Section 3.3.2 are written in ϵ-prediction form for readability; in practice the network predicts flow-matching velocities (noise minus clean latent), and the losses and linear guidance combinations apply to them unchanged. Stage 1 (context conflict resolution) runs for steps 1–10 with s_{ctx}=5.0; Stage 2 (temporal synchronization enhancement) runs for steps 11–50 with s_v=5.0, s_a=5.0. Negative anchors are: muted audio $\mathbf{z}_{a,0}^{\varnothing}$ (all-zero audio encoded by Audio VAE) and static video $\mathbf{z}_{v,0}^{\varnothing}$ (white image repeated over all frames, encoded by Video VAE). Stage 1 requires 2 forward passes per step and Stage 2 requires 3 (joint prediction shared), totaling $10\times2+40\times3=140$ forward passes. Single-sample inference (121 frames at 540p + audio) takes 588 s on a single H20 GPU; Table 10 compares this against the baselines under an identical protocol.

C SpongeBob-Bench Construction and Metrics

C.1 Construction Pipeline

SpongeBob-Bench is constructed through four steps: (1) candidate samples are collected from independent video sources with no overlap with the training set; (2) the same automated quality verification as training data is applied (audio separation quality, mask quality, ASR validity, base-audio cleanliness); (3) annotators write the target edit description for each sample, and human review confirms ground truth accuracy, mask precision, and text description correctness; (4) samples are assigned to three subsets based on scene characteristics, ensuring intra-subset diversity. The final benchmark contains 700 test samples with strict non-overlap from the training set.

C.2 Scene Partitioning

Speech-Video (400 samples) evaluates speaker editing, lip synchronization, and non-target speaker preservation; **Sound-Video** (100 samples) evaluates temporal synchronization between object actions and event sounds; **Complex Scene** (200 samples) evaluates context consistency when multi-person dialogues, non-target speaker voices, and ambient sounds coexist.

Subset	Samples	Core Evaluation Target
Speech-Video	400	Lip sync, speaker editing, non-target preservation
Sound-Video	100	Action-sound temporal alignment
Complex Scene	200	Multi-source context consistency

Table 6: **Speech-Video subset (400 samples).** SpongeBob achieves +17% Sync-C over the strongest baseline InstructAV2AV, benefiting from bidirectional cross-modal attention that enables lip generation to perceive audio phoneme rhythms in real time.

Method	FVD↓	MS↑	DD↑	BG↑	PQ↑	CLAP↑	Sync-C↑	Sync-D↓	IB↑	Ctx-F1↑	G↑
AvED	545.0	0.953	0.18	0.865	4.90	0.215	1.28	12.55	0.16	0.55	3.6
VACE-Foley	365.0	0.983	0.32	0.920	5.88	0.208	2.05	11.15	0.20	0.65	5.4
VACE-Coh.	365.0	0.983	0.32	0.920	6.19	0.222	3.78	9.91	0.22	0.77	6.4
AVI-Edit	310.0	0.986	0.35	0.935	6.15	0.225	3.85	9.85	0.22	0.75	6.3
InstructAV2AV	292.0	0.988	0.34	0.940	6.25	0.232	4.22	9.52	0.22	0.77	6.6
Ours	**280.5**	**0.991**	**0.36**	**0.952**	**6.48**	**0.240**	**4.95**	**8.25**	**0.24**	**0.84**	**7.7**

C.3 Metric Computation

Video quality. FVD (Fréchet Video Distance) (Unterthiner et al., 2019) measures Fréchet distance of I3D features. MS (motion smoothness) computes mean cosine similarity of DINOv2 features between adjacent frames. DD (dynamic degree) measures mean RAFT optical flow magnitude within the editing region. BG (background preservation) computes cosine similarity of DINOv2 features outside the mask before and after editing.

Audio quality. PQ uses AudioBox-Aesthetics (Tjandra et al., 2025) for perceptual quality scoring (1–10). CLAP computes cosine similarity between CLAP audio and text embeddings.

