EDIT IN SYNC: UNIFIED END-TO-END AUDIO-VISUAL EDITING FOR HUMAN-CENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-centric audio-visual editing requires precise alignment of lip movements, gestures, and expressions with speech. Existing methods process inputs through independent video and audio pipelines, preventing cross-modal interaction between visual editing and audio dubbing. Many of these pipelines further rely on mask-based inpainting. As a result, they suffer from poor lip-audio synchronization, cumulative pipeline errors, and an inherent mask-size dilemma. We present AVE, the first unified end-to-end framework that jointly performs reference-guided person swapping, insertion, and deletion across both audio and video modalities without requiring spatial masks. To overcome data scarcity, we design a synthesis pipeline and introduce AVE-Dataset, the first audio-visual editing dataset to provide 100K pairs with appearance and voice references. By combining Synthetic-to-Real reconstruction with preliminary audio warm-up, our framework transcends synthetic pipeline limits and mitigates audio-video token imbalance. AVE outperforms modality-separate audio-visual editing pipelines in editing realism, temporal consistency, and audio-visual alignment.

1 INTRODUCTION

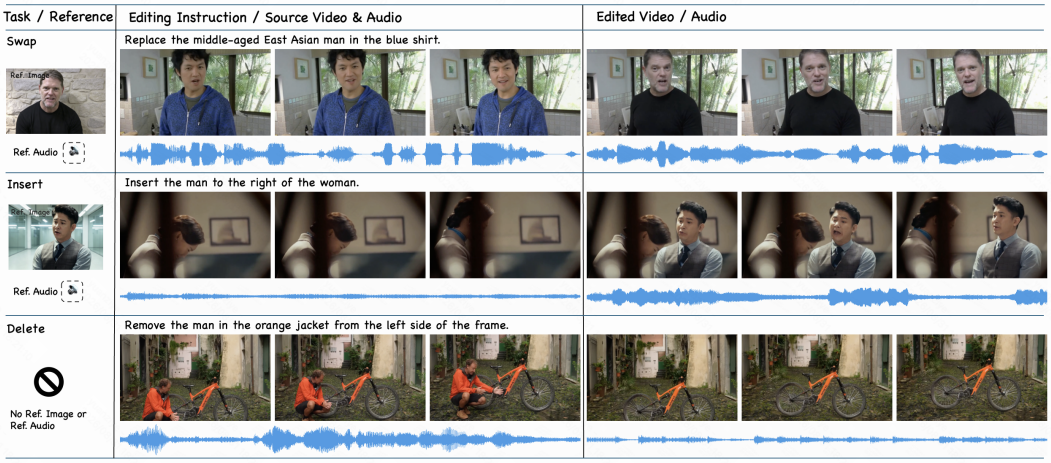

Figure 1: Overview of AVE across person swapping, insertion, and deletion. Appearance and voice references are provided for swapping and insertion, while deletion requires no reference.

Content editing has long been an important topic in both visual and audio fields. Representative visual editing tasks include object removal (Zhou et al., 2023; Zi et al., 2026; Fu et al., 2026; Miao et al., 2026), insertion (Bai et al., 2024; Tu et al., 2025; Zhao et al., 2025a; Jin et al., 2025), and replacement (Pan et al., 2025; Fu et al., 2025; Xu et al., 2026), while audio editing commonly involves voice conversion (Li et al., 2023; Liu, 2024), source separation (Subakan et al., 2021; Liu et al., 2024; Lu et al., 2024), and speech content editing (Peng et al., 2024; Ren et al., 2026). Despite progress in both modalities, independent audio-visual editing remains insufficient for human-centric scenarios where visual events and speech are intrinsically coupled across lip movements, facial expressions,

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.

and bodily actions (Lin et al., 2024; He et al., 2025). Human-centric editing such as insertion, removal, and replacement should be reflected coherently in both modalities, with consistent subject presence, identity, and timing (Guo et al., 2026; Liang et al., 2026). Thus, joint human-centric audio-visual editing has emerged as a key multimodal frontier and the primary focus of this work.

Existing approaches to audio-visual editing typically decouple video and audio processing into independent pipelines (Ishii et al., 2025; Wu et al., 2025b). Consequently, the visual editor remains speech-agnostic while the audio stage lacks visual guidance; this absent reciprocal conditioning degrades synchronization and compounds pipeline errors (Fig. 2). Emerging joint audio-video generators prove that unified cross-modal modeling naturally leads to more synchronized and realistic outputs (Low et al., 2025; HaCohen et al., 2026). Inspired by these advances, we develop an end-to-end framework that unifies the core suite of human-centric operations into a single model.

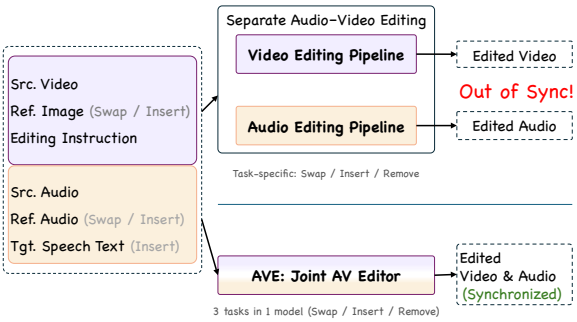

Figure 2: Separate editing pipelines versus AVE

However, training such a unified framework is severely constrained by data scarcity. End-to-end supervision requires paired audio-visual editing data, which is virtually non-existent in the wild and costly to collect manually at scale. A common workaround is to reformulate editing as a self-supervised mask-and-reconstruct task (Cheng et al., 2025; Guo et al., 2026; Liang et al., 2026), where the target subject and speech are masked to reconstruct the original content from the remaining context. Yet, such mask-dependent training imposes strict spatial constraints that often erase fine motion cues and surrounding context, causing visual artifacts and temporal instability. To overcome this bottleneck, we build a three-stage data construction pipeline. We leverage specialized visual and audio models to synthesize diverse editing pairs. To improve reference adherence across diverse visual and acoustic contexts, we construct cross-pair multimodal references by retrieving the target identity across distinct scenes or synthesizing it with generative augmentations, providing reliable identity and timbre guidance while preventing copy-and-paste shortcuts. Through multi-stage visual and acoustic filtering, we assemble AVE-Dataset, comprising 100K curated audio-visual pairs that support unified character *Delete*, *Insert*, and *Swap*.

