UNIAVC: UNIFIED AUDIO-VIDEO CREATION VIA PROGRESSIVE TRAINING AND ASYMMETRIC SCHEDULING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified audio-video creation requires a single model capable of text-conditioned joint synthesis, bidirectional cross-modal generation, and instruction-guided editing across both modalities. Existing methods rely heavily on paired corpora, which are scarce for joint generation and even rarer for editing. Consequently, they are spent learning single-branch generative and editing capabilities that abundant unimodal data could readily teach, rather than on the cross-modal coupling. Furthermore, prevailing systems apply a single shared noise schedule to both modalities, neglecting two critical limitations. First, due to the significant discrepancy in information density, video requires prolonged high noise to organize spatial and motion structure, whereas the sparser audio signal can settle earlier during the synthesis process, the dynamics a uniform schedule might compromise. Second, lockstep denoising leaves asymmetric workflows, such as conditioning on one modality to synthesize or edit another, untrained. We present UniAVC, which addresses both obstacles. It first trains each branch on unimodal dataset to full generative and editing competence, and subsequently dedicates paired data to exploit cross-modal correspondence. Modality-decoupled schedulers meanwhile warp the noise schedule independently from a shared timestep, while subtasks with one clean modality directly supervise asymmetric configurations. These settings respectively enable intent-aware editing, which preserves one modality intact while enlisting it to constrain the other, and post-inference self-refinement, which regenerates only an unsatisfactory branch against an accepted one. UniAVC thus spans multiple creation tasks from a single set of weights without task-specific adapters. Extensive experiments demonstrate that UniAVC achieves versatile audio-video generation and editing with competitive performance across different operating modes.

1 INTRODUCTION

Audio-video content creation has advanced along several largely independent lines. Joint generators synthesize a clip together with its soundtrack in a single pass, achieving a synchronization that cascaded pipelines struggle to match (Low et al., 2025; Wang et al., 2025a; Pham et al., 2026; HaCohen et al., 2026; Huang et al., 2025). Cross-modal models generate one modality conditioned on the other (Luo et al., 2023; Zhang et al., 2024; Cheng et al., 2025). Instruction-guided editors, extending a paradigm established for images (Brooks et al., 2023; Zhang et al., 2023), modify an existing clip according to a textual request (Chen et al., 2026c; Zheng et al., 2026), and control adapters attach conditioning to a pretrained backbone, typically one adapter per control (Ben-Yosef et al., 2026). Each line has matured on its own terms, and each ships its own weights, training corpus, and task coverage.

Operating across these decoupled paradigms carries a severe cost for audio-video creation. Because existing tools specialize in either generation or editing, practical workflows must cascade independent models sequentially, for instance, by generating a synchronized clip with a joint model and then modifying it with another instruction-guided editor. This sequential pipeline approach inherently degrades cross-modal coherence. When an editor modifies the visual stream without awareness of the soundtrack, it immediately breaks the precise synchrony established during initial generation. This drift accumulates through iterative edits, such as modifying ambient noise or restyling a visual scene, forcing users to deploy dedicated resynchronization modules or explicit alignment losses just to repair the audio-visual match (Chen et al., 2026c; Liang et al., 2026). A single model that natively unifies

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.

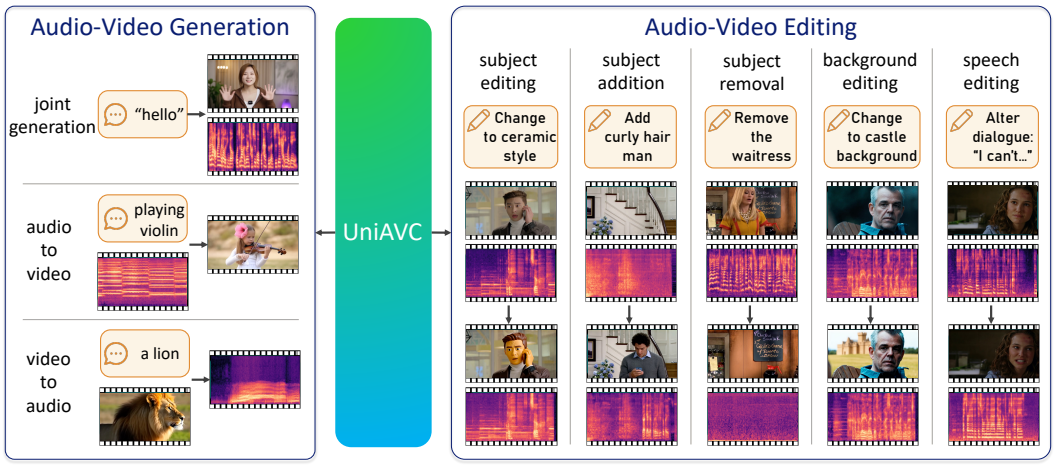

Figure 1: UniAVC performs joint audio-video generation, cross-modal generation in either direction, and instruction-guided editing of either or both modalities from one set of weights.

generation and editing across both modalities eliminates this vulnerability, establishing temporal correspondence once and retaining it across all downstream revisions (Huang et al., 2025). Realizing such a unified framework, however, remains markedly harder than assembling specialized tools due to two compounding challenges.

C1: Scarcity of paired audio-video supervision. High-quality datasets offering both precise audio-visual alignment for generation and instruction-guided editing pairs are far smaller than massive unimodal corpora. Training a unified model directly on this limited paired data yields poor convergence, as the network must simultaneously master multiple competing objectives, including source preservation, text instruction compliance, and cross-modal synchronization (Zheng et al., 2026). Exacerbating this bottleneck, much of this scarce paired supervision is inefficiently spent teaching the model basic single-modality capabilities, such as rendering realistic visual motion or synthesizing clear acoustic signals. Because unimodal datasets already provide these isolated features in abundance, relying on paired data to learn them dilutes supervision signal through redundant learning, leaving inadequate learning capacity to capture the intricate cross-modal coupling that unimodal data alone can never supply.

C2: Neglected information asymmetry between audio and video representations. Video latents maintain a dense spatio-temporal structure across every frame (Blattmann et al., 2023; HaCohen et al., 2024), whereas audio codecs compress signals into a far sparser sequence (Liu et al., 2023a; Evans et al., 2024). Consequently, a clip's visual token stream outnumbers its acoustic counterpart by more than an order of magnitude (HaCohen et al., 2026). Prevailing joint architectures nevertheless corrupt both modalities along a single shared noise schedule, leaving this structural disparity unaddressed. Because video requires a prolonged high-noise phase to establish global texture, identity, and motion before refining fine details, its optimal denoising trajectory differs fundamentally from that of audio, which settles into its final acoustic content much earlier (Fig. 2). Aligning the shared schedule to video leaves audio artificially noisy long after it has converged, depriving the visual stream of clean cross-modal conditioning. Conversely, optimizing the schedule for audio forces video to commit to details prematurely before its global structure has formed. Ultimately, a single rigid schedule fails both modalities.

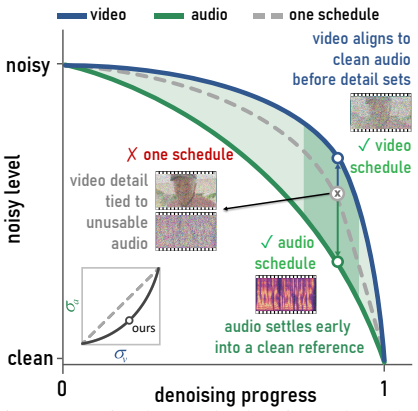

Figure 2: Single vs. dual noise schedules.

Beyond ignoring these distinct denoising trajectories, enforcing a shared schedule severely restricts how the model explores noise states during training. Because each modality carries an independent

noise level, the true state space during training is two-dimensional (Bao et al., 2023), rather than the single scalar timeline assumed in standard diffusion frameworks (Ho et al., 2020; Song et al., 2021). Locking both streams to equal noise levels forces training strictly along the diagonal of this 2D space, teaching the model only how to denoise symmetrically corrupted pairs. The core capabilities promised by a unified framework, however, depend on off-diagonal, asymmetric noise states. Tasks such as restyling a visual scene while preserving its soundtrack or generating video conditioned on existing audio require denoising a corrupted video stream against completely clean audio. Because a shared schedule never exposes the network to these off-diagonal configurations during training, the model encounters a severe distribution shift at inference, asking it to execute conditioning behaviors it was never trained to perform.

