KNOWING WHOM TO ERASE: INSTRUCTION-ROUTED AUDIO-VISUAL SUBJECT REMOVAL FOR HUMAN-CENTRIC SPEECH AND MUSIC VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual subject removal erases the person specified by an instruction and the sound they produce, while preserving other people and their sounds. Existing audio-visual editors learn text- or mask-conditioned edits from paired examples, but matching individual edits does not establish reliable target switching within the same recording. In multi-person scenarios, existing editors struggle to jointly remove the requested person and preserve competing sources. Single-target supervision can contribute to this failure by permitting an instruction-agnostic shortcut: when every instruction for an input refers to the same target, a fixed edit can fit the training data without learning whom to remove. To address this problem, we propose Audio-Visual Instruction-Routed Erasure (AV-IRE), a joint audio-video editor combining complementary-target supervision with explicit target routing. Source recordings and their teacher-edited counterparts are composed to supervise different target-specific outputs for the same input. The target is conveyed to the generator through a spatial routing map and modality-specific codes, while its text prompt remains target-neutral. Route-dependent output changes are supervised by a shared-state route-intervention loss. We also construct 54,705 automatically filtered, teacher-generated removal pairs from real recordings for task adaptation. AV-IRE outperforms strong audio-visual editing baselines in selective multi-person removal. Compared with Task-Adapted LTX-2.5, blind-rated joint success rises from 11.7% to 41.7% on natural two-musician recordings. On new recordings, stricter bidirectional success, requiring both target requests per input to succeed, rises from 3.3% to 26.7%. Audio-video samples are available at https://avire-removal.github.io/AV-IRE/.

1 INTRODUCTION

Joint audio-video generators (HaCohen et al., 2026; Lightricks, 2026) provide a shared backbone for editing both modalities under one instruction. We study *instruction-guided subject removal*: given a recording and target phrase, remove the specified speaker or performer and their sound while preserving all other people and sounds. For music, removal also includes the performer's instrument. In multi-person scenes, target suppression alone cannot distinguish selective editing from indiscriminate removal. Moreover, performers occupy distinct image regions, whereas their sounds overlap in time and frequency. Editors must therefore associate the requested person with their acoustic contribution (Zhao et al., 2018), rather than treat a visual region as an audio mask.

Recent approaches include video-conditioned audio editing (Ishii et al., 2025), mask-conditioned object removal (Xu et al., 2025), and instruction-guided joint editing (Zheng et al., 2026; Chen et al., 2026). Success on one request, however, does not establish target switching within the same recording: a source removed under one request must be preserved when another is selected.

Single-target supervision permits an input-only shortcut: when all instruction variants for a recording share the same target and edited output, a model can fit the training pairs without distinguishing targets. This is a form of shortcut learning (Geirhos et al., 2020): fitting the training pairs need not establish instruction-dependent target selection. Following counterfactually augmented data, which discourages spurious associations through minimally revised examples with changed la-

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.

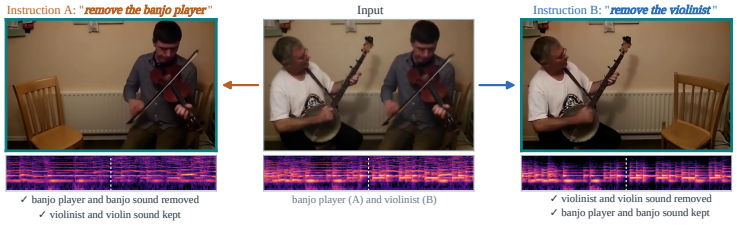

Figure 1: Target switching on the same recording. For each instruction, AV-IRE removes the named performer together with their instrument and sound while preserving the other performer and their sound; strips show the soundtracks as spectrograms. Appendix J gives the example-selection rule.

bels (Kaushik et al., 2020), we hold the recording fixed and vary the target request and edited output. This makes target dependence explicit in supervision and evaluation.

We propose *Audio-Visual Instruction-Routed Erasure* (AV-IRE), built on LTX-2.5 with a removal-specific LoRA (Hu et al., 2022) to edit both streams through joint denoising. We compose donor recordings and their teacher-edited counterparts to pair each input with two outputs, each removing a different target while preserving the other source. A spatial map and modality-specific codes route the target to the generator, whose text prompt stays target-neutral. Complementary outputs specify what to edit; a shared-state route-intervention objective supervises how generation responds to route changes with the reference input and noisy latent state fixed. An offline teacher pipeline supplies automatically filtered removal pairs from real recordings for task adaptation.

We introduce *AV-SwapTarget*, a benchmark testing different target requests on the same input across natural multi-musician recordings and controlled multi-source composites. Its primary selection test, Natural-2, contains 30 natural two-musician recordings with two requests each, yielding 60 queries. Under blind review, *joint success* requires visual and acoustic target removal, preservation of other people and sounds, and audio-visual consistency. *Bidirectional success* requires both requests for the same input to succeed. Figure 1 illustrates this behavior: AV-IRE removes each requested performer, instrument, and sound while preserving the other performer and their sound.

Compared with Task-Adapted LTX-2.5, our same-backbone control conditioned on the target phrase, AV-IRE raises blind-rated joint success from 11.7% to 41.7% on Natural-2. Visual target removal rises from 60.0% to 88.3%, and preservation of other people from 51.7% to 78.3%. The preregistered primary endpoint, bidirectional success, rises from 0.0% to 16.7% without reaching the prespecified significance threshold; the secondary graded score, counting criteria passed across both requests, improves significantly. A second independent rater reproduces the ordering. In a preregistered follow-up on 30 fresh recordings evaluated after model freezing, bidirectional success rises from 3.3% to 26.7%, a 23.3-percentage-point gain meeting the prespecified threshold. These results support complementary supervision and explicit routing for selective joint editing.

Our contributions are summarized as follows:

- AV-IRE is proposed as an instruction-routed joint audio-video editor combining a target-neutral generator prompt, explicit spatial routing, and modality-specific conditioning with shared-state route-intervention training for selective subject removal.
- Complementary-target supervision is introduced to train different removal outcomes for the same input, and AV-SwapTarget to evaluate instruction-dependent target switching across natural multi-musician recordings and controlled multi-source composites.
- An offline teacher pipeline and a provenance-tracked corpus of 54,705 automatically filtered removal pairs from real recordings are developed for joint audio-video adaptation.