Based on AVE-Dataset, we propose AVE, an end-to-end framework for human-centric joint audio-visual editing, supporting three tasks within a unified architecture. By taking complete source audio-video signals directly as input AVE preserves full spatiotemporal and acoustic context, producing harmoniously synchronized output under multimodal conditions. AVE builds on a joint audio-video backbone (Low et al., 2025), utilizing its bidirectional cross-modal attention for synchronized editing across modalities. To handle complex multimodal inputs, we introduce a Modality-Aware Sync-RoPE design that assigns role-specific positional coordinates to source, target, and reference tokens. This preserves source-target temporal correspondence while aligning audio and video along a shared timeline. Furthermore, we introduce a multi-stage training curriculum with bidirectional supervision: forward supervision (from a real source to a synthetic target) teaches the intended editing operations, while the backward direction exposes the model to real target distributions. Together with audio warm-up, bidirectional joint adaptation, and quality-focused refinement, this curriculum mitigates synthesis artifacts and enables the mask-free editor to bootstrap beyond the quality ceiling of its synthetic training data.

Our main contributions are as follows: (1) We present AVE, a unified mask-free framework for human-centric audio-visual editing. It employs Modality-Aware Sync-RoPE to enable comprehensive cross-modal interaction while preserving audio-visual temporal correspondence. (2) We develop a scalable data synthesis pipeline to construct AVE-Dataset, providing 100K verified audio-visual editing pairs equipped with appearance and voice references alongside a standardized benchmark. (3) We introduce a multi-stage training curriculum with bidirectional supervision, enabling the final model to progressively surpass the quality ceiling imposed by synthetic training data.

2 Related Work

Joint audio–video foundation models learn visual and acoustic content as a coupled distribution rather than separate streams. Early works explored coupled U-Nets for unconditional synthesis (Ruan et al., 2023), while recent DiT-based models scale cross-modal generation to open domains with strong multimodal priors (Low et al., 2025; HaCohen et al., 2026; Chen et al., 2026b). However, general generation capability alone does not naturally resolve the central editing problem: transferring requested attributes while preserving operation-irrelevant source content. Initial editing formulations adapted single-modality models to paired image–audio data (Liang et al., 2024) or relied on test-time optimization on video (Lin et al., 2026). Modular pipelines follow a decomposition that separately inverts, edits, or dubs audio and video (Fu et al., 2025; Wu et al., 2025b; Ishii et al., 2025; Zheng et al., 2026a), yet these one-way cascades lack bidirectional interaction and inevitably compound pipeline errors. Recent works move toward coupled end-to-end architectures (Xu et al., 2025; Zheng et al., 2026b; Chen et al., 2026c; Liang et al., 2026), yet they remain restricted to single operations (Xu et al., 2025) or rely on mask-guided routing (Liang et al., 2026). In contrast, we combine external appearance and timbre references with mask-free localization to unify person-level *Swap*, *Insert*, and *Delete* within a single generative process.

In human-centric scenarios, existing studies have separately investigated visual identity transfer (Pan et al., 2025; Xu et al., 2026) and speech-driven facial animation (Manela et al., 2025; He et al., 2025). While achieving promising portrait fidelity and lip synchronization, they primarily operate on the visual stream or require external driving audio, failing to jointly transfer appearance and voice. Conversely, joint dubbing and audio-video personalization methods couple appearance and voice for identity-consistent generation (Chen et al., 2026a; Li et al., 2026b; Chen et al., 2026d; Dahan et al., 2026; Chen et al., 2026e; Li et al., 2026a), but they synthesize new scenes from scratch rather than preserving source-video context. DreamID-Omni (Guo et al., 2026) represents the closest effort by supporting reference-based generation and person replacement. Nonetheless, it does not unify person insertion and deletion under a consistent editing framework, nor does it preserve source motion and linguistic context as editing invariants. In contrast, our approach enables holistic identity and timbre control across all three operations.

Audio editing provides essential capabilities for manipulating acoustic content in human-centric scenes. Zero-shot voice cloning synthesizes speech in an arbitrary timbre for character insertion (Du et al., 2024; Qin et al., 2023), while voice conversion preserves linguistic timing and prosody for character replacement (Liu, 2024). In parallel, speech separation and extraction isolate target voices from acoustic mixtures to facilitate character deletion (Shi et al., 2025; Wu et al., 2025b). However, these audio-centric methods cannot exploit visual cues to resolve speaker attribution and lip synchronization in talking videos, nor can separation alone reconstruct vacated visual regions. Our framework integrates these editing capabilities into a unified multimodal backbone, allowing acoustic and visual decisions to co-evolve seamlessly.

3 Method

We study instruction-guided audio-visual editing for talking videos across three person-level operations: *Swap*, *Insert*, and *Delete*. Given a source video $\mathbf{v}_s$, its synchronized audio $\mathbf{a}_s$, a natural-language instruction c, and optional multimodal references $(\mathbf{r}_v, \mathbf{r}_a)$ providing the target appearance and vocal timbre (with $\mathbf{r}_v, \mathbf{r}_a = \varnothing$ for *Delete*), our goal is to predict the edited audio-visual pair $(\hat{\mathbf{v}}, \hat{\mathbf{a}})$ using a single unified model f_θ:

$$(\hat{\mathbf{v}}, \hat{\mathbf{a}}) = f_\theta(\mathbf{v}_s, \mathbf{a}_s, \mathbf{r}_v, \mathbf{r}_a, c), \tag{1}$$

The core challenge lies in editing both modalities jointly without explicit spatial signals: f_θ must implicitly resolve the target referent from c, maintain precise lip-speech alignment across modalities, and faithfully preserve non-target context.

Building on this formulation, we present AVE. In this section, we first describe our data synthesis pipeline for obtaining complex editing pairs in **Sec. 3.1**. Then we present our editing framework with Modality-Aware Sync-RoPE design in **Sec. 3.2**. Finally, Sec. 3.3 introduces a progressive training curriculum with bidirectional supervision to overcome the quality ceiling of synthetic training data.

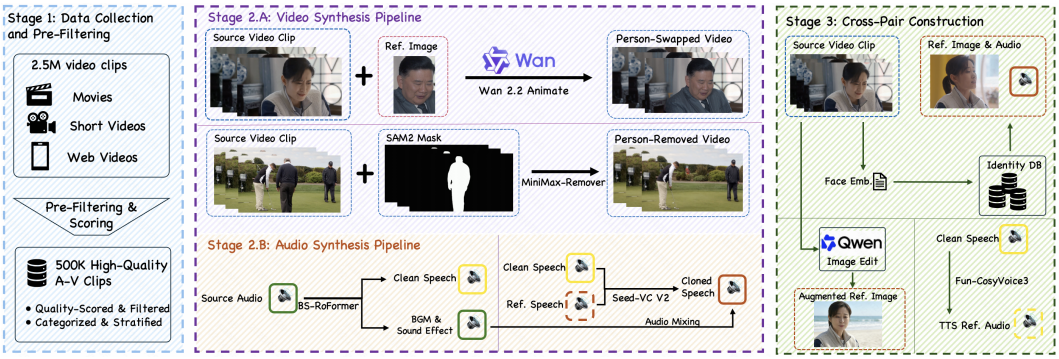

Figure 3: Automated data construction pipeline of AVE-Dataset. The workflow consists of three phases: (1) raw corpus collection and quality pre-filtering, (2) video-audio synthesis for editing pair creation, (3) identity retrieval and multimodal augmentation for cross-pair reference construction.