AV synchronization. Sync-C and Sync-D measure SyncNet (Chung & Zisserman, 2016) lip-sync confidence and distance respectively (only for subsets containing speakers). IB computes mean per-frame cosine similarity of ImageBind (Girdhar et al., 2023) audio-visual embeddings. IB captures semantic correspondence rather than temporal alignment, so for non-speech events we additionally report DeSync, the audio-visual temporal offset predicted by Synchformer (Iashin et al., 2024) (only for the Sound-Video subset; Tables 1 and 10).

Context consistency. Ctx-F1 is based on pyannote (Bredin et al., 2020) speaker detection: Precision $= 1-$ conflict rate (overlap between generated target audio and non-target speakers); Recall $=$ target audio coverage rate; F1 combines both, simultaneously penalizing conflicts and excessive silence. G-Score uses Gemini 2.5 Pro (Comanici et al., 2025) as a multimodal evaluator providing holistic scores (1–10), averaged over 3 evaluations per sample.

D Baseline Details

The five baselines are configured as follows. (1) **AvED** (Lin et al., 2026): a zero-shot cross-modal editing method based on pretrained text-to-image and text-to-audio diffusion models; (2) **VACE-Foley** (Jiang et al., 2025; Shan et al., 2025): a cascade that first edits video with VACE, then synthesizes audio from scratch with HunyuanVideo-Foley; (3) **VACE-Coherent** (Jiang et al., 2025; Ishii et al., 2025): likewise uses VACE for video editing but employs Coherent to edit audio conditioned on the source audio, preserving source audio structure; (4) **AVI-Edit** (Zheng et al., 2026a): uses Chatterbox-Turbo (speech TTS) and Stable Audio Open (non-speech SFX) to generate target audio, mixes with base audio, then drives the AVI-Edit video editing backbone with audio as condition; (5) **InstructAV2AV** (Zheng et al., 2026b): an end-to-end instruction-guided joint editing model, which denoises target audio and video together but does not explicitly model conflicts with the preserved base audio.

E Per-Scene Detailed Results

Table 1 in the main paper reports weighted averages across the three subsets (weights 400:100:200); FVD, although a set-level statistic, is aggregated in the same way. Sync-C, Sync-D, and Ctx-F1 are computed only on subsets containing speakers (Speech 400 + Complex 200 = 600). DeSync is computed only on the Sound-Video subset (100 samples). Below we provide per-subset breakdowns.

Table 7: **Sound-Video subset (100 samples).** No lip sync is involved, so Sync-C, Sync-D, and Ctx-F1 are not reported; IB measures semantic audio-visual correspondence, and event-level temporal alignment (DeSync) is reported in Table 10. SpongeBob achieves IB 0.28 vs. 0.25 for VACE-Coherent and InstructAV2AV (+12%).

Method	FVD↓	MS↑	DD↑	BG↑	PQ↑	CLAP↑	IB↑	G↑
AvED	520.0	0.958	0.24	0.878	5.15	0.238	0.19	4.0
VACE-Foley	348.0	0.986	0.38	0.928	6.05	0.228	0.22	5.8
VACE-Coh.	348.0	0.986	0.38	0.928	6.35	0.240	0.25	6.6
AVI-Edit	295.0	0.988	0.40	0.942	6.32	0.242	0.24	6.6
InstructAV2AV	281.5	0.990	0.39	0.946	6.38	0.250	0.25	6.8
Ours	**268.2**	**0.993**	**0.42**	**0.958**	**6.62**	**0.258**	**0.28**	**8.0**

Table 8: **Complex Scene subset (200 samples).** The most challenging subset, with every metric worse than on the other subsets. SpongeBob improves Ctx-F1 by +10% over the best baseline on this metric, VACE-Coherent, a larger margin than in Speech-Video (+9%), demonstrating greater contribution of the Context-Aware Module in complex multi-source scenarios.