2 Related Work

Joint audio-visual editing. Audio-visual editing follows sequential or joint approaches. CoherentAVEdit (Ishii et al., 2025) edits video first, then generates audio conditioned on the source audio, edited video, and text prompt. TV-AudioRemover (Guo et al., 2026) removes sound associated with a visually erased object by conditioning an audio editor on the edited video, original mixture, and instruction. Joint approaches edit both streams within one model. SAVE (Xu et al., 2025) performs parallel audio-visual object removal with a flow-matching Schrödinger bridge conditioned

on text and a target mask. InstructAV2AV (Zheng et al., 2026) adapts an audio-video generation backbone for instruction-guided editing, balancing source content and instructions through source-instruction gated attention. JAVEdit (Chen et al., 2026) adapts LTX-2.3 with LoRA for reference-conditioned editing of human-centric videos. AV-IRE instead conveys the requested target to the generator through an explicit route comprising a predicted spatial map and modality-specific codes, rather than a target-bearing generator text prompt or user-provided mask.

Spatial control and preservation. Local editing aims to confine changes while preserving unrelated content. Prompt-to-Prompt (Hertz et al., 2023) spatially controls text-driven image edits by manipulating cross-attention maps. Video object removal reconstructs the scene behind a masked object while addressing sampling artifacts (Zi et al., 2025), shadow effects (Miao et al., 2025), or imperfect masks (Hu et al., 2026). Instruction-guided video editors including Kiwi-Edit (Lin et al., 2026) support broader edits. AdaptEdit (Cai et al., 2026) augments a frozen diffusion transformer with region-aware adapters and a learned spatial gate to predict edit regions from instructions without inference-time user masks. Like AdaptEdit, AV-IRE combines a frozen generator, residual adaptation, and predicted spatial control in video. It extends this interface to joint audio-video editing through modality-specific conditioning and shared-state route-intervention training.

Audio-visual grounding and separation. Visually guided separation extracts audio mixture components associated with visible sources (Zhao et al., 2018; Ephrat et al., 2018). SAM Audio (Shi et al., 2025) separates targets using text, visual masks, or temporal prompts, while music source separation recovers predefined stems such as vocals, drums, and bass (Luo & Yu, 2023; Lu et al., 2024). Removing these estimated sources can yield edited audio but does not remove the corresponding subject from the video. AV-IRE shares the goal of visual-acoustic association but generates both edited streams in a single joint denoising process without explicitly producing separated stems.

Supervision and evaluation of target-dependent editing. Instruction-based editors commonly learn from model-generated pairs (Brooks et al., 2023). Synthesized and curated audio-visual datasets include SAVEBench (Xu et al., 2025), InsAVE-80K (Zheng et al., 2026), and JAVEdit-100k (Chen et al., 2026). Compositional supervision underlies mix-and-separate training on mixed recordings (Zhao et al., 2018) and SAVEBench's object-grounded removal pairs. Inspired by counterfactually augmented data, which discourages spurious associations through minimal edits and changed labels (Kaushik et al., 2020), we compose donor recordings and their teacher-edited counterparts into complementary removal targets for one input. These targets supervise route-dependent generation with fixed reference input and noisy latent state. AVE-Compass (Wen et al., 2026) assesses instruction following, fidelity preservation, realism, and editing intent in audio-video editing; AV-Remove-Bench (Guo et al., 2026) evaluates sounding-object removal across speech, music, and sound effects. AV-SwapTarget complements these benchmarks by testing each source's removal in turn; bidirectional success requires both requests on a two-source input to succeed.

3 METHOD

3.1 TASK AND COMPLEMENTARY TARGETS

Given an audio-visual recording $X = (V, A)$ with video V, audio A, and target phrase q, the editor F produces $(\hat{V}, \hat{A}) = F(X, q)$, removing the named person, their instrument if present, and their sound while preserving all other visible and audible sources. Phrases are category-level; selecting among same-category performers is outside our scope. Neither a target mask nor a retained-source description is provided at inference. Success requires five binary judgments: visual and acoustic target removal (j_1, j_2), preservation of other people and sounds (j_3, j_4), and audio-visual consistency (j_5). For two eligible sources specified by q_A and q_B,

$$J(X, q) = \prod_{k=1}^{5} j_k(X, q), \qquad P(X) = J(X, q_A)\, J(X, q_B), \tag{1}$$

where joint success J requires all five criteria and bidirectional success P requires both requests to succeed.

If all instruction variants for X_i share target output Y_i, the input-only mapping $H(X_i) = Y_i$ fits every corresponding training pair. To expose this shortcut through supervision (Kaushik et al., 2020),

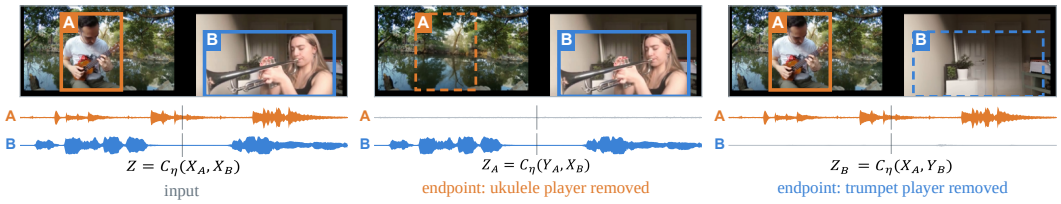

Figure 2: Complementary-target supervision. A shared composition produces input Z and target-specific outputs Z_A and Z_B, each removing one donor's subject and sound while retaining the other donor.

we assign one input two distinct target outputs. Given source recordings (donors) X_A, X_B and their teacher-removed counterparts Y_A, Y_B,

$$Z = C_\eta(X_A, X_B), \qquad Z_A = C_\eta(Y_A, X_B), \qquad Z_B = C_\eta(X_A, Y_B), \tag{2}$$

here Z is the composite input containing both sources; Z_A and Z_B are its removal targets for sources A and B, respectively, each leaving the other donor unchanged. Following mix-and-separate training (Zhao et al., 2018), C_η places donor videos in separate screen regions and mixes their audio; η collects the layout, source gains, and peak-scaling factor shared by all three compositions. Figure 2 shows a ukulele–trumpet example. Under a fixed random state, an editor ignoring q cannot reproduce both outputs from Z.

3.2 Routing the Target

We use LTX-2.5 (HaCohen et al., 2026; Lightricks, 2026), adapted for removal using LoRA (Hu et al., 2022). Figure 3 shows how target phrases condition this joint generator through explicit routing.