3.1 Data

To obtain scalable supervision for joint audio-visual editing, we build an automatic data construction pipeline illustrated in Fig. 3. In Phase 1, we filter 2.5M raw video clips to retain a 500K high-quality, dialogue-rich source corpus (see Appendix A.1 for details). In Phase 2, we synthesize paired examples for *Swap*, *Insert*, and *Delete*, and verify the resulting samples for visual, acoustic, and synchronization fidelity. In Phase 3, we construct cross-pair appearance and voice references through identity retrieval and augmentation, ensuring that the references preserve the desired identity and timbre while differing in pose, scene, and linguistic content. Additionally, structured editing instructions are annotated using Qwen3.5 (Qwen Team, 2026) (detailed in Appendix A.2) to guide cross-modal editing intents. To guarantee dataset fidelity, a multi-stage verification suite (detailed in Appendix A.3) rigorously filters out visual artifacts, lip-speech misalignments, and identity-timbre mismatches throughout the pipeline. The resulting AVE-Dataset contains 100K curated audio-visual editing pairs, including 40K *Swap* pairs, over 35K *Insert* pairs, and over 20K *Delete* pairs.

3.1.1 Edit Pair Synthesis

We construct high-quality editing pairs with aligned audio and video through a multi-stage synthesis and filtering pipeline (Fig. 3, Stage 2). The pipeline has two branches: *Swap*, and the complementary *Delete*/*Insert* operations. The resulting editing pairs support both distillation from synthetic targets and reconstruction of real targets.

Swapping. Person swapping replaces the appearance and voice timbre of a target person in the source video with those of a reference individual. It should faithfully preserves the original motion and poses, speech timing, and background audio. For video synthesis (Stage 2.A), we employ Wan 2.2 Animate (Cheng et al., 2025) with filtered source video clips serving as motion and pose drivers, combining them with reference facial images to synthesize videos with replaced appearances (*Person-Swapped Video*). For audio synthesis (Stage 2.B), we first utilize BS-RoFormer (Lu et al., 2024) to precisely separate the source audio into clean speech and background noise/sound effect tracks. Subsequently, Seed-VC V2 (Liu, 2024) performs voice conversion using the isolated speech for content and prosody guidance with the reference timbre. The cloned speech is then remixed with background audio, yielding a target track synchronized with the edited video.

Removal & Insertion. Removal and insertion are inverse audio-visual editing operations. For video-level erasing, we track the target person in the source video using SAM 2 (Ravi et al., 2025) to generate frame-wise binary masks, which are then fed into MiniMax-Remover (Zi et al., 2026) for inpainting to yield a background video with the target person removed (*Person-Removed Video*). For audio-level separation, we similarly strip the target voice via BS-RoFormer, retaining only ambient background sounds and audio effects (or silence if no background audio exists). From the resulting video and audio, we construct bidirectional editing pairs: 1) **Removal**: The original audio-visual clip serves as the source, while the person-removed video and background audio serve as the target; 2)

Insertion: We reverse the pair and provide reference images and audio of the person to be inserted, using the original clip as the target. Thus, insertion is supervised by real, synchronized audio-visual targets.

3.1.2 Cross-Pair Reference Construction

For reference-conditioned backward (using synthetic clips as inputs and real clips as targets) reconstruction, we need references for the original person. Simply cropping the real target would expose its pose, scene, or speech content, encouraging copy-and-paste shortcuts. Therefore, we construct cross-pair visual and audio references that match the target's identity and clothing but differ in background, pose, expression. For visual references, we sample and crop one face per second from each video. We then compute face embeddings and use them to index the crops in an identity retrieval database. We retrieve candidate references of the same identity and further verify them through speaker-timbre similarity. After suppressing background sounds and sound effects, the corresponding speech segments serve as candidate audio references. Qwen3.5 rejects neighboring excerpts from the same continuous scene and candidates with inconsistent clothing, accessories, or hairstyles, which commonly arise when the same actor appears in different productions. We complement retrieved references with generative augmentation via Qwen-Image-Edit (Wu et al., 2025a) and CosyVoice 3 (Du et al., 2024). For visual augmentation, an image editor swaps the subject's background with samples from a background library while introducing subtle variations in expression and pose. For audio augmentation, CosyVoice 3 synthesizes new utterances from a text corpus using clean speech segments as voice prompts. All synthesized references undergo a final quality check (detailed in Appendix A.3).

3.2 Framework of AVE

With the paired data in place, we build our editor on a joint audio–video generation backbone, so that editing inherits a strong prior on how speech, lip motion, and ambient sound co-occur instead of having to relearn cross-modal consistency from the editing pairs alone. Our design relies on only three properties shared across this family of models (Low et al., 2025; HaCohen et al., 2026; Chen et al., 2026b): separate video and audio VAEs map the two streams into their own latent spaces; two Diffusion Transformer (DiT) branches are coupled by bidirectional cross-modal attention, so that each modality conditions on the other at every block; and a shared frozen text encoder supplies the semantic embedding of the editing instruction. Because we assume nothing beyond these properties, our method is not tied to a particular backbone. We instantiate it with Ovi (Low et al., 2025) and adapt the conditional DiT blocks to source-conditioned editing.

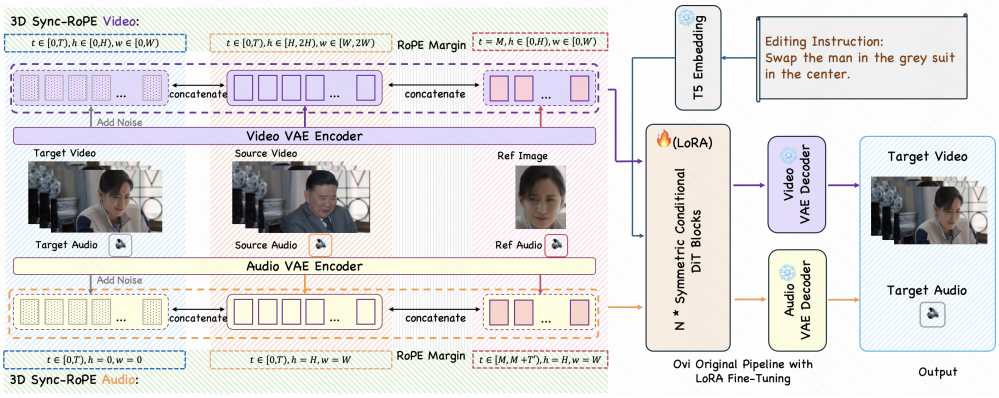

Figure 4: Overview of the AVE framework. To prevent token ambiguity, our Modality-Aware Sync-RoPE introduces role-specific 3D coordinate offsets that explicitly distinguish source, target, and reference tokens while maintaining temporal correspondence along a shared timeline.