Our approach: UniAVC. To address these challenges, we introduce UniAVC, a unified audio-video creation model built with progressive training and decoupled noise scheduling. We first establish full generative and editing competence in each modality using abundant unimodal corpora, then freeze both branches so that scarce paired data is used exclusively to train lightweight cross-modal bridges. This guarantees that paired supervision is dedicated entirely to cross-modal coupling rather than redundant single-modality learning. To address representation asymmetry and explore off-diagonal noise space, asymmetric modality-decoupled schedulers map a shared timestep index onto two independently warped noise trajectories, curving the training region away from the symmetric diagonal. Complementing this, clean-branch subtasks explicitly supervise extreme boundary configurations that continuous warped curves cannot reach. This joint coverage unlocks intent-aware editing, which holds the preserved modality as a clean reference to return the source material unchanged, as well as post-inference self-refinement, enabling the selective regeneration of an unsatisfactory branch conditioned on an accepted one.

Our primary contributions are summarized as follows:

- **Unified generation and editing framework.** A single model natively performs text-conditioned joint audio-video generation, bidirectional cross-modal synthesis, and instruction-guided editing of either or both modalities without requiring task-specific adapters (Fig.1).
- **Progressive training paradigm.** By first mastering unimodal generative and editing capabilities on single-modality corpora before freezing both branches, our framework dedicates scarce paired datasets entirely to establishing cross-modal correspondence and temporal synchronization.
- **Asymmetric noise-level scheduling.** Modality-decoupled schedulers derive per-branch noise trajectories from a shared timestep index to steer training off the diagonal of the 2D noise space, while clean-branch subtasks explicitly supervise extreme boundary states, together enabling intent-aware editing and post-inference self-refinement.
- **Extensive empirical evaluation.** Comprehensive experiments across diverse generation and editing benchmarks demonstrate that UniAVC achieves performance competitive with specialized state-of-the-art baselines across all supported creation modes.

2 Related Work

Joint Audio-Video Generation. Early joint synthesis coupled distinct modality-specific diffusion processes, either through shared cross-modal architectures (Ruan et al., 2023) or by aligning independently trained generators during inference (Xing et al., 2024). In parallel, cross-modal models synthesize one stream conditioned on the other, deriving alignment directly from the conditioning input rather than learning a joint prior (Luo et al., 2023; Zhang et al., 2024; Cheng et al., 2025; Pham et al., 2025). Contemporary joint frameworks scale these paradigms by fusing twin backbones block by block (Low et al., 2025), integrating specialized experts into a shared trunk (Wang et al., 2025a; OpenMOSS Team, 2026), estimating hierarchical spatio-temporal priors before generation (Liu et al., 2026a;b), or training end-to-end audio-visual foundation models (HaCohen et al., 2026). These advances leverage video diffusion transformers (Blattmann et al., 2023; HaCohen et al., 2024; Kong et al., 2024; Polyak et al., 2024) and audio latent diffusion models (Liu et al., 2023a; Evans et al., 2024) that made high-resolution synthesis feasible. While these systems differ in their cross-modal interaction mechanisms, they universally share a core design assumption: corrupting both modalities along a single shared noise schedule, a key limitation our work directly re-examines.

Joint Audio-Video Editing. Instruction-guided audio-video editing requires modifying specific target attributes while preserving unmentioned content intact. To address the resulting data scarcity,

JAVEdit (Chen et al., 2026c) employs an agent-in-the-loop curation pipeline and benchmark, whereas SpongeBob (Liang et al., 2026) reformulates editing as self-supervised inpainting on unannotated clips. InstructAV2AV (Zheng et al., 2026) trains an end-to-end joint editor using gated attention to balance instruction compliance against source preservation. For efficiency, AVControl (Ben-Yosef et al., 2026) deploys lightweight per-modality control adapters, whereas JoVA (Huang et al., 2025) and the concurrent MiniMax-H3 (MiniMax, 2026) unify generation and editing within a shared representation space. These existing methods rely heavily on scarce paired editing data to establish basic single-modality competence. In contrast, UniAVC derives single-modality capabilities from abundant unimodal corpora, dedicating its paired supervision exclusively to cross-modal coupling.

Non-uniform Noise Schedules. Diffusion models typically apply a single noise level across an entire sample, but two distinct lines of work depart from this assumption. The first varies noise levels across *positions* within a single modality, using independent per-frame noise (Chen et al., 2024), sliding temporal windows (Ruhe et al., 2024; Kim et al., 2024b), or monotonic token schedules (Wu et al., 2023a). The second varies noise across *modalities*: UniDiffuser (Bao et al., 2023) samples independent noise for image and text, while MoNL (Kim et al., 2024a) varies noise along temporal and modal axes to enable arbitrary conditional generation. Extending this to audio and video, MoVA (OpenMOSS Team, 2026) draws timesteps independently per stream. While this covers more of the joint noise space, uncoupled sampling breaks the deterministic link between noise pairs and the step counter, leaving the model without structured conditioning during inference. Orthogonally, rectified-flow models scale global schedules by sequence length (Esser et al., 2024), acknowledging density disparities while still enforcing a single trajectory per sample. In contrast, UniAVC assigns dedicated schedules per modality anchored to a shared step counter. Trajectories remain phase-locked as their magnitudes diverge, keeping the denser visual stream at high noise long after the acoustic stream converges. Finally, two clean-branch subtasks explicitly supervise boundary states where one stream remains clean, directly enabling single-modality editing and post-inference branch refinement.

3 METHOD

3.1 PRELIMINARY

Problem Formulation. We operate in the latent spaces of pretrained video and audio autoencoders. A sample is defined as a tuple $(c, \boldsymbol{v}_0, \boldsymbol{a}_0, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{a}})$ pairing a text condition c with clean target latents $(\boldsymbol{v}_0, \boldsymbol{a}_0)$ alongside source reference latents $(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{a}})$, which are supplied during editing and set to $\varnothing$ for generation from scratch. A specific task is determined by which references are present and which branches are supervised via binary masks $\boldsymbol{M}_v, \boldsymbol{M}_a \in \{\mathbf{0}, \mathbf{1}\}$. Joint generation and joint editing both set $\boldsymbol{M}_v = \boldsymbol{M}_a = \mathbf{1}$, differing only in whether non-null references are provided. Conversely, cross-modal and single-modality tasks supervise one branch while setting the opposing mask to zero to hold that stream clean. Consequently, generation and editing are unified under a single formulation, distinguished solely by their conditioning inputs and supervision masks.

Rectified Flow. Rectified flow learns a velocity field by regressing the derivative of a prescribed linear path (Liu et al., 2023b; Lipman et al., 2023). For a clean latent $\boldsymbol{x}_0$ and noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the path is $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\,\boldsymbol{\epsilon}$ with $t \sim \mathcal{U}(0,1)$, whose target velocity is $\boldsymbol{u}(\boldsymbol{x}_t, t \mid \boldsymbol{x}_0, \boldsymbol{\epsilon}) = \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$. Under this path, the absolute noise level σ of $\boldsymbol{x}_t$ is numerically interchangeable with t, *i.e.*, $\sigma \triangleq t$, hence we use σ throughout for clarity. Assigning each modality its own level gives

$$\boldsymbol{v}_{\sigma_v} = (1-\sigma_v)\,\boldsymbol{v}_0 + \sigma_v\,\boldsymbol{\epsilon}_v, \qquad \boldsymbol{a}_{\sigma_a} = (1-\sigma_a)\,\boldsymbol{a}_0 + \sigma_a\,\boldsymbol{\epsilon}_a, \qquad \boldsymbol{\epsilon}_v, \boldsymbol{\epsilon}_a \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

with $\boldsymbol{\epsilon}_v, \boldsymbol{\epsilon}_a$ drawn independently. A velocity network v_θ is optimized with

$$\mathcal{L}_{\mathrm{FM}} = \sum_{i \in \{a,v\}, \boldsymbol{x} \in \{\boldsymbol{a}, \boldsymbol{v}\}} \mathbb{E}_{\boldsymbol{x}_0, c, \boldsymbol{\epsilon}_i, \sigma_i, \tilde{\boldsymbol{x}}} \left[\boldsymbol{M}_i\, w_i \left\| v_\theta^{(i)} - (\boldsymbol{\epsilon}_i - \boldsymbol{x}_0) \right\|_2^2 \right], \tag{2}$$

where $v_\theta^{(v)}$ and $v_\theta^{(a)}$ are the branch velocities predicted from both noised latents, both noise levels, the text and the references, and w_v, w_a are the branch loss weights. The expectation is taken over the data distribution of the tuple, the independent noise draws, and the noise-level pair.