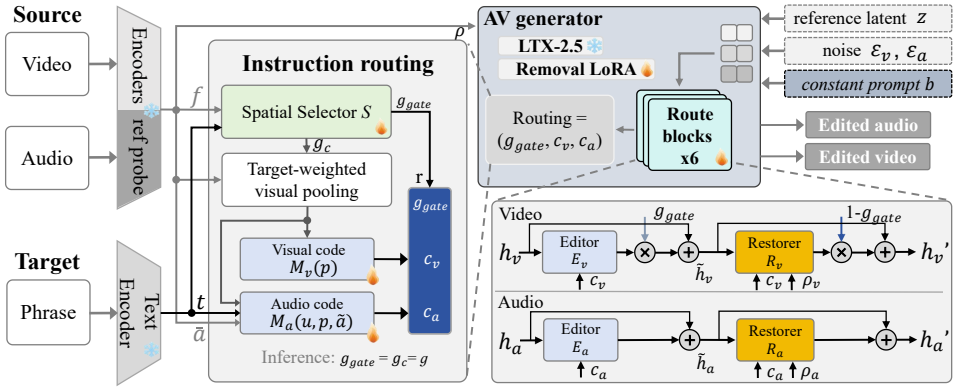

Figure 3: AV-IRE architecture. The target phrase conditions the joint generator through a spatial routing map and modality-specific codes, while the text prompt stays target-neutral. Residual route blocks apply spatially weighted video edits and unmasked audio edits.

A feature-extraction run without the trainable executor (the code networks and route blocks) extracts block-32 video features $f = \{f_n\}_{n=1}^N$, averaged over latent frames at $N = 160$ spatial positions, and a mean audio feature $\bar{a}$. A frozen text encoder maps the target phrase to t. The generator receives a constant, target-neutral prompt b, so instruction-specific target conditioning enters through the explicit route. The spatial selector predicts

$$g_n = \text{sigmoid}\left(S([P_v(f_n); P_t(t)])\right). \tag{3}$$

Here g_n is the target weight at spatial position n, $[\cdot;\cdot]$ denotes concatenation, P_v, P_t are layer-normalization–projection maps, and S is an MLP. We supervise the selector on box-derived regions in natural two-person instrument clips. Region-balanced binary cross-entropy trains localization; an instruction-swap term encourages each source region to score higher under its own phrase than under the other phrase; and a background term penalizes changes between the two queries outside both sources. The selector is then frozen during executor training; Appendix D gives the loss and region-labeling rules.

The map pools a target feature,

$$p = \textstyle\sum_n g_n f_n \big/ \sum_n g_n, \tag{4}$$

with the spatial mean used when map mass is at most 10^{-6}. Learned projections P'_t, P_a, P_h, feature-wise scale and shift functions γ, β (Perez et al., 2018), and MLPs M_v, M_a produce visual and audio route codes c_v, c_a:

$$u = P'_t(t), \quad \tilde{a} = (1 + \gamma(u)) \odot P_a(\bar{a}) + \beta(u), \quad c_v = M_v(p), \quad c_a = M_a([u; P_h(p); \tilde{a}]). \tag{5}$$

Here u is the projected target-text feature and $\tilde{a}$ the phrase-modulated audio feature; $\odot$ denotes elementwise multiplication. The visual code depends on the phrase only through the spatial map; the audio code combines text, pooled visual features, and audio context. The route tuple $r = (g_{\text{gate}}, c_v, c_a)$ comprises the execution map and both codes. We distinguish g_{gate} from the pooling map g_c, denoted g in Equation (4). Training interventions can vary these maps independently; at inference, $g_{\text{gate}} = g_c = g$. Both 256-dimensional codes are broadcast over tokens, and the execution map over latent frames and channels.

3.3 Joint execution

At six backbone depths, route blocks add residual edits to generated video and audio token states h_v, h_a, using depth-aligned reference states ρ_v, ρ_a. These token-wise reference states are distinct from the pooled probe features used to construct the route:

$$\begin{aligned} \tilde{h}_v &= h_v + g_{\text{gate}} \odot E_v([h_v; c_v]), & h'_v &= \tilde{h}_v + (1 - g_{\text{gate}}) \odot R_v([\tilde{h}_v; \rho_v; c_v]), \\ \tilde{h}_a &= h_a + E_a([h_a; c_a]), & h'_a &= \tilde{h}_a + R_a([\tilde{h}_a; \rho_a; c_a]). \end{aligned} \tag{6}$$

For modality $m \in \{v, a\}$, $\tilde{h}_m$ and h'_m are the intermediate and output states. The editing branch E_m and reference-conditioned preservation branch R_m are two-layer MLPs trained jointly through edited-output supervision. Their output projections are zero-initialized, leaving the task-adapted generator unchanged at initialization under identical prompts and sampling states (Zhang et al., 2023). In video, the map weights the editing residual toward the target and the preservation residual toward the remainder, while subsequent attention can propagate changes across regions. In audio, target and retained sources can sound simultaneously: a complementary temporal gate would suppress preservation when both are active. Both audio branches therefore use c_a without masking, with token-state inputs enabling token-dependent edits.

3.4 Training objective and inference

Training has three stages: task-LoRA adaptation on teacher-generated pairs with their instructions; separate selector training; and training only the code networks and route blocks, with the backbone, LoRA, and selector frozen and the prompt fixed to b.

Target flow matching. Let $z = (z_v, z_a)$ and $y = (y_v, y_a)$ encode the input and edited target. Gaussian noise $\varepsilon_m \sim \mathcal{N}(0, I)$ and noise level σ_m define $x_{\sigma_m, m} = (1 - \sigma_m) y_m + \sigma_m \varepsilon_m$ and target velocity $v^\star_m = \varepsilon_m - y_m$, with I the identity covariance. Write $x_\sigma = (x_{\sigma_v, v}, x_{\sigma_a, a})$ and $\boldsymbol{\sigma} = (\sigma_v, \sigma_a)$. Ordinary executor updates use flow matching (Lipman et al., 2023):

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\sum_{m \in \{v,a\}} \alpha_m \left\langle \left(v_{\theta,m}(x_\sigma, \boldsymbol{\sigma}; z, b, r) - v^\star_m \right)^2 \right\rangle_{\mathcal{M}_m} \right], \tag{7}$$

here $v_{\theta,m}$ is the joint model's modality-m velocity prediction, with trainable executor parameters θ, reference latents z, fixed prompt b, and route $r = (g_{\text{gate}}, c_v, c_a)$. The expectation averages over training examples, noise levels, and Gaussian draws; $\alpha_v = \alpha_a = 1$. The supervision mask $\mathcal{M}_m$ selects tokens and channels for the mean $\langle \cdot \rangle_{\mathcal{M}_m}$, defined as zero for an empty mask. Audio-only remix targets use an empty video mask, so unchanged video is not supervised as a removal target.