3.2.1 SOURCE- AND REFERENCE-CONDITIONED EDITING

As illustrated in Fig. 4, the source and reference inputs are encoded as clean context, whereas the target video and audio streams are initialized from noise. Within each modality, the target, source, and reference tokens are concatenated and jointly processed by the coupled DiT backbones. After joint denoising, the video and audio decoders produce the edited pair.

3.2.2 MODALITY-AWARE SYNC-ROPE

Inspired by DreamID-Omni (Guo et al., 2026), we design a Modality-Aware Sync-RoPE layout that distinguishes the roles of target, source, and reference tokens without disrupting source–target temporal correspondence. Both streams share the same three coordinate axes, but each modality interprets them according to its own structure. Let T, H, and W denote the temporal, height, and width dimensions of the video latent grid. We assign the target and source video tokens to

$$\Omega^v_{\text{tar}} = [0, T) \times [0, H) \times [0, W), \qquad \Omega^v_{\text{src}} = [0, T) \times [H, 2H) \times [W, 2W). \tag{2}$$

Thus, corresponding source and target frames share the same temporal coordinates, while offsets along both spatial axes prevent their tokens from being interpreted as the same latent canvas. For an optional reference image, we use

$$\Omega^v_{\text{ref}} = \{M\} \times [0, H) \times [0, W), \tag{3}$$

where the RoPE margin M separates reference information from the source and target sequences.

We apply the same principle to the audio stream. Let T_a be the number of audio latent tokens; following Ovi (Low et al., 2025), we scale their temporal RoPE frequencies by $\gamma = T/T_a$, such that $\gamma T_a = T$ and the audio and video streams span the same temporal RoPE range. Since audio has no intrinsic spatial grid, we use the two auxiliary RoPE axes to encode token roles:

$$\Omega^a_{\text{tar}} = [0, T_a) \times \{0\} \times \{0\},\ \Omega^a_{\text{src}} = [0, T_a) \times \{H\} \times \{W\},\ \Omega^a_{\text{ref}} = [M, M + T_a) \times \{H\} \times \{W\} \tag{4}$$

The source and target audio tokens therefore remain temporally aligned, while the reference audio is isolated by the same margin. We retain Ovi's (Low et al., 2025) temporal RoPE scaling between the audio and video branches so that their different token rates correspond to the same physical timeline.

3.3 PROGRESSIVE TRAINING

3.3.1 BIDIRECTIONAL SUPERVISION

Training only in the forward direction (from a real source to a synthetic target) would primarily distill the capabilities and artifacts of the data-synthesis pipeline. We therefore additionally reverse each editing-pair: the synthesized sample is used as the source and its real counterpart as the supervision target. This backward direction exposes the model to real target distributions and improves generation quality. We train primarily on backward pairs, supplemented by forward pairs. Forward supervision teaches the intended edits from real source videos without compromising the pretrained generation prior, whereas backward supervision improves output fidelity without encouraging the model to exploit residual synthesis artifacts as shortcuts. Our stringent filtering makes the synthetic targets used in forward training and the synthetic inputs used in backward training closely approximate real audio–video data.

3.3.2 MULTI-STAGE TRAINING CURRICULUM

Stage 0: Audio Warm-up. We first isolate the audio branch and train it on audio-only synthetic data for TTS, voice conversion, and speaker-timbre disentanglement. This warm-up is critical for mitigating the severe audio–video token imbalance and substantially accelerates subsequent joint optimization, allowing the audio editor to be efficiently integrated into the joint model.

Stage 1: Bidirectional Joint Adaptation. We train with both forward and backward directions. We first freeze the audio branch and optimize the video and cross-modal fusion modules on joint audio-video data, and subsequently unfreeze the audio branch to jointly optimize all trainable modules. This enables the model to acquire all three editing operations within a unified framework.

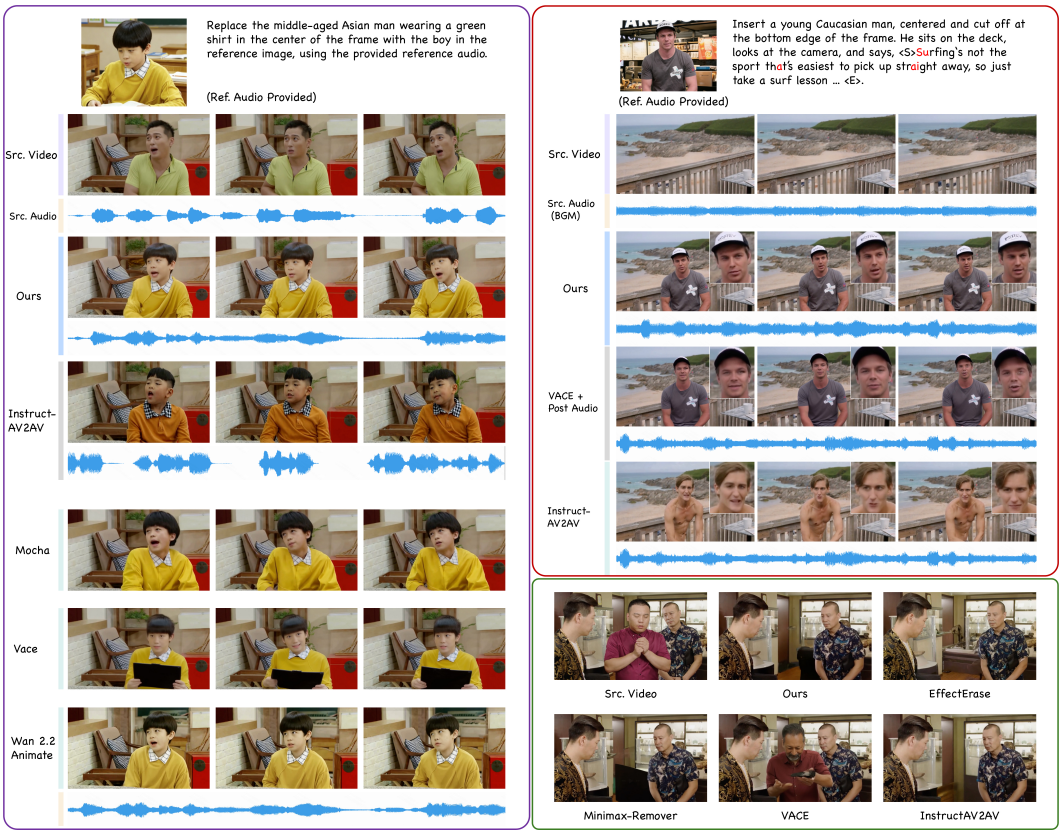

Figure 5: Qualitative comparisons across person swapping, insertion, and removal. Please zoom in for more details.