3.2 UNIFIED GENERATION AND EDITING

Backbone. We build upon LTX-2.3 (HaCohen et al., 2026), where video and audio tokens are processed by separate transformer stacks, each equipped with dedicated self-attention, feed-forward

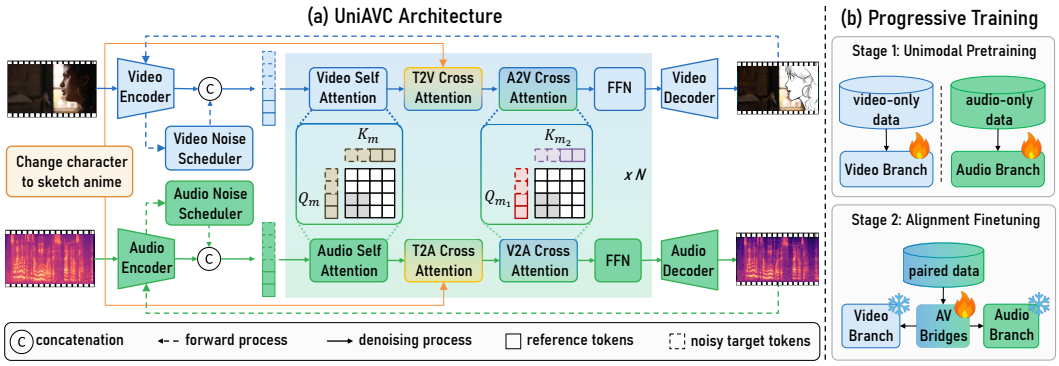

Figure 3: Illustration of the UniAVC framework. (a) Two transformer stacks process video and audio, intersecting exclusively at per-block cross-attention bridges whose modulation is driven by the partner branch's noise level. Reference latents enter as in-context tokens concatenated to target tokens, with attention masks applied so that reference tokens do not attend to noisy tokens in either self- or cross-attention blocks. (b) Each branch first attains full generative and editing competence on unimodal data in parallel; both branches are subsequently frozen, and the paired corpus is used solely to train the cross-modal bridges.

blocks, and adaptive normalization (Peebles & Xie, 2023), while both attend to a shared text embedding. The two branches intersect exclusively at per-block bridges, consisting of an audio-to-video cross-attention layer where video queries attend to audio keys and values, and a symmetric video-to-audio layer with swapped roles. Each bridge operates as a gated residual module whose scale, shift, and gate are derived via adaptive normalization from a noise level. The critical design choice lies in which noise level modulates each path. Crucially, each bridge is modulated by the noise level of the opposing branch, such that the gate regulating information flow from audio into video is a function of σ_a, while its mirror in the visual branch depends on σ_v (Fig. 3a).

In-Context Reference Conditioning. Rather than adding dedicated input branches, we adopt in-context conditioning for diffusion transformers (Huang et al., 2024) by concatenating reference latents, patchified by the same encoder, onto the target sequence of their corresponding modality. A temporal offset applied to the reference rotary embeddings and a learned token-type embedding separate the two sequence spans. Prior in-context video editing works assign reference tokens the exact positions of their target counterparts, enforcing strict spatiotemporal alignment between input and output (Chen et al., 2026a; Polaczek et al., 2025). However, this strict alignment fails when editing instructions require relocating or transforming content. Furthermore, naive concatenation allows reference tokens to attend to noisy target tokens, causing the conditioning signal to drift and degrade as noise increases. We prevent this by applying attention masks that allow target queries to attend to the full sequence while restricting reference queries to read only the uncorrupted reference span. The same isolation extends to cross-modal bridges, where reference queries in one branch can attend to the partner modality's reference span but not its noisy target (Fig. 3a).

3.3 Progressive Training Paradigm

Two Stages. Paired audio-video editing datasets are scarce compared to the abundant unimodal corpora used to pretrain single-modality generators. Training a joint editor directly on limited paired data forces the model to learn source preservation, instruction following, and cross-modal synchronization simultaneously, a multi-task burden known to cause poor convergence (Zheng et al., 2026). To resolve this bottleneck, we decompose training into two stages: Unimodal Pretraining and Alignment Finetuning (Fig. 3b). The first stage establishes full generative and editing competence within each independent branch using modality-specific data. The second stage integrates both pretrained branches into the dual-branch backbone and updates only the cross-modal bridges alongside their adaptive normalization parameters. This strategy dedicates scarce paired datasets entirely to learning cross-modal alignment and temporal synchronization, which cannot be acquired from unimodal corpora alone.

Conditioning Dropout. Classifier-free guidance (Ho & Salimans, 2022) is enabled by randomly dropping the text prompt and reference latents independently during training. This exposes two distinct guidance axes at inference rather than a single scale: the text axis governs instruction compliance, while the reference axis enforces source fidelity, allowing the two signals to be traded off explicitly during sampling. Furthermore, we include a reconstruction subtask where the reference is designated as the target latent under a null-edit instruction. This subtask provides explicit supervision for source preservation, directly counteracting the tendency of instruction-tuned models to alter regions unmentioned by the prompt.

3.4 Asymmetric Modality-Decoupled Schedulers

Noise-Level Pair. Standard diffusion models typically sample a single noise level per trajectory across all modalities (Bao et al., 2023; Kim et al., 2024a), imposing the constraint $\sigma_v = \sigma_a$ on Eq. 1. In contrast, with independently drawn noise vectors $\boldsymbol{\epsilon}_v$ and $\boldsymbol{\epsilon}_a$, Eq. 1 defines a family of corrupted distributions parameterized by the noise-level pair (σ_a, σ_v):

$$q_{\sigma_a,\sigma_v}(\boldsymbol{v}, \boldsymbol{a} \mid \boldsymbol{v}_0, \boldsymbol{a}_0) = \mathcal{N}\big(\boldsymbol{v}; (1-\sigma_v)\boldsymbol{v}_0, \sigma_v^2 \boldsymbol{I}\big)\mathcal{N}\big(\boldsymbol{a}; (1-\sigma_a)\boldsymbol{a}_0, \sigma_a^2 \boldsymbol{I}\big), \quad \sigma_a, \sigma_v \in [0,1], \tag{3}$$

from which each branch regresses its target velocity. Specifically, the video branch targets $\mathbb{E}_{\boldsymbol{\epsilon}_v \mid \boldsymbol{v}_{\sigma_v}, \boldsymbol{a}_{\sigma_a}, c}[\boldsymbol{\epsilon}_v - \boldsymbol{v}_0]$, while the audio branch targets the analogous vector for its own stream. Because each branch conditions on the partner stream at its respective noise level, three distinct operational regimes emerge. On the diagonal ($\sigma_a = \sigma_v$), the model observes a partner stream that is informative yet noisy, regressing a velocity that averages over the partner's uncertainty. When $\sigma_a = 0$, the partner stream is completely clean, allowing the video branch to treat audio as an uncorrupted constraint, which is essential for cross-modal generation and audio-preserving editing. When $\sigma_a = 1$, Eq. 1 yields pure noise ($\boldsymbol{a}_{\sigma_a} = \boldsymbol{\epsilon}_a$), making the partner independent of $(\boldsymbol{v}_0, \boldsymbol{a}_0, c)$ so that the branch relies solely on text conditioning. Standard shared schedules visit only the diagonal (Fig. 2), preventing the model from encountering the clean-partner boundary regimes needed for downstream editing and cross-modal generation tasks.

Dual-Schedule Shifting. A shared synchronization clock does not imply a shared noise level. We maintain a single step index $i \in \{0, \dots, T-1\}$ drawn uniformly per sample while decoupling its mapping to each branch's noise level by warping the base linear grid $\bar{\sigma}_i = 1 - i/T$ per modality:

$$S_k(\sigma) = \frac{k\,\sigma}{1 + (k-1)\,\sigma}, \quad \sigma_v^{(i)} = \mathcal{R}\big[S_{k_v}\big](\bar{\sigma}_i), \quad \sigma_a^{(i)} = \mathcal{R}\big[S_{k_a}\big](\bar{\sigma}_i), \tag{4}$$

where k_v and k_a are per-branch shift strengths and $\mathcal{R}$ rescales the warped grid to a shared terminal noise level. Because S_k is an increasing bijection on $[0,1]$ and $\mathcal{R}$ preserves terminal boundaries, both schedules span identical endpoints. Substituting Eq. 4 into Eq. 3 yields a dual schedule $i \mapsto (\sigma_a^{(i)}, \sigma_v^{(i)})$ that departs from the diagonal throughout its interior. Because S_k is strictly increasing with respect to k, setting $k_v > k_a$ ensures $\sigma_a^{(i)} < \sigma_v^{(i)}$ at every interior step, positioning the schedule strictly below the diagonal (Fig. 2). Audio thus resolves ahead of video, allowing cross-modal bridges to pass a progressively cleaner acoustic representation into the video branch, establishing a natural temporal ordering from two warped schedules on a single index rather than manual staging.