Shared-state route interventions. Microbatches from the original complementary composites use either flow matching or route intervention, not their sum; all others use flow matching. For route intervention, two model evaluations share the *same* noised input $\xi_{\sigma,m} = (1 - \sigma) z_m + \sigma \varepsilon_m$, changing only the route to obtain one-step clean-latent estimates:

$$\hat{y}^{(i)}_m = \xi_{\sigma,m} - \sigma \, v_{\theta,m}(\xi_\sigma, \sigma; z, b, r^{(i)}), \qquad i \in \{1, 2\}. \tag{8}$$

Here $\xi_\sigma = (\xi_{\sigma,v}, \xi_{\sigma,a})$ uses the same noise level σ for both modalities, and $r^{(i)}$ denotes the route in evaluation i. Unlike flow matching, which noises the target, this auxiliary update regresses edited

targets from a shared noised input. The reference, prompt, noise level, and per-modality noise draws are identical across evaluations, isolating the route change.

Let g_s be the execution map and $c_s = (c_{v,s}, c_{a,s})$ the codes for source $s \in \{A, B\}$. We sample diagonal routes $(g_A, c_A), (g_B, c_B)$ or crossed routes $(g_A, c_B), (g_B, c_A)$ with equal probability. Each evaluation's video target follows its map and its audio target its codes. Thus, (g_A, c_B) is supervised to remove A visually and B acoustically. Crossed assignments are used only during training; at inference, the map and codes specify the same target for joint removal.

Difference and midpoint matching. Let $y_m^{\star(i)}$ be the target latent assigned to evaluation i. Define the predicted difference $\hat{d}_m = \hat{y}_m^{(1)} - \hat{y}_m^{(2)}$ and midpoint $\hat{\mu}_m = \frac{1}{2}(\hat{y}_m^{(1)} + \hat{y}_m^{(2)})$, with target counterparts $d_m^\star = y_m^{\star(1)} - y_m^{\star(2)}$ and $\mu_m^\star = \frac{1}{2}(y_m^{\star(1)} + y_m^{\star(2)})$. The route loss is

$$\mathcal{L}_{\text{route}} = \frac{\lambda_{\text{route}}}{2} \sum_{m \in \{v,a\}} w_m \Big[\frac{\langle (\hat{d}_m - d_m^\star)^2 \rangle_m}{D_m} + \lambda_{\text{mid}} \frac{\langle (\hat{\mu}_m - \mu_m^\star)^2 \rangle_m}{B_m} \Big], \tag{9}$$

where λ_{route} scales the loss, λ_{mid} weights midpoint matching, and w_m are modality weights distinct from α_m. The mean $\langle \cdot \rangle_m$ covers the union of source regions in video and all output tokens in audio, including channels. The stop-gradient operator sg preserves its argument's forward value but blocks gradients through it. The scale normalizers are

$$D_m = \text{sg}\big(\langle (d_m^\star)^2 \rangle_m\big) + \epsilon, \qquad B_m = \max\big(\text{sg}(\langle (\mu_m^\star - z_m)^2 \rangle_m),\, 10^{-4}\,\text{sg}(\langle (d_m^\star)^2 \rangle_m)\big) + \epsilon,$$

where $\epsilon = 10^{-12}$; pairs with zero target-difference energy are rejected. Difference matching supervises route-dependent changes. An identical offset in both predictions cancels in their difference, so midpoint matching additionally constrains this shared error. The normalizers account for target-dependent scales; together with λ_{mid}, they determine the relative weighting of the two terms. Both terms reuse the complementary targets without additional edited outputs. Appendix D details the route assignments and loss decomposition.

Inference. The selector and code networks construct a route from the input recording and target phrase. This route remains fixed throughout joint audio-video generation; Appendix C gives the feature-extraction and sampling configuration.

4 Experiments

4.1 Data and experimental setup

Teacher corpus. We construct 54,705 instruction-conditioned removal pairs from real OpenHumanVid, AVSpeech, MUSIC-AVQA, and VGGSound recordings (Li et al., 2025; Ephrat et al., 2018; Li et al., 2022; Chen et al., 2020). The offline cascade combines SVOR (Hu et al., 2026), edited-video-conditioned TV-AudioRemover (Guo et al., 2026), and a SAM Audio replacement branch (Shi et al., 2025). This corpus, with 89.96% speech-removal pairs, supports removal-task adaptation; subsequent complementary-target training supervises multi-source target selection. Automatic admission tests suppression, not preservation, without systematic human auditing. Appendix A details the pipeline and full source and target distributions.

Selector and executor data. The selector uses 1,739 box-labeled natural two-person instrument clips. The executor uses 1,836 query-target examples: 1,200 from complementary composites, 440 from additional pairs, and 196 from single-target natural clips. Appendix B details construction.

Evaluation subsets. AV-SwapTarget tests each source separately in Natural-2 and Natural-3 recordings with two and three musicians, and Composite-2 and Composite-N mixtures with two and two to five sources. Natural-2 is primary; all subsets exclude component-training and validation source videos. Single-speaker transfer uses JAVEditBench (Chen et al., 2026) subject removal and InsAVE (Zheng et al., 2026) person removal.

Baselines and metrics. One blind review round covers six configurations on 60 Natural-2 queries: Task-Adapted LTX-2.5 with the target phrase in its prompt, AV-IRE, its box-map oracle variant, and three ablations. InstructAV2AV (Zheng et al., 2026) was rated earlier; Kiwi-Edit → CoherentAVEdit (Lin et al., 2026; Ishii et al., 2025) and the unadapted backbone use automatic evaluation. SAVE (Xu et al., 2025) and JAVEdit (Chen et al., 2026) have no released weights. Metrics include TSSR and PSF with removal and preservation criteria (Guo et al., 2026); SAJ, PQ, IB-AV,

and DeSync (Shi et al., 2025; Tjandra et al., 2025; Girdhar et al., 2023; Iashin et al., 2024); and composite PROVE and slot-masked PSNR, SSIM, and LPIPS (Li et al., 2026; Zhang et al., 2018). Appendices I and E.5 give settings and definitions.