Method	FVD↓	MS↑	DD↑	BG↑	PQ↑	CLAP↑	Sync-C↑	Sync-D↓	IB↑	Ctx-F1↑	G↑
AvED	569.3	0.947	0.15	0.848	4.60	0.204	0.89	13.45	0.11	0.46	3.4
VACE-Foley	398.5	0.978	0.29	0.909	5.69	0.198	1.45	11.96	0.16	0.56	4.85
VACE-Coh.	398.5	0.978	0.29	0.909	6.00	0.213	2.61	11.17	0.18	0.68	5.95
AVI-Edit	347.5	0.982	0.32	0.921	5.96	0.216	2.65	11.14	0.18	0.66	5.80
InstructAV2AV	327.0	0.983	0.30	0.927	6.04	0.220	2.90	10.85	0.19	0.65	6.10
Ours	**305.7**	**0.987**	**0.33**	**0.945**	**6.30**	**0.224**	**3.60**	**9.69**	**0.22**	**0.75**	**7.2**

F AvED-Bench Generalization

To verify generalization, we evaluate on the external AvED-Bench which focuses on non-speech environmental sound editing. For fairness and reproducibility, we use each method's original pretrained weights for inference, and AVI-Edit's target audio is generated with the open-source Stable Audio Open (Appendix D). As shown in Table 3, SpongeBob leads across all metrics. Relative to the best baseline on each metric, AC improves by 1.7% (22.15 vs. 21.79) and FVD decreases by 2.7% (338.62 vs. 347.85).

G User Study

We conduct a user study to validate SpongeBob's perceptual advantages as a complement to automatic metrics. We randomly sample 30 test cases from SpongeBob-Bench (15 Speech-Video, 5 Sound-Video, 10 Complex Scene) and compare SpongeBob with all five baselines. Twenty evaluators with audio-visual professional backgrounds independently view all results in randomized anonymous order and rate each on a 1–5 scale across four dimensions: **AV-Sync** (temporal alignment naturalness), **Audio-Q** (clarity and realism), **Context** (whether new sounds conflict with preserved background audio and non-target speakers), and **Overall** (holistic editing quality).

Table 9: **User study (MOS, 1–5).** SpongeBob significantly outperforms all baselines ($p < 0.01$, paired t-test).

Method	AV-Sync↑	Audio-Q↑	Context↑	Overall↑
AvED	2.12	2.45	2.28	2.18
VACE-Foley	2.98	3.25	2.68	2.88
VACE-Coh.	3.36	3.63	3.41	3.40
AVI-Edit	3.42	3.65	3.28	3.38
InstructAV2AV	3.66	3.74	3.30	3.55
Ours	**4.28**	**4.15**	**4.32**	**4.25**

SpongeBob significantly outperforms all baselines across all dimensions ($p < 0.01$, paired t-test). Compared with the strongest baseline InstructAV2AV, the advantage is most pronounced on Context (+1.02), mirroring the Ctx-F1 gap, while the AV-Sync margin narrows to +0.62 because InstructAV2AV also denoises the two modalities jointly. Notably, VACE-Coherent scores above VACE-Foley

on all four dimensions, with the widest gap on Context (3.41 vs. 2.68), in line with its higher Sync-C, PQ, and Ctx-F1 in automatic metrics: conditioning on the source track respects the existing acoustic scene, whereas generating audio from scratch does not.

H Additional Analyses

H.1 Event-Level Synchronization on Non-Speech Content

IB measures semantic correspondence between audio and visual content and does not reflect their temporal alignment. We therefore restrict IB to semantic consistency and add an event-level temporal evaluation on the 100-sample Sound-Video subset, using the official AudioSet-pretrained Synchformer (Iashin et al., 2024) under MMAudio's (Cheng et al., 2025) public av-benchmark protocol. As shown in Table 10, SpongeBob attains the lowest DeSync at 0.58 s, reducing the predicted temporal offset by 21.6% relative to the strongest baseline, InstructAV2AV. The two metrics yield the same ranking but support different conclusions: IB verifies semantic correspondence, whereas DeSync estimates event-level temporal alignment.