Stage 2: Quality Refinement. We select the highest-quality subset from each data category based on VBench (Huang et al., 2024), SyncNet (Chung & Zisserman, 2016), and other quality scores, retaining only samples with retrieved cross-pair references for continued backward training. This quality-focused stage improves local details and reduces the residual artifacts from Stage 1.

4 Experiments

4.1 Experimental Setup

Implementation Details. We fine-tune Ovi (Low et al., 2025) using LoRA (Hu et al., 2021); trainable modules are detailed in Appendix B. Training videos are standardized to a resolution of 480×832 at 24 fps with a duration of 5.04 seconds (121 frames), aligning with Ovi's temporal grid. Audio is sampled at 16 kHz. For Modality-Aware Sync-RoPE, we set the shared RoPE margin to $M = 350$ for both the video and audio branches. Following the curriculum in Sec. 3.3: In Stage 0, we train the audio-only LoRA parameters with a batch size of 512 for 2k steps. In Stage 1, we train on audio-video pairs with a batch size of 64, running 2k steps with frozen audio LoRA parameters followed by 2k steps of joint optimization across all LoRA parameters. In Stage 2, we continue backward training on the top-50% quality-refined subset with a batch size of 64 for 2k steps.

Benchmark Evaluation. Publicly available benchmarks closest to our setting remain fragmented. InstructAV2AV and JAVEdit provide multi-operation evaluation sets based on text instructions, without external appearance or voice references (Zheng et al., 2026b; Chen et al., 2026c). IDBench-Omni provides reference-conditioned evaluation, but its released source-video editing set comprises only 25 full-human-swap cases and 25 face-swap cases, with target dialogue specified by text prompts; person insertion and deletion into existing source videos are not evaluated (Guo et al., 2026).

Table 1: Quantitative comparison on AVE-Dataset-Bench across three editing tasks. Reconstruction metrics are computed over the background region. Best results are shown in bold. "+PA" denotes cascaded baselines paired with our audio post-processing pipeline.

	VBench (%)					Audio Quality & AV Sync			Ref Similarity			BG Fidelity	
Task / Method	Subj. Cons. ↑	BG Cons. ↑	Temp. Flick. ↑	Motion Sm. ↑	Imaging Q. ↑	PQ ↑	Sync-C ↑	Sync-D ↓	CAM++ ↑	Arc ↑	CLIP ↑	PSNR ↑	LPIPS ↓
Swap													
VACE+PA	95.26	94.30	96.48	98.11	68.13	6.586	4.86	9.29	0.665	0.614	0.705	**35.37**	**0.0481**
Wan2.2+PA	94.09	**96.12**	97.76	97.71	66.41	6.586	5.02	8.76	0.665	0.634	0.710	33.27	0.0505
MoCha+PA	96.52	93.16	**97.97**	97.30	65.11	6.586	5.05	8.72	0.665	0.639	**0.712**	32.42	0.0584
InstructAV2AV	95.86	93.88	95.56	98.10	66.60	6.255	3.87	9.65	–	–	–	31.84	0.0640
Ours	**97.35**	95.69	97.67	**98.54**	**69.74**	**6.709**	**5.39**	**8.53**	**0.682**	**0.647**	**0.712**	34.62	0.0549
Delete													
VACE+PA	93.06	92.72	95.61	97.59	65.64	**6.813**	–	–	–	–	–	29.94	0.1129
MiniMax-Remover+PA	97.39	95.90	97.97	97.98	59.07	**6.813**	–	–	–	–	–	27.99	0.1317
Effect Erase+PA	97.69	96.24	98.04	96.20	62.65	**6.813**	–	–	–	–	–	31.27	0.1212
InstructAV2AV	96.53	93.82	96.32	97.56	60.63	6.573	–	–	–	–	–	29.74	0.0991
Ours	**98.49**	**96.64**	**98.12**	**98.43**	**67.15**	6.810	–	–	–	–	–	**31.74**	**0.0926**
Insert													
VACE+PA	96.04	94.71	97.61	97.70	68.20	6.384	3.93	9.73	0.638	0.614	**0.705**	28.66	0.1271
InstructAV2AV	95.71	93.52	**98.10**	**98.63**	63.72	6.090	3.84	10.60	–	–	–	**29.38**	**0.0950**
Ours	**98.36**	**96.87**	98.05	98.44	**72.82**	**6.491**	**4.43**	**8.83**	**0.651**	**0.630**	0.698	27.70	0.1023

To address these gaps, we construct AVE-Dataset-Bench, comprising held-out video clips strictly disjoint from our training set. Specifically, AVE-Dataset-Bench contains 64 carefully curated test samples for each of the three editing tasks, yielding 192 test cases in total. Models receive real-world audio-visual clips as source inputs, paired with reference images and audio as conditioning signals.

Baselines. We compare AVE with two categories of baselines: separate audio-video editing pipelines and the text-driven joint audio-video editing model InstructAV2AV (Zheng et al., 2026b). For separate audio–video editing, each visual editor is paired with the corresponding task-specific audio editing pipeline. Specifically, we use VACE (Jiang et al., 2025) as the general-purpose visual baseline across all three tasks: *Swap*, *Delete*, and *Insert*. In addition, for *Swap*, we evaluate Wan-Animate (Cheng et al., 2025) and MoCha (Xu et al., 2026); for *Delete*, we evaluate MiniMax-Remover (Zi et al., 2026) and EffectErase (Fu et al., 2026). To benchmark against end-to-end joint editing, we further compare with InstructAV2AV. Since InstructAV2AV is conditioned solely on text instructions and does not natively accept external visual or vocal references, we convert the reference subject's salient appearance and vocal attributes into detailed text descriptions and append them to the instruction prompt.