Blind review and statistics. Blind review covers Natural-2, InsAVE, and ten JAVEditBench clips. Independent non-author raters without project involvement see inputs and targets but not system identities, scoring Equation (1)'s five criteria. Partial judgments fail; wrong-source removal is recorded separately. The preregistered Natural-2 AV-IRE–Task-Adapted LTX-2.5 comparison uses first-rater judgments and recording-level pairs: exact McNemar for primary P and paired Wilcoxon for the secondary 10-point graded score. InsAVE tests J. Appendix E details protocols and statistics.

Implementation details. Inference uses two feature-extraction and 30 sampling steps, producing 121 frames at 512×320 and 30 fps. The 53.0M-parameter executor trains for 1,000 updates on composites, then 1,000 on the full set. Appendix C gives schedules and hyperparameters.

4.2 Comparison with baselines

Table 1: Blind review: 30 recordings, two requests each per panel. Joint (J): all five criteria pass; Both (P, primary): both requests pass; Score (secondary): mean criteria passed per recording. Other counts are out of 60; C: audio-visual consistency. Bold: column best per panel, including ties, not significance. †: earlier review round; panel (a) otherwise uses the first reviewer.

	Overall ↑			Removal ↑		Keep others ↑		AV ↑	Wrong source ↓	
Method	Joint (/60)	Both (/30)	Score (0–10)	Video	Audio	Video	Audio	C	Video	Audio
(a) Natural-2: main comparison										
InstructAV2AV†	1	0	3.67	6	33	30	24	17	13	**14**
Task-Adapted LTX-2.5	7	0	4.53	36	36	31	13	20	7	19
AV-IRE (ours)	**25**	**5**	**6.83**	**53**	**37**	**47**	**36**	**32**	**2**	**14**
(b) Natural-2: same outputs, second reviewer										
Task-Adapted LTX-2.5	5	1	4.70	35	**40**	34	13	19	24	30
AV-IRE (ours)	**26**	**6**	**7.00**	**52**	**40**	**48**	**37**	**33**	**9**	**17**
(c) Fresh recordings: new reviewer, frozen models										
Task-Adapted LTX-2.5	7	1	4.23	41	35	23	11	17	31	43
AV-IRE (ours)	**34**	**8**	**8.07**	**55**	**40**	**56**	**48**	**43**	**4**	**10**

Blind review on AV-SwapTarget. Table 1 shows AV-IRE increases joint success from 11.7% to 41.7% over Task-Adapted LTX-2.5 on 60 Natural-2 queries. It improves visual removal and non-target preservation, raising acoustic preservation from 21.7% to 60.0%. Similar acoustic removal (61.7% versus 60.0%) and fewer wrong-performer visual edits indicate improved selectivity, consistent with complementary-target supervision and spatially routed, reference-conditioned edits. InstructAV2AV's 10.0% visual removal indicates a bottleneck. Across 30 inputs, preregistered bidirectional success rises from 0.0% to 16.7% without statistical significance ($p = 0.0625$); the graded secondary endpoint improves from 4.53 to 6.83 ($p = 4.9 \times 10^{-4}$).

Second rater and fresh test set. Table 1 presents two preregistered follow-ups supporting AV-IRE. Second-reviewer ratings reproduce this ordering across 120 AV-IRE and Task-Adapted LTX-2.5 outputs (joint-success agreement $\kappa = 0.85$). Bidirectional success remains non-significant ($p = 0.0625$); the graded score improves ($p = 1.1 \times 10^{-4}$). The fresh set comprises 30 recordings admitted after model freezing, excluding training and development source videos. A new independent rater finds bidirectional success of 26.7% versus 3.3% ($p = 0.039$), meeting the prespecified threshold. The fresh-set primary analysis retains all inputs, including five with mislabeled targets. Appendix E.2 details admission and sensitivity analyses; Appendix H records preregistration times.

Automatic metrics. Among the four editors in Table 2, AV-IRE leads on preservation, PSF, three of four judge fields, and IB-AV, with production quality comparable to Task-Adapted LTX-2.5. The Kiwi cascade leads on removal and TSSR but ranks lowest on PQ and IB-AV. AST-based removal measures target-class suppression, whereas PQ assesses audio quality (Tjandra et al., 2025) and IB-AV cross-modal semantic correspondence (Girdhar et al., 2023). Thus, suppression can score well despite degraded or mismatched remaining audio. Separate video editing followed by conditioned audio generation (Ishii et al., 2025) is a plausible source of this mismatch. Conversely, the unadapted

Table 2: Automatic scores on the Natural-2 outputs of Table 1; TSSR and PSF are medians over valid queries, other scores are means. Bold/underline: best/second best among the four editors; the no-op reference and the oracle are not ranked.

	AST-based				SAM Audio Judge				Other	
Method	rem./60	kept/60	TSSR ↑	PSF ↑	Ov. ↑	Fa. ↑	Re. ↑	Pr. ↑	PQ ↑	IB-AV ↑
LTX-2.5, no LoRA (no-op ref.)	3	55	0.007	0.997	2.67	4.72	4.96	2.72	6.64	0.373
Kiwi cascade	**50**	44	**0.994**	0.957	2.06	3.43	**4.62**	2.15	4.16	0.064
InstructAV2AV	29	46	0.739	0.981	2.91	3.73	3.84	3.38	5.87	0.154
Task-Adapted LTX-2.5	34	41	0.927	0.971	2.81	3.18	3.31	3.34	**6.91**	0.227
AV-IRE (ours)	34	**50**	0.891	**0.999**	**3.35**	**3.94**	4.33	**3.47**	6.88	**0.281**
AV-IRE (ours), oracle map	34	49	0.891	1.005	3.34	3.99	4.39	3.46	6.90	0.294

backbone leads on kept and IB-AV by making few edits. Blind joint review distinguishes selective removal from suppression or inaction.

Table 3: Single-speaker public benchmarks, scored as in Table 2; J is from blind review (†: earlier round) and superscripts give the valid VTSS clips. Bold/underline: best/second best among the four editors of each panel; –: not computed.