Table 10: **Event-level synchronization, base-audio preservation, and inference cost.** IB and DeSync (the temporal offset predicted by Synchformer) are measured on the Sound-Video subset; LPAPS compares the base tracks separated from the source and the edited audio; time is measured on a single H20 GPU with identical samples and output duration, averaged over 50 samples.

		Sound-Video subset		Base audio	Cost
Method	Composition	IB↑	DeSync (s)↓	LPAPS↓	Time (s)↓
AvED	Stable Diffusion 2.1 + AudioLDM2	0.19	1.28	1.94	412
VACE-Foley	VACE-14B + HunyuanVideo-Foley	0.22	0.96	2.22	1218
VACE-Coh.	VACE-14B + Coherent	0.25	0.80	1.98	1220
AVI-Edit	Stable Audio Open + AVI-Edit	0.24	0.82	1.88	456
InstructAV2AV	Ovi-initialized dual-stream DiT	0.25	0.74	1.85	436
Ours	Dual-stream audio-video DiT	**0.28**	**0.58**	**1.81**	588

H.2 Robustness to Imperfect Audio Separation

SpongeBob consumes a base audio track produced by automatic source separation, which is never error-free. A typical failure is that part of the target sound remains in the base track. Because base audio conditions target-audio denoising through Acoustic Context Attention, the model is exposed to such correlated leakage during training and may tolerate it at test time.

To quantify this, we align the RMS level of the source target track a_{tgt} with that of the base audio, obtaining $\tilde{a}_{\text{tgt}}$, and construct a leaked base track $b_\alpha = b + \alpha\,\tilde{a}_{\text{tgt}}$ followed by peak normalization, so that α is the amplitude ratio of the injected target sound relative to the base audio. The same trained checkpoint is evaluated on all 700 test samples without adaptation. Table 11 shows that performance is essentially unchanged up to 20% leakage and degrades gradually beyond it.

Table 11: **Robustness to target-sound leakage in the base audio.** α is the amplitude ratio of the injected target sound; α=0% is the standard setting.

Leakage α	PQ↑	CLAP↑	Sync-C↑	Ctx-F1↑	G-Score↑
0%	6.45	0.238	4.50	0.81	7.6
10%	6.44	0.237	4.51	0.81	7.6
20%	6.38	0.234	4.42	0.80	7.4
30%	6.24	0.227	4.29	0.75	7.0
40%	5.91	0.218	4.11	0.68	6.2
50%	5.68	0.204	3.89	0.60	5.5

Table 12: **Long-video editing via overlapped sliding windows.** 50 videos of 10–20 s. Boundary similarities are compared against non-boundary adjacent frames.

Metric	Reference	Long video
Sync-C ↑	4.33	4.21
FAD ↓	2.31	2.38
Boundary DINOv2 similarity ↑	0.95	0.93
Boundary ArcFace identity ↑	0.92	0.90

H.3 Preservation of the Base Audio

Our task expects the target audio to change while the base audio is preserved, so an LPAPS (Iashin & Rahtu, 2021) computed over the full track would conflate a successful target edit with damage to preserved content. We therefore measure LPAPS on the base track alone: using the same SAM-Audio checkpoint and separation settings, we apply the corresponding target prompt to extract base-audio tracks from both the source and the edited audio and compare them. As reported in Table 10, SpongeBob preserves the base audio best despite regenerating the target. Part of this advantage comes from our design, which adds the base audio back in the waveform domain rather than regenerating it. Since the metric absorbs separation error from both sides, we use it only for relative comparison among these methods.