Evaluation Metrics. We evaluate the edited results across five dimensions: (1) Visual quality: five VBench dimensions (Huang et al., 2024), including subject consistency, background consistency, temporal flickering, motion smoothness, and imaging quality; (2) Acoustic quality: AudioBox PQ (Tjandra et al., 2025) (↑) measuring the acoustic quality of generated speech; (3) Audio-visual synchronization: Sync-C (↑, confidence) and Sync-D (↓, feature distance) evaluated via SyncNet (Chung & Zisserman, 2016); (4) Reference fidelity: ArcFace (Deng et al., 2019) (↑) for facial identity similarity, CAM++ (Wang et al., 2023) (↑) for speaker timbre similarity, and CLIP-I (Radford et al., 2021) (↑) for image-level subject similarity; (5) Background fidelity: PSNR (↑) and LPIPS (Zhang et al., 2018) (↓) computed over the background region. For *Swap* and *Delete*, the background is obtained by segmenting the edited subject via SAM 2 (Ravi et al., 2025) followed by a moderate mask dilation. For *Insert*, the inserted subject is segmented from the edited output, and the unmasked region is taken as the background.

4.2 Main Results

Table 1 summarizes the quantitative comparisons across the three editing tasks, and Figure 5 shows representative qualitative examples. Overall, AVE delivers competitive visual and acoustic quality, strong audio–visual synchronization, and high reference fidelity.

Swapping. In Figure 5, AVE seamlessly replaces the target subject with the reference child. In contrast, VACE exhibits lip-speech misalignment, Wan-Animate alters the background layout, and MoCha degrades facial fidelity. InstructAV2AV modifies appearance via text prompts, but cannot

Table 2: Ablation study on AVE-Dataset-Bench across tasks. Best results are shown in bold.

	VBench (%)					Audio Quality & AV Sync		
Task / Variant	Subj. Cons. ↑	BG Cons. ↑	Temp. Flick. ↑	Motion Sm. ↑	Imaging Q. ↑	PQ ↑	Sync-C ↑	Sync-D ↓
Swap								
w/o Audio Pretrain	96.60	95.34	**97.94**	97.95	68.58	6.107	4.71	9.25
w/o Multistage	95.90	94.57	97.28	98.21	67.88	6.580	5.28	8.66
Ours	**97.35**	**95.69**	97.67	**98.54**	**69.74**	**6.709**	**5.39**	**8.53**
Delete								
w/o Audio Pretrain	97.87	96.20	97.93	**98.44**	65.57	6.304	–	–
w/o Multistage	96.57	95.08	97.61	98.13	63.27	6.704	–	–
Ours	**98.49**	**96.64**	**98.12**	98.43	**67.15**	**6.810**	–	–
Insert								
w/o Audio Pretrain	96.75	94.95	98.37	97.97	68.21	5.980	3.54	9.75
w/o Multistage	96.69	95.21	**98.47**	97.60	68.22	6.320	4.15	9.18
Ours	**98.36**	**96.87**	98.05	**98.44**	**72.82**	**6.491**	**4.43**	**8.83**

recover the specific reference identity. Quantitatively, AVE achieves the highest identity fidelity (ArcFace 0.682, CAM++ 0.647) and best audio–visual synchronization (Sync-C 5.39, Sync-D 8.53).

Insertion. For person insertion, AVE generates a new speaker with precise lip synchronization on key syllables and clear facial details. VACE suffers from blurred eyes and poor lip sync, while InstructAV2AV produces an unnatural pose and disobey the prompt. In quantitative evaluation, AVE substantially improves audio–visual alignment (Sync-C 4.43, Sync-D 8.83 versus 3.93, 9.73 for VACE+PA) and achieves a higher imaging quality score (72.82).

Removal. For subject removal, AVE naturally infills the vacated region without explicit masks. EffectErase generates discordant background textures, MiniMax-Remover leaves black artifacts, VACE fails to delete the target, and InstructAV2AV distorts the adjacent person while blurring the background. AVE achieves both the best background preservation (PSNR 31.74, LPIPS 0.0926) and the highest imaging quality (67.15).

4.3 Ablation Study and Analysis

We conduct ablation studies on AVE-Dataset-Bench to validate our core training strategies, additional ablation of Modality-Aware Sync-RoPE is provided in Appendix C.

Effect of Audio Warm-up (Stage 0). To verify the necessity of isolated audio warm-up, we compare our full model against a variant that skips Stage 0 and directly initiates joint training from Stage 1 through Stage 2 (with extra training in Stage 1 to ensure equal total training iterations). As demonstrated in our experiments, omitting Stage 0 causes severe cross-modal token imbalance and leads to degraded speech quality (AudioBox PQ) and poorer AV synchronization (Sync-C), confirming that preliminary audio warm-up provides crucial acoustic priors for joint optimization.

Effect of Multi-Stage Refinement (Stage 2). We evaluate the impact of Stage 2 quality refinement by comparing against a variant that continues training exclusively on Stage 1 data for an equal number of iterations without quality filtering. As shown in Table 2, omitting Stage 2 leads to a performance degradation, particularly in visual quality and consistency metrics evaluated by VBench.

5 Conclusion

We present AVE, an end-to-end framework for mask-free human-centric audio-visual editing that unifies person *Swap*, *Insert*, and *Delete* within a single generative process while preserving external appearance and voice reference control. To overcome data scarcity, we construct AVE-Dataset, a 100K-pair audio-visual editing dataset with multimodal references, and train the model primarily with backward reconstruction to leverage authentic targets as supervision. Experimental results across all three tasks demonstrate that AVE achieves stronger editing fidelity, temporal stability, and lip-speech synchronization than cascaded editing baselines.

Reproducibility Statement

To ensure the reproducibility of our results, all key architectural designs, mathematical formulations, and coordinate assignments of Modality-Aware Sync-RoPE are presented in Section 3.2 and Section 3.3. Detailed training hyperparameters, stage schedules, and optimization configurations are provided in Appendix B. The data curation workflow, filtering criteria, and benchmark protocols are described in Section 3.1 and Appendix A. The source code, evaluation benchmark (AVE-Dataset-Bench), and model weights will be made publicly available for research purposes upon publication.

Ethics Statement

This work explores human-centric audio-visual editing for creative content generation and multimedia post-production. We recognize the potential risks of identity manipulation, deepfake generation, and unauthorized voice cloning. To mitigate these risks, all video and audio datasets used in this research were gathered exclusively from publicly available sources for academic research purposes. We strictly oppose malicious uses of identity manipulation and advocate the integration of synthetic media detection frameworks and digital watermarking alongside generative models.

AI Use Statement

In this work, generative foundation models and specialized tools were used as algorithmic components of our data curation and training pipeline: Qwen3.5 for instruction annotation and visual filtering, Wan-Animate and MiniMax-Remover for video synthesis, and BS-RoFormer, Seed-VC, and CosyVoice 3 for audio separation, voice conversion, and speech augmentation. Generative LLMs were also used solely for text polishing and proofreading of the manuscript. All synthesized datasets and final manuscript texts were carefully verified and audited by the authors. We take full responsibility for the content and artifacts presented in this paper.