Clean-Branch Subtasks. Because continuous schedules rarely reach the boundary faces where $\sigma = 0$ for an individual branch, we explicitly supervise these boundary conditions to support cross-modal generation and single-modality editing. Specifically, we assign a small fraction of training samples to subtasks where one stream is pinned to its uncorrupted latent and its loss is masked, such as audio-conditioned video generation ($\sigma_a = 0$, $\boldsymbol{M}_a = 0$ in Eq. 2) or its visual counterpart. Masking the loss reflects that the uncorrupted stream serves as a fixed conditioning signal rather than a regression target, which guarantees that single-modality edits leave the non-target stream untouched (Sec. 3.5). Concurrent work (Liang et al., 2026) adopts similar modes by switching between fixed discrete configurations; in contrast, our unified step index dynamically positions each denoising step along distinct off-diagonal pairs within a continuous noise-level space.

3.5 Selective Regeneration

Dual-schedule training expands the domain of observed noise-level pairs, directly enabling two specialized inference modes. By holding one branch at $\sigma = 0$ while denoising the other, these modes remain in-distribution, as clean-branch subtasks explicitly supervise these boundary regimes.

Intent-Aware Editing. Users frequently require targeted edits that modify one modality while preserving the other, such as restyling video while retaining original dialogue. Single-modality editors treat the preserved stream as static context uncoupled from the output, causing the edited result to drift out of alignment (Huang et al., 2025; Ben-Yosef et al., 2026; Liang et al., 2026). In contrast, we pin the preserved branch at $\sigma_a = 0$ as a reference latent. The visual branch thus conditions directly on clean latents $\boldsymbol{a}_0$ rather than a corrupted trajectory $\boldsymbol{a}_{\sigma_a}$, referencing acoustic timing, onsets, and lip motion at every step. Because no denoising steps are applied to the pinned branch, the sampler returns the exact source latent rather than an approximated resynthesis, guaranteeing exact preservation by construction. Both editing directions operate without task-specific adapters, whereas evaluating this setup on a single-schedule baseline constitutes an out-of-distribution extrapolation.

Self-Refinement. When a joint generation yields only one satisfactory modality, users can retain the accepted stream and regenerate the other rather than resampling the entire pair. For instance, to refine a generated audio latent $\hat{\boldsymbol{a}}_0$ while keeping the video branch fixed, a second denoising trajectory is executed after the initial pass completes. Unlike two-stage inference guidance (Liang et al., 2026), which adjusts conditioning within a single trajectory, self-refinement re-noises $\hat{\boldsymbol{a}}_0$ to a noise level $\sigma_a = \lambda$ and denoises along a rescaled schedule $\lambda\sigma_a^{(i)}$ while holding the visual branch clean at $\sigma_v = 0$. This mechanism mirrors the pinning and in-context conditioning of intent-aware editing, differing only in the source of the pinned stream, which uses the first-pass generated latent rather than user-supplied input.

4 Experiments

4.1 Experimental Setup

Implementation Details. We initialize each branch from the pretrained weights of LTX-2.3 (HaCohen et al., 2026) and train it in the two stages of Sec. 3.3. Each stage runs at 720p over 121 frames at 24 fps with 16 kHz mono audio on H800 GPUs, with AdamW optimizer at a learning rate of 1×10^{-5} and weight decay 0.01 in `bf16`. Table 1 gives the sample budget of each training stage, where alignment finetuning spends under 1% of the unimodal pretraining in total. Schedulers run with $k_v = 9$ and $k_a = 2$. The text and reference dropout probabilities are both set to 0.1. Inference uses 30 denoising steps, with text and reference CFG scales of $(3.0, 2.0)$ and $(1.5, 3.0)$ for video and audio respectively, and $\lambda = 0.5$ when self-refinement is enabled.

Table 1: Sample budget by stage.

Stage	Data	Corpus	Seen
Pretrain (V)	gen	14.5M	33.4M
	edit	0.8M	0.8M
	total	**15.3M**	**34.3M**
Pretrain (A)	gen	14.0M	27.4M
	edit	0.9M	27.4M
	total	**14.9M**	**54.9M**
Align (AV)	gen	0.46M	0.31M
	edit	0.28M	0.36M
	total	**0.74M**	**0.67M**

Datasets. Stage 1 trains each branch on unimodal data, and consumes no paired sample. The video branch draws 34.3M samples from a 15.3M pool built from Boogu-Image (Chen et al., 2026b), Ditto (Bai et al., 2025) and OpenVE-3M (He et al., 2025); the audio branch draws 54.9M samples from a 14.9M pool built from AudioSet (Gemmeke et al., 2017), VGGSound (Chen et al., 2020), WavCaps (Mei et al., 2024), Emilia (He et al., 2024), LibriTTS (Zen et al., 2019), AudioEdit (Tian et al., 2026) and the ISCSLP 2026 CoT-TTS Challenge (Xue et al., 2026). Stage 2 then finetunes for alignment on 0.67M paired samples drawn from a 0.74M pool built from TAVGBench (Mao et al., 2024), JAVEdit (Chen et al., 2026c) and InsAVE-80K (Zheng et al., 2026). Table 1 summarizes both stages.

Benchmarks. We evaluate editing and generation capabilities of UniAVC on separate benchmarks, in which we follow their evaluation protocol for fair comparison against selected baselines unless otherwise specified. For joint editing, we use JAVEditBench Chen et al. (2026c) and the held-out evaluation split of InsAVE-80K (Zheng et al., 2026). For joint generation, we employ Verse-Bench (Wang et al., 2025a), while for cross-modal generation, we adopt the test set of VGGSound (Chen et al., 2020) with protocol from AVTok (Pham et al., 2026) for assessment.

Baselines. We select baselines per capability. For joint editing, we compare against AVED (Lin et al., 2025), AVI-Edit (Zheng et al., 2025), JAVEdit (Chen et al., 2026c), InstructAV2AV (Zheng et al., 2026), MiniMax-H3 (MiniMax, 2026), and cascaded methods (Ishii et al., 2025; Chen et al., 2026c). For joint generation, we compare against Ovi (Low et al., 2025), JavisDiT (Chen et al., 2026c), UniVerse-1 (Wang et al., 2025a), JoVA (Huang et al., 2025), LTX-2.3 (HaCohen et al.,

Table 2: Quantitative comparisons for joint audio-video editing on JAVEditBench and the test split of InsAVE-80K. Column names abbreviate each benchmark's official metrics for brevity. † denotes results retrieved from (Chen et al., 2026c). Best results are in **bold** and second best are underlined.

Method	JAVEditBench						InsAVE-80K					
	VQ↑	AQ↑	AVS↑	IC↑	VF↑	AVQ↑	FVD↓	TV-A↑	FAD↓	TA-A↑	AV-A↑	PEAVS↑
AVED	0.0590	1.72	0.1641	2.95	3.87	2.93	305.41	24.68	2.86	29.89	26.44	3.15
AVI-Edit	0.0604	2.34	0.2721	3.49	3.89	3.86	205.34	25.15	5.09	30.07	26.37	2.54
Cascaded	0.0563	2.35	0.2925	3.99	4.08	3.51	326.63	25.18	8.01	32.47	22.67	3.20
JAVEdit†	0.0596	2.42	**0.3688**	4.07	4.22	**3.88**	–	–	–	–	–	–
InstructAV2AV	0.0588	1.40	0.1759	2.88	3.93	3.61	180.38	25.23	2.75	**34.96**	27.72	3.24
MiniMax-H3	**0.0605**	2.42	0.2307	4.11	4.17	3.61	211.24	20.37	3.25	27.32	25.11	**3.75**
UniAVC (ours)	0.0599	**2.59**	0.2975	**4.42**	**4.48**	3.85	**150.86**	**25.66**	**1.99**	27.36	**28.09**	3.51

Table 3: Quantitative comparisons for joint generation on Verse-Bench (Left), and cross-modal generation on VGGSound test split (Right). † denotes results retrieved from (Huang et al., 2025). Best results are in **bold** and second best are underlined.