	Blind	AST-based		SAM Audio Judge				Quality		Sync	Official	
Method	J	rem.	TSSR ↑	Ov. ↑	Fa. ↑	Re. ↑	Pr. ↑	PQ ↑	IB-AV ↑	DeSync ↓	UTMOS ↑	VTSS ↑
(a) JAVEditBench subject removal: 25 clips; J over 10, rem. over the 22 with an audible target												
LTX-2.5, no LoRA	–	–	–	–	–	–	–	–	–	–	–	–
Kiwi cascade	$0^{\dagger}$	21/22	0.909	2.01	3.32	4.06	2.22	3.84	0.007	1.31	1.69	0.0310^{23}
InstructAV2AV	$3^{\dagger}$	9/22	0.304	2.48	**4.06**	**4.55**	2.46	5.78	0.115	0.86	1.35	0.0372^{24}
Task-Adapted LTX-2.5	**10**	**22/22**	**0.972**	**3.19**	3.25	3.16	**4.60**	**7.28**	0.133	**0.62**	1.30	0.0262^{18}
AV-IRE (ours)	9	17/22	0.962	3.04	3.70	3.74	3.62	7.05	**0.178**	0.94	**2.34**	0.0305^{16}
(b) InsAVE person removal: 49 queries, all rated blind												
LTX-2.5, no LoRA	–	0/49	0.018	2.31	4.35	4.88	2.18	6.26	0.382	0.05	–	–
Kiwi cascade	1	40/49	0.882	2.86	3.60	4.35	3.16	4.21	0.020	0.75	–	–
InstructAV2AV	29	29/49	0.857	2.96	3.68	4.31	3.20	6.05	0.111	0.78	–	–
Task-Adapted LTX-2.5	**47**	**48/49**	**0.983**	**4.05**	4.06	4.34	**4.73**	**7.41**	0.156	0.93	–	–
AV-IRE (ours)	31	28/49	0.959	3.58	**4.24**	**4.46**	3.75	6.83	**0.169**	**0.63**	–	–

Public single-speaker benchmarks. Table 3 tests single-speaker transfer (protocols and criterion-level results in Appendix I.1). On ten blind-rated JAVEditBench clips (Chen et al., 2026), AV-IRE achieves 90.0% joint success versus 100.0% for Task-Adapted LTX-2.5 and at most 30.0% for earlier-rated baselines. A separate preregistered review of 49 InsAVE queries (Zheng et al., 2026) gives AV-IRE 63.3%, Task-Adapted LTX-2.5 95.9%, and InstructAV2AV 59.2%; Task-Adapted LTX-2.5's advantage is significant ($p = 1.5 \times 10^{-4}$). AV-IRE passes visual removal and both preservation criteria on all 49 queries, with residual speech accounting for its 18 failures. It leads the four editors on IB-AV on both benchmarks (0.178 and 0.169), extending its Natural-2 advantage. This pattern is consistent with routing and reference-conditioned residual edits maintaining source context within the joint generator: transfer yields competitive observed joint success relative to InstructAV2AV and strong semantic coherence, not uniformly stronger target suppression.

Qualitative comparison. In Figure 4, AV-IRE passes all five criteria on both queries; InstructAV2AV leaves both targets visible, while Task-Adapted LTX-2.5 retains the accordionist and damages the other performer and sound in the cello example.



Figure 4: Qualitative comparison on two illustrative Natural-2 queries. ✓/×: pass/fail per blind criterion; teal border: full pass; †: earlier review round.

4.3 Ablation study

Table 4: Ablation on Natural-2; rows are cumulative. N.P.: target-neutral prompt; Map: routing map for gating and pooling; Cr.: crossed pairs. Bold/underline: best/second best.

	Setting			Joint success			Criteria passed (/60)					Wrong source ↓	
Configuration	N.P.	Map	Cr.	*J*/60	*P*/30	gr./10	VR	AR	VP	AP	C	video	audio
Task-Adapted LTX-2.5	–	–	–	7	0	4.53	36	<u>36</u>	31	13	20	<u>7</u>	19
+ executor, fixed target	–	–	–	8	<u>2</u>	4.80	<u>43</u>	34	25	26	16	8	18
+ complementary targets	–	–	–	11	1	5.03	35	35	33	24	24	8	18
+ route-intervention loss	–	–	–	<u>16</u>	<u>2</u>	<u>5.63</u>	38	26	<u>43</u>	<u>33</u>	<u>29</u>	<u>7</u>	**14**
+ routing (**AV-IRE**, ours)	✓	✓	✓	**25**	**5**	**6.83**	**53**	**37**	**47**	**36**	**32**	**2**	**14**

Table 4 compares cumulative configurations under matched executor training, with one run each (configuration details in Appendix Table 10). Rows two to four retain target-bearing prompts, uniform pooling without spatial gating, and, where enabled, diagonal interventions. Full routing jointly adds target-neutral prompting, map-based gating and code construction, and crossed interventions. Complementary targets improve visual preservation and consistency; route-intervention training further improves preservation but lowers acoustic removal from 58.3% to 43.3%. Full routing raises acoustic removal to 61.7% and visual removal from 63.3% to 88.3%, with fewer wrong-performer visual edits and the highest joint success and graded score. These gains are consistent with complementary roles: paired targets specify which source to remove and preserve, shared-state interventions supervise route-dependent changes, spatial maps and codes identify the target, and crossed assignments supervise modality-specific responses. The final comparison measures the routing choices jointly, not individually; bidirectional success is non-monotonic, and adjacent graded-score comparisons are not significant after Holm correction.

4.4 AV-SwapTarget composites and more sources

Table 5: AV-SwapTarget subsets beyond Natural-2: classifier criteria (rem./kept) and, on Composite-N, PROVE temporal coherence and retained-slot LPIPS; other metrics are in Appendix Table 9. Bold/underline: best/second best.