References

Chen Bai, Zeman Shao, Guoxiang Zhang, Di Liang, Jie Yang, Zhuorui Zhang, Yujian Guo, Chengzhang Zhong, Yiqiao Qiu, Zhendong Wang, et al. Anything in any scene: Photorealistic video object insertion. *arXiv preprint arXiv:2401.17509*, 2024.

Anthony Chen, Naomi Ken Korem, Gal Zeevi, Tavi Halperin, Matan Ben Yosef, Urska Jelercic, Ofir Bibi, Or Patashnik, and Daniel Cohen-Or. Just-dub-it: Video dubbing via joint audio-visual diffusion. *arXiv preprint arXiv:2601.22143*, 2026a.

Guibin Chen, Dixuan Lin, Jiangping Yang, Youqiang Zhang, Zhengcong Fei, Debang Li, Sheng Chen, Chaofeng Ao, Nuo Pang, Yiming Wang, et al. SkyReels-V4: Multi-modal video-audio generation, inpainting and editing model. *arXiv preprint arXiv:2602.21818*, 2026b.

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*, 2026c.

Yingjie Chen, Shilun Lin, Cai Xing, Binxin Yang, Long Zhou, Qixin Yan, Wenjing Wang, Dingming Liu, Hao Liu, Chen Li, et al. Identity as presence: Towards appearance and voice personalized joint audio-video generation. *arXiv preprint arXiv:2603.17889*, 2026d.

Yuheng Chen, Qingdong He, Teng Hu, Yuji Wang, Yabiao Wang, Lizhuang Ma, and Jiangning Zhang. Omni-Customizer: End-to-end multimodal customization for joint audio-video generation. *arXiv preprint arXiv:2605.17488*, 2026e.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-Animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 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.

Aviad Dahan, Moran Yanuka, Noa Kraicer, Lior Wolf, and Raja Giryes. ID-LoRA: Identity-driven audio-video personalization with in-context lora. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. Effecterase: Joint video object removal and insertion for high-quality effect erasing. *arXiv preprint arXiv:2603.19224*, 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.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. DreamID-Omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xu He, Haoxian Zhang, Hejia Chen, Changyuan Zheng, Liyang Chen, Songlin Tang, Jiehui Huang, Xiaoqiang Liu, Pengfei Wan, and Zhiyong Wu. From inpainting to editing: A self-bootstrapping paradigm for context-rich visual dubbing. *arXiv preprint arXiv:2512.25066*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

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.

Hoiyeong Jin, Hyojin Jang, Jeongho Kim, Junha Hyung, Kinam Kim, Dongjin Kim, Huijin Choi, Hyeonji Kim, and Jaegul Choo. Insertanywhere: Bridging 4d scene geometry and diffusion models for realistic video object insertion. *arXiv preprint arXiv:2512.17504*, 2025.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Liyang Li, Wen Wang, Canyu Zhao, Tianjian Feng, Zhiyue Zhao, Hao Chen, and Chunhua Shen. MMControl: Unified multi-modal control for joint audio-video generation. In *European Conference on Computer Vision*, pp. 650–667. Springer, 2026a.

Maomao Li, Zhen Li, Kaipeng Zhang, Guosheng Yin, Zhifeng Li, and Dong Xu. OmniCustom: Sync audio-video customization via joint audio-video generation model. *arXiv preprint arXiv:2602.12304*, 2026b.

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 *Asian Conference on Computer Vision*, pp. 123–139. Springer, 2024.

Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*, 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.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE Transactions on Audio, Speech and Language Processing*, 33:458–471, 2024.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. Music source separation with band-split rope transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 481–485. IEEE, 2024.

Binyamin Manela, Sharon Gannot, and Ethan Fetyaya. Video editing for audio-visual dubbing. *arXiv preprint arXiv:2505.23406*, 2025.

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.

Tianrui Pan, Lin Liu, Jie Liu, Xiaopeng Zhang, Jie Tang, Gangshan Wu, and Qi Tian. RASA: Replace anyone, say anything—a training-free framework for audio-driven and universal portrait video editing. *arXiv preprint arXiv:2503.11571*, 2025.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12442–12462, 2024.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. OpenVoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Yong Ren, Jiangyan Yi, Jianhua Tao, Tao Wang, Le Xu, and Zhengqi Wen. Edit content, preserve acoustics: Imperceptible text-based speech editing via self-consistency rewards. *arXiv preprint arXiv:2602.00560*, 2026.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10219–10228. IEEE, 2023.

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.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25. IEEE, 2021.

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.

Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Kam Man Wu, Zeyue Tian, Liya Ji, and Qifeng Chen. VAInpaint: Zero-shot video-audio inpainting framework with llms-driven module. *arXiv preprint arXiv:2509.17022*, 2025b.

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. Schrödinger audio-visual editor: Object-level audiovisual removal. *arXiv preprint arXiv:2512.12875*, 2025.

Zhengbo Xu, Jie Ma, Ziheng Wang, Zhan Peng, Jun Liang, and Jing Li. MoCha: End-to-end video character replacement without structural guidance. *arXiv preprint arXiv:2601.08587*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Qi Zhao, Zhan Ma, and Pan Zhou. Dreaminsert: Zero-shot image-to-video object insertion from a single image. *arXiv preprint arXiv:2503.10342*, 2025a.

Shengkui Zhao, Zexu Pan, and Bin Ma. Clearervoice-studio: Bridging advanced speech processing research and practical deployment. *arXiv preprint arXiv:2506.19398*, 2025b.

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.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. Minimax-remover: Taming bad noise helps video object removal. *Advances in Neural Information Processing Systems*, 38:75518–75547, 2026.

A Dataset Profile and Curation Details

A.1 Source Corpus and Pre-Filtering

We collect videos featuring human subjects from films, television programs, and short-form video platforms. We filter the collection along multiple dimensions, including file specifications (duration, resolution, frame rate, and audio sampling rate), visual quality, temporal continuity and motion, visual contamination, suitability of the person and speaking scene, audio quality, audio-visual consistency, and content safety. We then stratify the retained clips and balance their distributions across key attributes.

A.2 Editing Instruction Generation

To enable precise model comprehension of cross-modal editing intents, we leverage Qwen3.5 (Qwen Team, 2026) to jointly analyze source video frames, bounding boxes of edited/erased regions, and reference images, automatically generating structured instructions: 1) **Unique Referring Description**: Details the target person's absolute and relative spatial position along with unique appearance traits, grounding the referring expression to prevent referent ambiguity during inference; 2) **Action Prompts**: concise editing directives (e.g., "*Replace the man in suit with the reference person*" or "*Remove the speaker*"). These two components are combined into a complete instruction via templates and polished using Qwen3.5 to ensure strong generalization across diverse user inputs in real-world inference scenarios.