Method	MS↑	AS↑	PQ↑	CS↑	LSE-C↑	WER↓
Joint Audio-Video						
UniVerse-1	0.20	0.47	5.20	0.16	1.62	0.18
JavisDiT	0.27	0.42	5.49	0.19	0.85	1.00
Ovi	0.55	0.47	6.20	0.20	6.61	0.12
JoVA†	**0.75**	**0.49**	6.23	0.29	6.65	0.11
LTX-2.3	0.37	0.34	6.21	0.26	4.07	0.09
MiniMax-H3	0.34	0.48	**6.87**	**0.36**	**6.96**	0.13
UniAVC (ours)	0.40	**0.49**	6.26	0.30	6.72	**0.08**

Method	FVD↓	FAD↓	DeSync↓	IB↑
Audio-to-Video				
TempoTokens	786.61	–	1.359	0.132
AVTok	150.26	–	1.317	0.143
UniAVC (ours)	**103.17**	–	**0.773**	**0.291**
Video-to-Audio				
MMAudio	–	17.09	0.813	0.291
VinTAGe	–	80.06	1.294	0.044
UniAVC (ours)	–	**10.12**	**0.650**	**0.305**

2026), and MiniMax-H3 (MiniMax, 2026). For cross-modal generation, we compare against some representative specialists in each direction: MMAudio (Cheng et al., 2025), VinTAGe (Kushwaha & Tian, 2025) for video-to-audio; and AVTok (Pham et al., 2026) and TempoTokens (Yariv et al., 2024) for audio-to-video.

4.2 Main Results

Across both evaluations, UniAVC performs competitively against selected baselines across all supported tasks. Specifically, UniAVC outperforms prevailing methods on most joint audio-video editing metrics (Tab. 2) and simultaneously achieves robust results in both joint and cross-modal generation settings (Tab. 3). Qualitative comparisons in Fig. 4 further demonstrate that UniAVC maintains precise instruction following, high perceptual fidelity, and robust preservation of character identity and unedited regions, across various tasks, addressing key limitations often observed in the representative InstructAV2AV and MiniMax-H3.

4.3 Ablation Study

To evaluate the impact of the progressive training recipe, the two inference modes and the asymmetric dual schedulers proposed in Sec. 3, we conduct an ablation study on JAVEditBench, of which the results are shown in Tab. 4.

Training Paradigm. Progressive training outperforms single-stage finetuning across all metrics, primarily on cross-modal ones (AVQ, AQ) rather than visual quality alone, confirming that reserving paired data for the second stage effectively dedicates capacity to cross-modal alignment.

Inference Mode. Intent-aware editing improves audio quality by conditioning on uncorrupted source latents, whereas self-refinement yields the largest gains in instruction compliance by re-noising a single branch against an accepted counterpart. The two modes operate complementarily, with their combination achieving peak cross-modal synchronization (AVS, AVQ) and overall performance.

Scheduler Shift. Removing schedule asymmetry causes notable drops in cross-modal metrics (AQ, AVQ) with minimal impact on visual quality, demonstrating the limitation of symmetric schedules that restrict conditioning to equally corrupted streams. Excessively increasing shift strength to

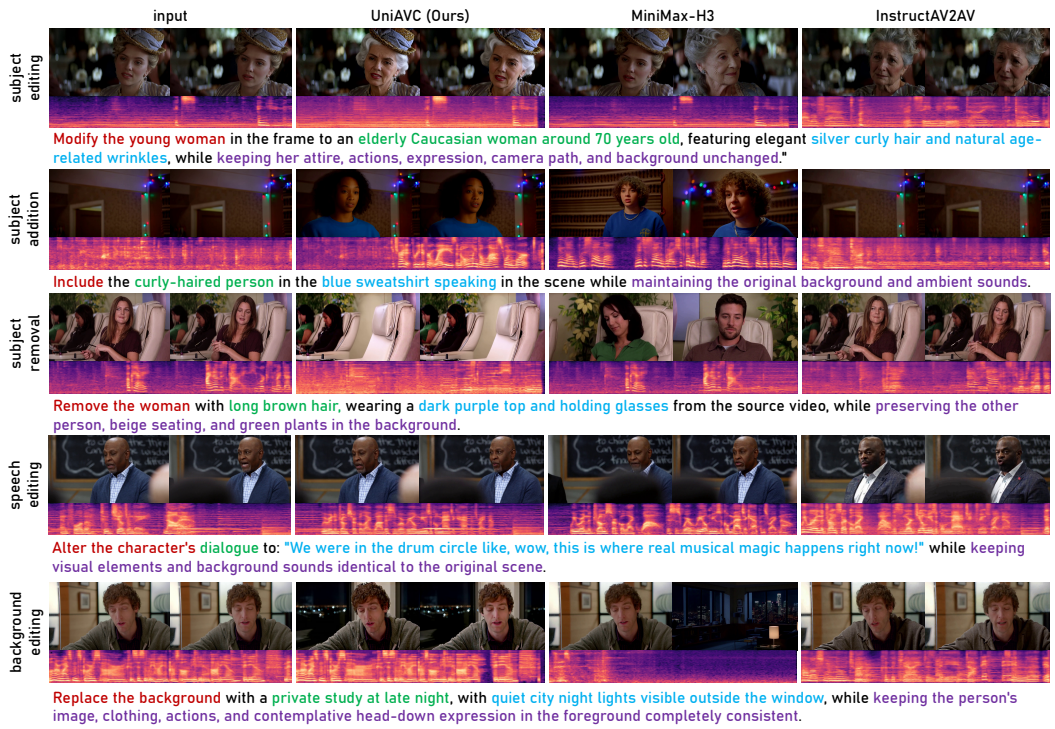

Figure 4: Qualitative comparisons for joint editing on JAVEditBench.

Table 4: Ablation study on JAVEditBench.

	Setting	VQ↑	AQ↑	AVS↑	IC↑	VF↑	AVQ↑
Training paradigm	Single-stage	0.0584	2.27	0.2446	4.21	4.24	3.53
	Progressive	0.0599	2.59	0.2975	4.42	4.48	3.85
Inference mode	Base	0.0584	2.37	0.2643	4.22	4.31	3.68
	Intent-aware	0.0596	2.53	0.2728	4.31	4.39	3.75
	Self-refinement	0.0592	2.52	0.2643	4.39	4.43	3.75
	Full	0.0599	2.59	0.2975	4.42	4.48	3.85
Scheduler shift	$(k_a, k_v) = (1, 1)$	0.0589	2.48	0.2545	4.34	4.35	3.67
	$(k_a, k_v) = (3, 12)$	0.0592	2.51	0.2618	4.41	4.40	3.80
	$(k_a, k_v) = (2, 9)$	0.0599	2.59	0.2975	4.42	4.48	3.85

$(k_a, k_v) = (3, 12)$ degrades performance across most metrics, confirming that schedule warping saturates beyond a moderate shift.

5 Conclusion

We present UniAVC, a unified framework for audio-video generation and editing, capable of text-conditioned joint generation, bidirectional cross-modal synthesis, and instruction-guided editing of either or both streams without requiring task-specific adapters or secondary checkpoints. This capability rests on two key designs: progressive training establishes unimodal competence from single-modality corpora to reserve paired data for cross-modal bridges, while asymmetric schedulers drive independently warped noise trajectories to access the off-diagonal configurations required for editing and cross-modal generation, complemented by clean-branch subtasks that supervise boundary states. Together, these mechanisms enable intent-aware editing to return preserved modalities unchanged and self-refinement to selectively regenerate flawed modalities. These operational modes differ solely in conditioning and supervision rather than architecture, though coverage remains confined to a 1D trajectory and two boundary faces rather than the full noise-pair space, and cross-modal bridges align representations without repairing unimodal deficits.

AI Use Statement. Generative AI tools were employed exclusively to polish the manuscript and enhance text clarity and readability. This assistance was strictly limited to linguistic refinement; generative AI played no role in research ideation, data collection or generation, mathematical proofs, experimental design or implementation, data analysis, or the formulation of scientific claims. All AI-assisted revisions were thoroughly reviewed and verified by the authors to guarantee accuracy. The authors retain full responsibility for the entire final manuscript, including all methods, data, analyses, and conclusions.

REFERENCES

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. *arXiv preprint arXiv:2409.09305*, 2024.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Ditto: Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, 2023.