	Composite-2		Composite-N				Natural-3	
Method	rem.	kept	rem.	kept	RC-T ↓	LPIPS ↓	rem.	kept
Kiwi cascade	**17/20**	8/20	**65/70**	141/178	0.317	0.157	**66/81**	117/158
InstructAV2AV	<u>13/20</u>	<u>11/20</u>	25/70	<u>161/178</u>	0.104	<u>0.114</u>	31/81	**129/158**
Task-Adapted LTX-2.5	10/20	9/20	36/70	140/178	<u>0.051</u>	0.124	<u>49/81</u>	117/158
AV-IRE (ours)	12/20	**17/20**	<u>41/70</u>	**167/178**	**0.031**	**0.099**	38/81	<u>125/158</u>

On Composite-N (Table 5), AV-IRE leads on automatic preservation (167/178 retained sources), RC-T, and LPIPS; the Kiwi cascade leads on removal but has the worst temporal coherence. On Composite-2, AV-IRE preserves 17/20 sources versus 9/20 for Task-Adapted LTX-2.5. On Natural-3, the two-source selector pairs each target phrase with a fixed phrase naming no source in the clip. AV-IRE preserves more sources than Task-Adapted LTX-2.5 (125/158 versus 117/158) but passes fewer removal queries (38/81 versus 49/81); InstructAV2AV leads on preservation (129/158).

5 Discussion and limitations

AV-IRE combines complementary-target supervision and explicit routing to improve instruction-dependent joint removal on natural two-musician recordings, with better target-performer removal and preservation of other people and their sounds. Second-reviewer agreement and fresh-set bidirectional gains support this advantage, emphasizing target switching and preservation rather than suppression alone. Residual speech in single-speaker transfer motivates source-specific audio supervision for stronger suppression without sacrificing non-target sounds. The preregistered audio-only intervention in Appendix E.3 detected no contribution of the pooled visual feature to audio success, testing one audio-code input rather than all visual pathways through the joint generator. Future interventions could clarify how untested visual pathways influence acoustic removal and preservation.

ETHICS STATEMENT

Removing a person and their voice from a recording can misrepresent an event or erase evidence of someone's participation, so edited outputs should be labeled as edited and not presented as authentic recordings. Our recordings come from public research datasets used under their published terms, namely AVSpeech and VGGSound under CC BY 4.0, MUSIC-AVQA under CC BY-NC 4.0 and OpenHumanVid under its access agreement, and from the JAVEditBench and InsAVE benchmarks. We redistribute no source media except the short MUSIC-AVQA clips and edited derivatives shown in the anonymous qualitative demo under CC BY-NC 4.0; all other recordings are referenced only by their public identifiers, and we add no identity annotations. The human study consists of blind judgments of model outputs by independent raters who are not authors of this paper and were never involved in the project, and no personal data about the raters were collected. For research integrity, the primary endpoints were pre-registered, and the deviations, the retired predecessor system and the known limitations of the ratings are reported in Appendices E, G, and H.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and objectives, and Section 4.1 describes the training data, evaluation subsets, baseline settings, and blind-review protocol. Appendices C, B, F, and H document training schedules, hyperparameters, checkpoints, hardware, source exclusions, audibility audits, and time-stamped analysis records. We will release the code, the task LoRA and the selector and executor weights, the AV-SwapTarget query manifests, the blind-review answers with their code mappings, the scoring scripts and a per-item table of every automatic score, for research use under the backbone's license and the terms of the source datasets.

AI USE STATEMENT

In this work, we used generative AI tools to help implement methods, to give feedback on the experimental design, and for translation and language editing. We have not used them to formulate claims or proofs or to produce any experimental measurement or human judgment. We reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

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 (CVPR)*, pp. 18392–18402, 2023.

Honghao Cai, Xiangyuan Wang, Yunhao Bai, Haohua Chen, Tianze Zhou, Runqi Wang, Wei Zhu, Yibo Chen, Xu Tang, Yao Hu, and Zhen Li. Edit Where You Mean: Region-Aware Adapter Injection for Mask-Free Local Image Editing. *arXiv preprint arXiv:2604.23763*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 138846–138923, 2026.

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. doi: 10.1109/ICASSP40776.2020.9053174.

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*, 2026.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. *ACM Transactions on Graphics*, 37 (4):112:1–112:11, 2018. doi: 10.1145/3197517.3201357.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. LLMDet: Learning Strong Open-Vocabulary Object Detectors under the Supervision of Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14987–14997, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Xinyue Guo, Jianxuan Yang, Daiguo Zhou, Jiagao Hu, Yuxuan Chen, Fei Wang, and Jian Luan. TV-AudioRemover: Joint Text-Visual Guided Sound Removal with Multi-Task Hard-Mixture Curriculum. *arXiv preprint arXiv:2609.25864*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233*, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. In *International Conference on Learning Representations (ICLR)*, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2022.

Jiagao Hu, Yuxuan Chen, Fuhao Li, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. From Ideal to Real: Stable Video Object Removal under Imperfect Conditions. *arXiv preprint arXiv:2603.09283*, 2026.

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 (ICASSP)*, pp. 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

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.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations (ICLR)*, 2020.

Fuhao Li, Shaofeng You, Jiagao Hu, Yu Liu, Yuxuan Chen, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. PROVE: A Perceptual RemOVal cohErence Benchmark for Visual Media. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19108–19118, 2022.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762, 2025.

Lightricks. LTX-2.5. Hugging Face model repository, Lightricks/LTX-2.5, 2026. URL https://huggingface.co/Lightricks/LTX-2.5.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-Edit: Versatile Video Editing via Instruction and Reference Guidance. *arXiv preprint arXiv:2603.02175*, 2026.

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 (ICLR)*, 2023.

Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. Music Source Separation with Band-Split RoPE Transformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 481–485, 2024. doi: 10.1109/ICASSP48485.2024.10446843.

Yi Luo and Jianwei Yu. Music Source Separation with Band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023. doi: 10.1109/TASLP.2023.3271145.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 149140–149162, 2025. doi: 10.52202/085713-4988.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. SAM Audio: Segment Anything in Audio. *arXiv preprint arXiv:2512.18099*, 2025.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound. *arXiv preprint arXiv:2502.05139*, 2025.

Yuqing Wen, Yukai Huang, Qianqian Xie, Jiangtao Wu, Yibin Lin, Yikai Gu, Jialu Chen, Yuanxing Zhang, and Jiaheng Liu. AVE-Compass: Towards Holistic Evaluation for Audio-Video Editing Abilities. *arXiv preprint arXiv:2607.24821*, 2026.

Weihan Xu, Kan Jen Cheng, Koichi Saito, Muhammad Jehanzeb Mirza, Tingle Li, Yisi Liu, Alexander H. Liu, Liming Wang, Masato Ishii, Takashi Shibuya, Yuki Mitsufuji, Gopala Anumanchipalli, and Paul Pu Liang. Schrodinger Audio-Visual Editor: Object-Level Audiovisual Removal. *arXiv preprint arXiv:2512.12875*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The Sound of Pixels. In *European Conference on Computer Vision (ECCV)*, pp. 570–586, 2018.