For *Insert*, since the reference image and audio specify only the target person's identity (appearance and timbre) but not their action or speech content in the video, we additionally annotate this information within the instruction: actions are annotated by Qwen3.5, while speech content is primarily derived from subtitle OCR and pre-obtained scripts, with ASR serving as an auxiliary check. These annotations preserve the temporal synchronization between action and speech, alternately describing the action and the corresponding utterance at each moment rather than listing all speech content at once. All instructions strictly omit explicit identity or timbre descriptions present in the reference inputs (e.g., specific facial features or vocal traits), preventing the model from exploiting textual shortcuts to bypass reference encoders.

A.3 Multi-Stage Quality Verification Suite

To ensure pristine quality and fidelity across the synthesized dataset, we incorporate a rigorous quality verification suite throughout the entire synthesis workflow: 1) **Raw Data Pre-Filtering**: VBench (Huang et al., 2024) evaluates video aesthetics and temporal smoothness, while SyncNet (Chung & Zisserman, 2016) filters out audio-visual synchronization offsets; 2) **Synthesis Quality Control**: Qwen3.5 checks the precision of SAM 2 masks and prompt bounding boxes, while ClearVoice (Zhao et al., 2025b) performs denoising and enhancement on cloned speech; 3) **Cross-Pair Audit**: ArcFace verifies facial cosine similarity, CAM++ audits speaker identity consistency, and ASR combined with subtitle OCR cross-verifies spoken dialogue accuracy, ultimately yielding high-fidelity audio-visual editing pairs.

For generatively augmented cross-pair references, we apply additional modality-specific checks. Qwen3.5 verifies visual fidelity and the absence of visible artifacts, while ASR-based checks assess the linguistic correctness of synthesized speech. We then compute CLIP-based visual similarity (Radford et al., 2021) and CAM++-based timbre similarity (Wang et al., 2023) between each synthesized reference and the original identity, retaining only samples with a CLIP score above 0.75 and a CAM++ score above 0.70.

B Training Details

Table 3 summarizes the data mixtures, optimization settings, trainable modules, and hardware for the curriculum in Sec. 3.3. All stages optimize rank-256 LoRA adapters; the pre-trained backbone weights, VAEs, and T5 text encoder remain frozen.

The filtered, camera-captured VOR subset (Fu et al., 2026) provides real removal supervision in Stages 1 and 2. Quality refinement selects approximately the top 50% of our constructed data pool as a whole while retaining this external subset.

Table 3: Stage-wise training configuration. Dataset sizes and task mixtures are approximate. The VOR subset is included in, not added to, the reported *Delete* counts.

Configuration	**Stage 0** Audio Warm-up	**Stage 1** Bidirectional Joint Adaptation	**Stage 2** Quality Refinement
Training data	200K audio examples	120K editing pairs	70K editing pairs
Task mixture	30% voice conversion; 40% text-to-speech; 30% vocal separation	40K *Swap*; 37K *Insert*; 41K *Delete*	25K *Swap*; 15K *Insert*; 30K *Delete*
Reference source	–	Retrieved + generated	Retrieved only
VOR real-captured subset	–	20K *Delete*	20K *Delete*
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-5}	1×10^{-5}	1×10^{-5}
LoRA rank	256	256	256
Batch size	512	64	64
Training steps	2K	2K + 2K	2K
Trainable LoRA modules	A only	First 2K: V + F; next 2K: A + V + F	A + V + F
GPUs	8× NVIDIA H800	24× NVIDIA H800	24× NVIDIA H800
Distributed training	DeepSpeed ZeRO Stage 1	DeepSpeed ZeRO Stage 2	DeepSpeed ZeRO Stage 2

A and V denote LoRA adapters in the audio and video DiT blocks, respectively; F denotes adapters in the bidirectional audio-video cross-modal fusion attention.

C Additional Ablation Studies

C.1 Effect of Modality-Aware Sync-RoPE

We compare our full model with a conventional sequential RoPE variant that distinguishes token roles only by their positions along the first RoPE axis. For video, the target and source tokens occupy $[0, T)$ and $[T, 2T)$ on this axis, respectively, and the reference image is placed at $2T$; all three use the same spatial coordinates $[0, H) \times [0, W)$. For audio, the target, source, and reference tokens occupy $[0, T_a)$, $[T_a, 2T_a)$, and $[2T_a, 3T_a)$, respectively, while sharing auxiliary coordinates $\{0\} \times \{0\}$. Thus, this variant removes both the role-specific offsets along the H and W axes and the fixed reference margin M used by Modality-Aware Sync-RoPE. The token concatenation order, network architecture, training data, optimization schedule, and all other settings remain unchanged.

As shown in Table 4, replacing Modality-Aware Sync-RoPE with sequential RoPE consistently degrades imaging quality and AudioBox PQ across all three tasks. The full design improves imaging quality by 8.78, 4.68, and 3.27 points and AudioBox PQ by 0.659, 0.575, and 0.952 for *Swap*, *Delete*, and *Insert*, respectively. It also substantially improves audio-visual alignment for *Swap* and *Insert*, increasing Sync-C by 2.25 and 1.44 while reducing Sync-D by 1.31 and 1.18. Visual consistency generally improves as well, despite isolated exceptions in background consistency for *Swap* and motion smoothness for *Insert*. These results demonstrate that jointly preserving source–

Table 4: Ablation study of Modality-Aware Sync-RoPE on AVE-Dataset-Bench across tasks. Best results are shown in bold.

	VBench (%)					Audio Quality & AV Sync		
Task / Variant	Subj. Cons. ↑	BG Cons. ↑	Temp. Flick. ↑	Motion Sm. ↑	Imaging Q. ↑	PQ ↑	Sync-C ↑	Sync-D ↓
Swap								
w/o Modality-Aware Sync-RoPE	96.46	**96.01**	95.38	98.43	60.96	6.050	3.14	9.84
Ours	**97.35**	95.69	**97.67**	**98.54**	**69.74**	**6.709**	**5.39**	**8.53**
Delete								
w/o Modality-Aware Sync-RoPE	94.24	93.73	96.09	97.80	62.47	6.235	–	–
Ours	**98.49**	**96.64**	**98.12**	98.43	**67.15**	**6.810**	–	–
Insert								
w/o Modality-Aware Sync-RoPE	95.87	93.96	97.44	**98.63**	69.55	5.539	2.99	10.01
Ours	**98.36**	**96.87**	**98.05**	98.44	**72.82**	**6.491**	**4.43**	**8.83**

target temporal correspondence, separating token roles on the auxiliary axes, and isolating references with a fixed margin yields more reliable multimodal editing than sequential role encoding alone.