Matan Ben-Yosef, Tavi Halperin, Naomi Ken Korem, Mohammad Salama, Harel Cain, Asaf Joseph, Anthony Chen, Urska Jelercic, and Ofir Bibi. Avcontrol: Efficient framework for training audio-visual controls. *arXiv preprint arXiv:2603.24793*, 2026.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Anthony Chen, Naomi Ken Korem, 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.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom S. Y. Li, Jintao Lin, Bowen Dong, Zian Qian, Yuhao Liu, Yuxuan Hu, Weikang Shi, Bin Zou, Bowen Zheng, Haoxuan Che, Chang Chen, Yuyang He, Heyang Sun, Tianyu Huang, Chong Hou Choi, Cheng Gong, Han Shi, Haoli Bai, Xihui Liu, Hongsheng Li, Qifeng Chen, Chao Huang, Rui Liu, and Chenyang Lei. Boogu-image-0.1: Boosting open agentic multimodal generation via understanding under a minimal budget. *arXiv preprint arXiv:2607.13125*, 2026b.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

Yinan Chen, Chuming Lin, Zhennan Chen, Yuxiang Zeng, Junwei Zhu, Yali Bi, Xijie Huang, Chengming Xu, Donghao Luo, Zhucun Xue, Xiaobin Hu, Chengjie Wang, Yong Liu, Jiangning Zhang, and Shuicheng Yan. Javedit: Joint audio-visual instruction-guided video editing with agentic data curation. *arXiv preprint arXiv:2606.03168*, 2026c.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Asian Conference on Computer Vision*, 2016.

Mandar Pramod Diwakar and Brijendra Gupta. VGGish deep learning model: Audio feature extraction and analysis. In *International Conference on Data Management, Analytics and Innovation*, pp. 59–70, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J. Han. PEAVS: Perceptual evaluation of audio-visual synchrony grounded in viewers' opinion scores. *arXiv preprint arXiv:2404.07336*, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890, 2024.

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. *arXiv preprint arXiv:2512.07826*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

Xiaohu Huang, Haoyang He, Hao Zhou, Qiangpeng Yang, Shilei Wen, and Kai Han. Jova: Unified multimodal learning for joint video-audio generation and editing. *arXiv preprint arXiv:2512.13677*, 2025.

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

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*, 2021.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, and Krishna Somandepalli. A versatile diffusion transformer with mixture of noise levels for audiovisual generation. In *Advances in Neural Information Processing Systems*, 2024a.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems*, 2024b.

Weijie Kong et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13529–13539, 2025.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. LatentSync: Taming audio-conditioned latent diffusion models for lip sync with SyncNet supervision. *arXiv preprint arXiv:2412.09262*, 2025.

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

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. *arXiv preprint arXiv:2503.20782*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, 2023a.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, 2026a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *International Conference on Learning Representations*, 2026b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b.

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

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Tavgbench: Benchmarking text to audible-video generation. In *ACM International Conference on Multimedia*, 2024.

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

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities, 2026. https://www.minimax.io/blog/minimax-h3.

OpenMOSS Team. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Kien T Pham, Yingqing He, Yazhou Xing, Qifeng Chen, and Long Chen. Spa2v: Harnessing spatial auditory cues for audio-driven spatially-aware video generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Kien T Pham, I Chieh Chen, Qifeng Chen, and Long Chen. Avtok: 1d unified tokenization for holistic audio-video generation. In *European Conference on Computer Vision*, 2026.

Sagi Polaczek, Or Patashnik, Ali Mahdavi-Amiri, and Daniel Cohen-Or. In-context sync-lora for portrait video editing. *arXiv preprint arXiv:2512.03013*, 2025.

Adam Polyak et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *International Conference on Machine Learning*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020.

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

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges, 2019.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025a.

Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. VideoCLIP-XL: Advancing long description understanding for video clip models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 16061–16075, 2024.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8428–8437, 2025b.

Jianzong Wu, Hao Lian, Jiongfan Yang, Dachao Hao, Ye Tian, Yunhai Tong, Jingyuan Zhu, Biaolong Chen, Qiaosong Qi, Aixi Zhang, Wanggui He, Mushui Liu, Pipei Huang, and Hao Jiang. Loomvideo: Unifying multimodal inputs into video generation and editing. *arXiv preprint arXiv:2606.06042*, 2026.

Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. Ar-diffusion: Auto-regressive diffusion model for text generation. In *Advances in Neural Information Processing Systems*, 2023a.

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

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jin Xu, Zhifang Guo, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Wei Xue, Junlan Feng, Shilei Zhang, Yue Wang, Ruosong Yang, Bei Liu, Liumeng Xue, Sitong Cheng, Jiahao Pan, Weizhen Bian, et al. Iscslp 2026 cot-tts challenge: Chain-of-thought reasoning for context-aware text-to-speech. *arXiv preprint arXiv:2606.21933*, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*, pp. 1526–1530, 2019.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024.

Haojie Zheng, Shuchen Weng, Jingqi Liu, Siqi Yang, Boxin Shi, and Xinlong Wang. Audio-sync video instance editing with granularity-aware mask refiner. *arXiv preprint arXiv:2512.10571*, 2025.

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

APPENDIX

A1 Additional Implementation Details **16**

A1.1 Training pools and sampling 16

A1.2 Subtask mixtures 16

A1.3 Evaluation metrics 17

A2 Additional Experimental Results **17**

A2.1 Qualitative results 17

A2.2 Quantitative results 17

A2.3 Ablations 18

A2.4 Limitations 18

Table A1: Detailed decomposition of training pools by stage. "Seen" is the number of samples the optimizer consumes during training.

Stage	Task	Source	Corpus	Prob.	Seen
Unimodal Pretraining (V)	T2I	Boogu-Image	9.5M	0.33	16.3M
	TI2I	Boogu-Image	3.4M	0.33	16.3M
	T2V	Boogu-Image	1.6M	0.17	0.8M
	TV2V	Ditto, OpenVE-3M	0.8M	0.17	0.8M
	stage total		**15.3M**		**34.3M**
Unimodal Pretraining (A)	T2A	sound: AudioSet, VGGSound, WavCaps	2.2M	0.50	27.4M
		speech: Emilia, LibriTTS	11.8M		
	TA2A	AudioEdit	0.9M	0.50	27.4M
		CoT-TTS Challenge	1.0M		
	stage total		**14.9M**		**54.9M**
Alignment Finetuning (AV)	T2AV	TAVGBench	0.46M	0.30	0.31M
	TAV2AV	JAVEdit, InsAVE-80K	0.28M	0.70	0.36M
	stage total		**0.74M**		**0.67M**

A1 Additional Implementation Details

A1.1 Training pools and sampling

Training is divided into two stages based on the target supervision. Unimodal Pretraining optimizes the video and audio branches independently on modality-specific corpora without paired audio-video samples. Alignment Finetuning consumes paired data to establish cross-modal alignment, which unimodal pretraining cannot provide.

Unimodal Pretraining (V). We draw 34.3M samples from a 15.3M pool constructed from Boogu-Image (Chen et al., 2026b), Ditto (Bai et al., 2025), and OpenVE-3M (He et al., 2025). Following Wu et al. (2026), still-image tasks are retained in the mixture to improve frame fidelity and edit diversity. Because image samples consume fewer tokens than 10s video clips, we assign sampling probabilities of 0.33 to each image task and 0.17 to each video task, maintaining an effective balance between spatial fidelity and temporal training dynamics respectively.

Unimodal Pretraining (A). We draw 54.9M samples from a 14.9M pool built from AudioSet (Gemmeke et al., 2017), VGGSound (Chen et al., 2020), WavCaps (Mei et al., 2024), Emilia (He et al., 2024), LibriTTS (Zen et al., 2019), AudioEdit (Tian et al., 2026), and the ISCSLP 2026 CoT-TTS Challenge (Xue et al., 2026), sampling the two audio tasks with equal probability of 0.50. Because speech samples outnumber environmental sound samples by roughly 5:1 within a single task identifier, we reweight the sampler toward domain-level parity rather than task-level parity.

Alignment Finetuning. We draw 0.67M paired samples from a 0.74M pool comprising TAVGBench (Mao et al., 2024), JAVEdit (Chen et al., 2026c), and the held-out split of InsAVE-80K (Zheng et al., 2026), assigning sampling probabilities of 0.30 to joint generation and 0.70 to joint editing. This stage operates in a sub-epoch regime, with 0.67M draws completing roughly 0.90 of a single pass over the 0.74M pool.

Tab. A1 details the per-task breakdown of these dataset sizes and sampling counts.