H.4 Inference Cost

All methods are timed on a single H20 GPU with identical test samples and output duration, excluding model loading and averaged over 50 samples (Table 10). Like InstructAV2AV, the other end-to-end model, SpongeBob edits audio and video in one pass rather than chaining a video editor with an audio generator, and therefore costs less than half the time of the two cascades built on VACE-14B. Against InstructAV2AV and the two lighter pipelines, AVI-Edit and AvED, it is roughly 1.3–$1.4\times$ slower, but produces substantially better edits (Sync-C 4.50 versus 3.78, 3.45, and 1.15; Ctx-F1 0.81 versus 0.73, 0.72, and 0.52). The overhead comes almost entirely from two-stage SPTG guidance, which needs 140 forward passes against 100 for standard CFG; scaled by forward-pass count, SpongeBob under standard CFG would take roughly 420 s, on par with InstructAV2AV. The cost is therefore a tunable property of the guidance schedule rather than of the architecture, and training-free DiT caching, sparser activation of the Stage-2 sync branch, and step distillation all offer further headroom.

H.5 Long-Video Editing

Training clips contain at most 121 frames, so longer inputs are processed with overlapped sliding-window inference: the overlapping frames of the previous window condition the next one, and audio latents are cross-faded at segment boundaries. We collected 50 videos of 10–20 s and compare against a single-window reference computed on the first 121 frames of the same videos. As shown in Table 12, Sync-C drops only from 4.33 to 4.21 and Fréchet Audio Distance (FAD) is essentially unchanged, while DINOv2 visual similarity and ArcFace identity similarity at window boundaries stay close to those measured on non-boundary adjacent frames. Under manual inspection 48 of 50 videos show no visible or audible seam; the remaining two exhibit slight identity drift when the speaker turns their head exactly at a boundary, consistent with the boundary-consistency risk noted in Appendix J.

I Data Pipeline Details

The main paper (Section 3.4) outlines the six-stage pipeline. Here we supplement key technical details.

Fine-grained acoustic category taxonomy. We design 50+ fine-grained categories covering animals (cat/dog/bird/frog breeds), instruments (string/wind/percussion/ethnic), and human speech. Each category precisely describes the sounding subject type and acoustic characteristics, used for candidate video retrieval and driving subsequent separation.

Quality verification. Gemini executes all quality assessments in an "LLM-as-Judge" paradigm across five dimensions: match score ($\geq$5, semantic consistency), completeness ($\geq$5, full extraction), quality ($\geq$5, no artifacts), leakage score ($\geq$ threshold, no target leakage in base audio), and ASR validity (speech scenes only). Only samples passing all criteria enter the final dataset. To verify that this automated filtering is reliable, we manually reviewed 200 samples drawn at random from the filtered training set: 191 passed the audio-separation check (95.5%), 196 passed the mask check (98.0%), and 187 passed both (93.5%). The remaining samples carry small separation errors, so the model sees realistically imperfect base audio during training rather than an oracle separation, which is consistent with the leakage tolerance reported in Table 11.

Instance segmentation. Source description text → Grounding DINO first-frame detection (box threshold 0.4, text threshold 0.3) → SAM2 first-frame segmentation → SAM2 full-video mask propagation → Gemini mask quality verification.

Dataset statistics. The final dataset contains approximately 400K variable-length samples of at most 121 frames, totaling 390 hours at 24 FPS (about 3.5 s per sample on average). Speech samples account for approximately 60% and animal/instrument/environmental samples for 40%, spanning 50+ fine-grained categories.

J LIMITATIONS

(1) **Cross-category generalization boundary.** Training is based on mask-and-reconstruct self-supervision where the model only sees same-category reconstruction. Cross-category editing (e.g., dog→cat) relies on compositional generalization of text conditions and reference images; quality may degrade when visual appearance differs substantially. (2) **Inference overhead.** The two-stage SPTG guidance requires 140 total forward passes (vs. 100 for standard CFG), increasing inference time by approximately 40%, which may bottleneck real-time or interactive applications. (3) **Long video limitation.** Training clips are limited to 121 frames (approximately 5 s); longer videos require segmented processing, potentially introducing audio-visual discontinuities at segment boundaries.