A1.2 Subtask mixtures

In both training stages, each batch samples a task according to the probabilities in Tab. A1. Alignment Finetuning further mixes multiple objectives within each task, allowing a single paired sample to supervise off-diagonal noise configurations in the (σ_a, σ_v) space. Specifically, the joint generation task assigns weights of 0.7 to the standard objective, 0.1 to reconstruction, and 0.1 to each clean-branch subtask. The joint editing task assigns weights of 0.8 to the standard objective and 0.1 to each self-refinement subtask.

The reconstruction objective uses the reference as the target alongside an instruction requesting no modification. This supplies the sole supervisory signal for leaving content unaltered, counteracting instruction-following losses that reward target changes without penalizing collateral edits. Clean-branch subtasks explicitly supervise the boundary regime where $\sigma = 0$, which is essential for cross-modal generation and single-modality editing, and differ based on whether reference conditioning is present.

A1.3 Evaluation metrics

Each benchmark scores a different metric set, and the same abbreviation denotes different quantities across benchmarks. The metrics are listed below under the benchmark they come from.

JAVEditBench. Six metrics span five dimensions, combining learned perceptual models with a multimodal judge. *Visual quality* (VQ) is scored by VTSS, the aesthetic model adopted by Koala-36M (Wang et al., 2025b). *Audio quality* (AQ) is scored by UTMOSv2 (Baba et al., 2024). *Audio-visual synchrony* (AVS) is scored by SyncNet as implemented in LatentSync (Li et al., 2025). *Instruction compliance* (IC), *video fidelity* (VF) and *AV quality* (AVQ) are judged by Qwen3-Omni (Xu et al., 2025): whether the edit executes the instruction, whether unrelated content is preserved, and the overall joint quality of the edited clip.

InsAVE-80K. Six metrics span video, audio and audio-visual properties. *Fréchet Video Distance* (FVD) measures visual realism and spatiotemporal coherence using a VideoMAE (Tong et al., 2022) backbone. *Fréchet Audio Distance* (FAD) is the analogous measure for audio, over VGGish (Diwakar & Gupta, 2024) embeddings. *Text-video alignment* (TV-A) and *text-audio alignment* (TA-A) measure agreement between the instruction and each generated modality, using VideoCLIP-XL (Wang et al., 2024) and CLAP (Wu et al., 2023b). *Audio-video alignment* (AV-A) measures semantic coherence between the two generated streams using ImageBind (Girdhar et al., 2023). *PEAVS* predicts a mean opinion score for audio-visual synchrony (Goncalves et al., 2024).

VGGSound. Four metrics are reported under the protocol of Pham et al. (2026). FVD (Unterthiner et al., 2019) and FAD (Kilgour et al., 2019) measure video and audio quality, DeSync (Iashin et al., 2024) measures temporal synchronization, and the ImageBind score (IB) (Girdhar et al., 2023) measures semantic alignment. Because the Fréchet distances are computed against ground truth, FVD requires generated video and is reported only for audio-to-video, while FAD requires generated audio and is reported only for video-to-audio; DeSync and IB are evaluated for both directions. The empty cells in Tab. 3 reflect this, rather than missing measurements.

Verse-Bench. Six metrics are reported following Huang et al. (2025). *Motion score* (MS) is computed from RAFT (Teed & Deng, 2020) optical flow. *Aesthetic score* (AS) averages MANIQA (Yang et al., 2022), aesthetic-predictor-v2-5 and Musiq (Ke et al., 2021). *Production quality* (PQ) comes from AudioBox-Aesthetics (Tjandra et al., 2025). *Text-audio consistency* (CS) is a LAION-CLAP (Wu et al., 2023b) score. *Lip-sync confidence* (LSE-C) is produced by SyncNet (Chung & Zisserman, 2016). *Word error rate* (WER) is computed by transcribing generated audio with Whisper-large-v3 (Radford et al., 2023). LSE-C is scored only on human-speech samples, since the metric requires visible mouth motion.

A2 Additional Experimental Results

A2.1 Qualitative results

Fig. A1 illustrates joint audio-video generation on Verse-Bench, where both streams are generated simultaneously from text prompts alone to yield synchronized soundtracks aligned with the visual scene. Figs. A2 and A4 compare UniAVC against baseline methods across five task families on matched inputs, demonstrating accurate instruction-guided editing while leaving unmentioned modalities unchanged.

A2.2 Quantitative results

Tab. A2 further breaks the aggregate of Tab. 2 down by task family, where UniAVC attains the best or second best score in the majority of cells.



Figure A1: Qualitative results for joint audio-video generation on VerseBench. Captions are omitted for brevity.

A2.3 Ablations

Tab. 4 isolates individual design factors relative to the default configuration, while Fig. A3 illustrates the performance of both inference modes on matched inputs. We observe that intent-aware editing and self-refinement modes both allow for high-quality editing or enhancing target modality, while losslessly preserving non-target counterpart and maintaining their alignment.

A2.4 Limitations

UniAVC incurs notable computational overhead during both training and inference. The two-stage recipe requires substantial resources, consuming large-scale unimodal data during pretraining before alignment finetuning begins. Inference latency also precludes real-time deployment because every request requires full denoising trajectories across both streams, with self-refinement adding a second pass over the target modality. Furthermore, performance is constrained by the complexity of instruction prompts in the training set. Because paired corpora contain few compositional requests specifying multiple concurrent edits, the model may execute multi-goal instructions imperfectly. Future work will focus on reducing staged training costs, accelerating inference trajectories, and expanding instruction mixtures to improve compositional editing.

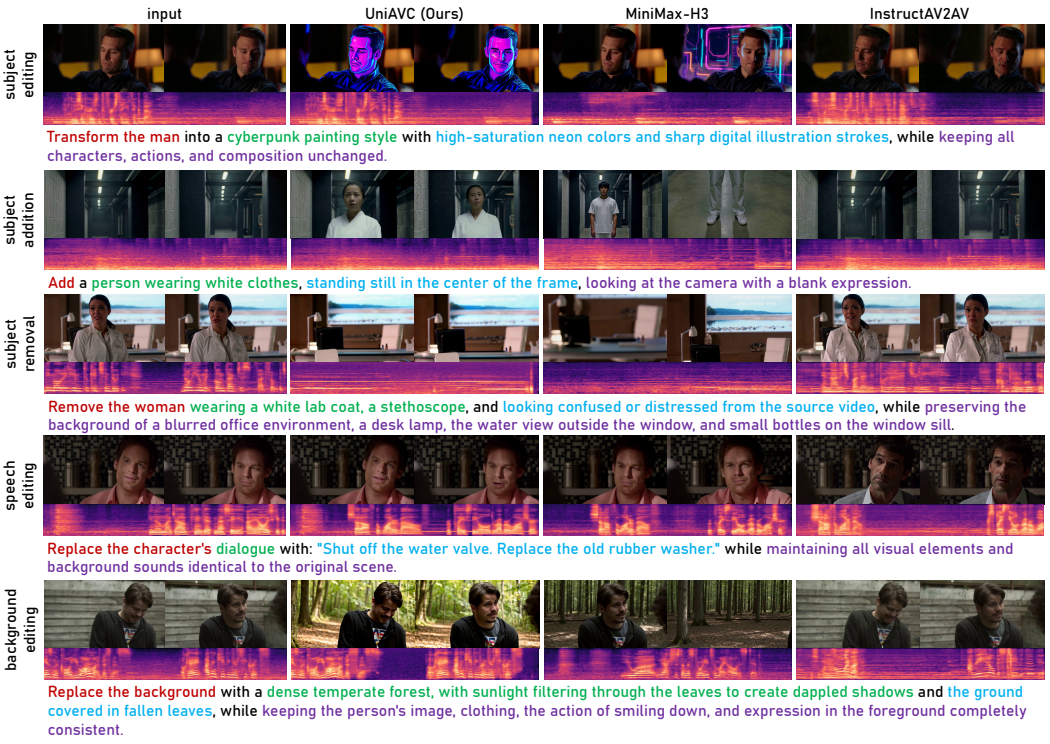

Figure A2: Additional qualitative comparisons for joint editing on JAVEditBench.

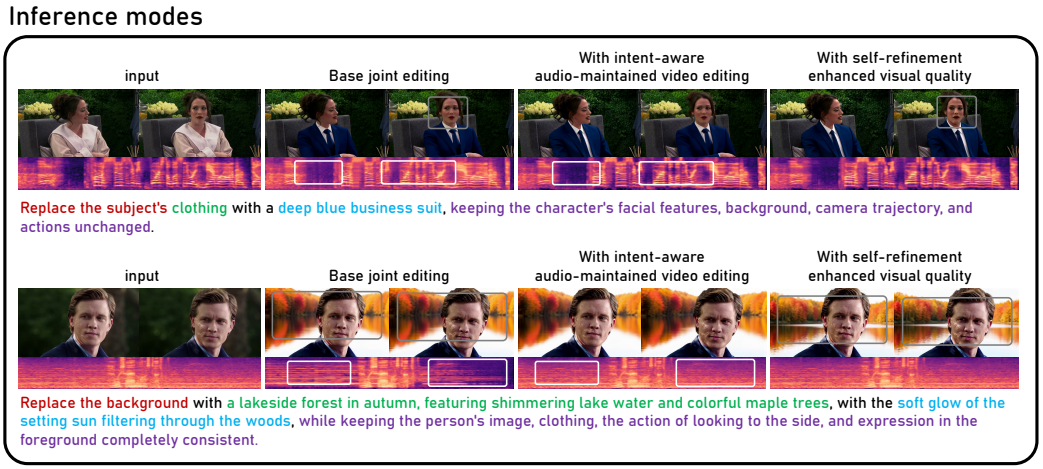

Figure A3: Inference modes: intent-aware editing preserves the pinned stream unchanged, while self-refinement improves one branch against the accepted other.

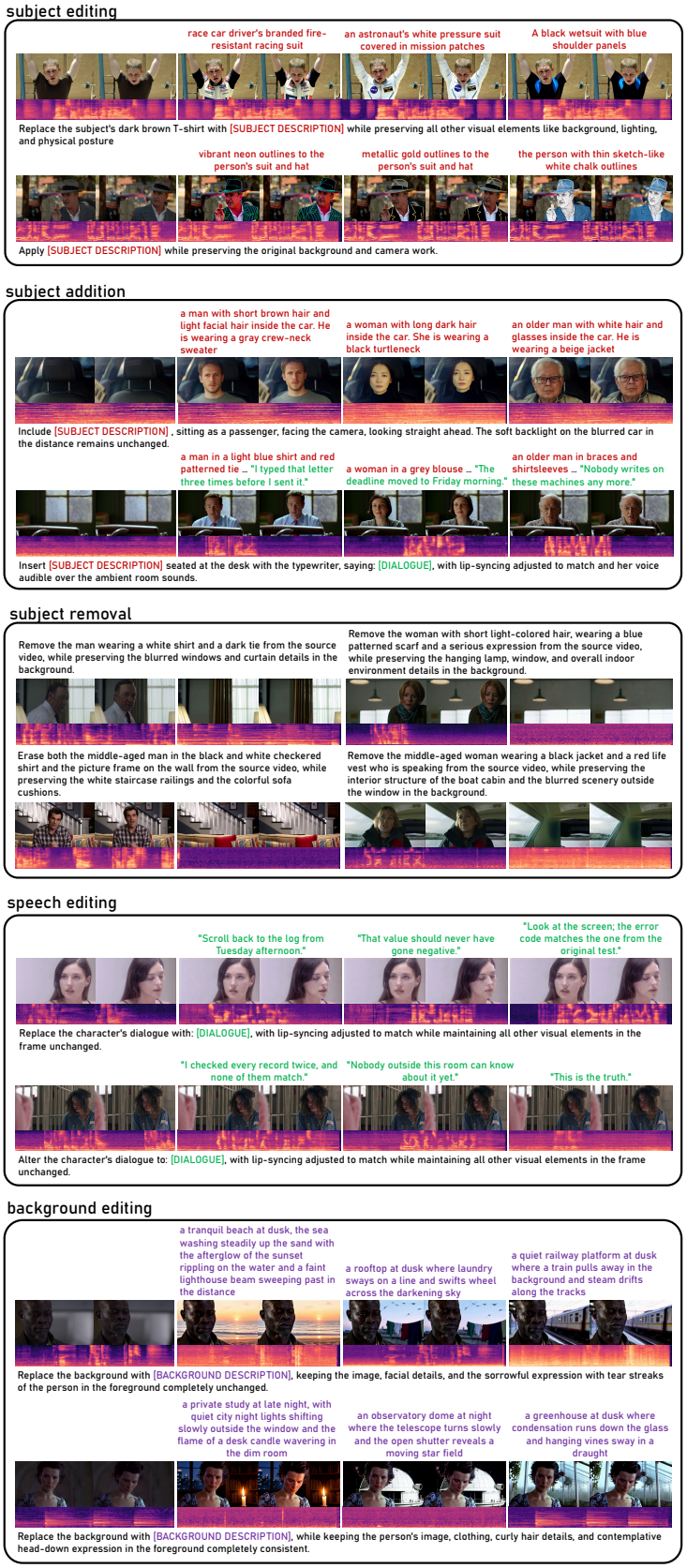

Figure A4: Audio-video editing versatility of UniAVC across five task families.

Table A2: Per-task results on JAVEditBench. Best per column in **bold** and second best underlined.

Metric	Method	Overall	Sub.Edit	BG.Edit	Speech	Sub.Add	Sub.Rm
VQ↑	AVED	0.0590	0.0603	0.0612	0.0603	0.0551	0.0562
	AVI-Edit	**0.0604**	0.0625	**0.0650**	0.0599	0.0492	**0.0638**
	JAVEdit	0.0596	0.0609	0.0636	0.0625	**0.0633**	0.0447
	InstructAV2AV	0.0588	0.0620	0.0618	0.0619	0.0417	0.0628
	MiniMax-H3	0.0562	0.0605	0.0553	0.0551	0.0579	0.0490
	UniAVC (ours)	0.0599	**0.0625**	0.0624	**0.0627**	0.0613	0.0416
AQ↑	AVED	1.72	1.71	1.49	1.96	1.62	1.75
	AVI-Edit	2.34	2.43	2.28	2.33	2.17	**2.44**
	JAVEdit	2.42	**2.48**	2.23	2.62	2.48	2.24
	InstructAV2AV	1.40	1.24	1.35	2.05	1.13	1.20
	MiniMax-H3	2.36	2.25	2.20	**2.66**	**2.52**	2.21
	UniAVC (ours)	**2.59**	2.44	**2.32**	2.64	2.25	2.29
AVS↑	AVED	0.1641	0.1333	0.0613	0.1050	**0.5459**	0.1367
	AVI-Edit	0.2721	0.3236	**0.3584**	0.2453	0.0365	**0.2658**
	JAVEdit	**0.3688**	**0.3779**	0.3097	**0.4569**	0.3264	0.0146
	InstructAV2AV	0.1759	0.1681	0.1531	0.2212	0.0293	0.1661
	MiniMax-H3	0.3009	0.3124	0.2614	0.3836	0.2379	0.2635
	UniAVC (ours)	0.2975	0.2972	0.2764	0.2625	0.2500	0.0544
IC↑	AVED	2.95	2.60	2.87	3.67	2.20	3.50
	AVI-Edit	3.49	3.64	3.63	3.30	3.20	3.58
	JAVEdit	4.07	**4.22**	3.83	4.33	3.88	3.96
	InstructAV2AV	2.88	3.00	2.47	4.10	1.50	3.12
	MiniMax-H3	4.15	4.10	3.97	**4.40**	**4.24**	**4.04**
	UniAVC (ours)	**4.42**	4.12	**4.23**	4.23	4.12	4.04
VF↑	AVED	3.87	3.92	4.07	4.00	3.48	3.79
	AVI-Edit	3.89	4.10	3.93	3.70	3.72	3.88
	JAVEdit	4.22	4.30	4.07	4.33	4.04	4.30
	InstructAV2AV	3.93	4.25	**4.10**	3.80	4.04	3.22
	MiniMax-H3	4.17	4.25	3.97	**4.43**	4.04	4.08
	UniAVC (ours)	**4.48**	**4.42**	4.07	4.40	**4.16**	**4.36**
AVQ↑	AVED	2.93	3.10	3.13	3.73	1.72	2.68
	AVI-Edit	3.86	4.03	3.93	3.87	3.40	**3.96**
	JAVEdit	**3.88**	**4.08**	**4.07**	3.63	3.72	3.82
	InstructAV2AV	3.61	3.60	3.63	3.90	3.28	3.60
	MiniMax-H3	3.85	4.03	3.80	**4.10**	3.84	3.32
	UniAVC (ours)	3.85	3.98	3.90	3.93	**4.04**	2.44