Haojie Zheng, Yixin Yang, Siqi Yang, Shuchen Weng, and Boxin Shi. InstructAV2AV: Instruction-Guided Audio-Video Joint Editing. *arXiv preprint arXiv:2605.18467*, 2026.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 75518–75547, 2025. doi: 10.52202/085713-2536.

Appendix

Contents

A Teacher data and provenance **15**

B Training compositions and source exclusions **16**

C Implementation and checkpoint lineage **18**

D Method details **19**

E Evaluation and statistical reporting **21**

E.1 Blind review of the final system 21

E.2 Preregistered follow-ups 23

E.3 Routing-map diagnostics 23

E.4 Earlier review round 24

E.5 Automatic metrics 25

F Audibility audits and eligibility rulings **26**

G Predecessor system and historical evaluation **26**

G.1 Results in the earlier review round 27

H Time-stamped analysis records **27**

I Baseline interfaces and public subsets **29**

I.1 Public single-target transfer 30

J Qualitative selection **31**

A TEACHER DATA AND PROVENANCE

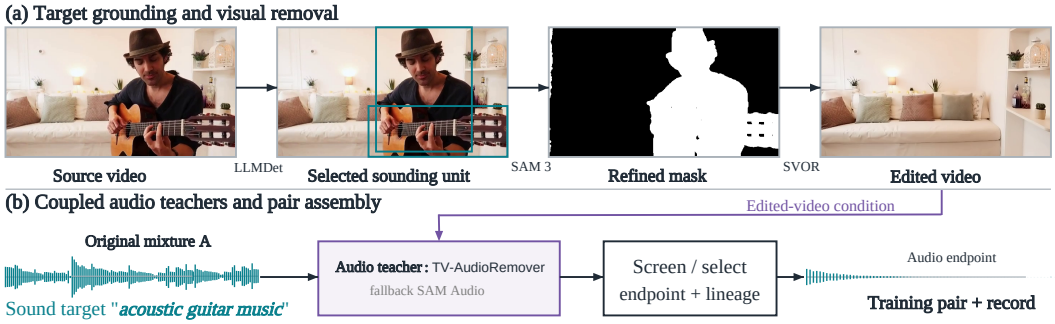

Figure 5: Teacher pipeline for removal pairs, shown on one training record. (a) Target construction: detection and association boxes, the refined mask, and the SVOR video removal. (b) Audio: TV-AudioRemover edits the mixture given the edited video and the sound phrase; a SAM Audio branch replaces rejected audio outputs while keeping the video output. The record is illustrative, not a quality estimate.

Table 6: Evaluation subsets; the first four form AV-SwapTarget. Counts are manifest entries, not independent replicates.

Subset	Role	Inputs	Queries	Evaluation
AV-SwapTarget: two-source instruction selection				
Natural-2	Natural recordings	30	60	blind
Composite-2	Controlled composites	10	20	automatic; slot-masked video
AV-SwapTarget: more sources				
Natural-3	Natural, three sources	27	81	automatic only
Composite-N	Composite, two to five	20	70	automatic; slot-masked video
Public single-target transfer				
JAVEditBench subject removal	Speech	25	25	blind on 10; automatic
InsAVE person removal	Speech	49	49	blind; automatic

Evaluation subsets. Table 6 lists the evaluation subsets. No subset has paired audio-visual removal targets, and InsAVE's video-only references are not scored. Composites have slot regions and exact donor labels. InstructAV2AV's Natural-2 ratings come from an earlier review round.

Teacher pipeline. The offline pipeline localizes the target with LLMDet (Fu et al., 2025), tracks it with SAM 3 (Carion et al., 2026), and removes it from the video with SVOR (Hu et al., 2026). TV-AudioRemover (Guo et al., 2026) edits the audio conditioned on the edited video; text-conditioned SAM Audio (Shi et al., 2025) replaces outputs that fail admission (16,341 of 54,705 pairs). Source identifiers and audio-teacher provenance are retained for tracing the paired supervision.

Training manifest. Figure 5 outlines the teacher pipeline. The frozen task-adaptation manifest contains 54,705 unique record identifiers and 135 target phrases. The production inventory of 76,593 identifiers and the candidate pool of 56,157 records are not training denominators, and the manifest shares no identifier with the pool's 3,795 held-out identifiers. Frame buckets are the largest member of $\{97, 105, 113, 121\}$ not exceeding the recorded frame count.

Composition. Tables 7 and 8 break the corpus down by source and audio teacher and by target phrase. Upstream keys collapse repeated MUSIC-AVQA windows by recording identifier and

Table 7: Task-adaptation corpus by source and audio teacher (unique training records; upstream keys are dataset-scoped identifiers). SAM: audio replaced by SAM Audio, video from SVOR.

Source	Records	Upstream keys	Share (%)	TV-Audio Remover	SAM
OpenHumanVid	28,227	28,227	51.60	26,827	1,400
AVSpeech	20,988	18,505	38.37	6,162	14,826
MUSIC-AVQA	4,614	2,477	8.43	4,614	0
VGGSound	876	876	1.60	761	115
Total	54,705	50,085	100.00	38,364	16,341

Table 8: Most frequent target phrases of the training corpus; the remaining 125 phrases are pooled. Phrases keep their original wording.

Target phrase	Records	Share (%)
a person speaking	49,215	89.96
accordion music	657	1.20
acoustic guitar music	548	1.00
saxophone music	412	0.75
flute music	374	0.68
cello music	348	0.64
piano music	347	0.63
clarinet music	265	0.48
ukulele music	246	0.45
guzheng music	227	0.41
Other 125 phrases	2,066	3.78
Total	54,705	100.00

AVSpeech/VGGSound clips by their stored upstream identifiers; they are not a cross-dataset content-deduplication or person-identity audit. A complete scan finds the same LLMDet checkpoint identifier and SAM 3 concept-head designation throughout.

Admission screens. Semantic admission is source-dependent: some records use a judge, some source metadata, and some a recorded bypass. The target precheck read source metadata directly for 28,227 records, was skipped for 21,864, and is unrecorded for 4,614. Candidate admission requires a stored target-sound suppression ratio $(p_{\text{in}} - p_{\text{out}})/p_{\text{in}}$ of at least 0.7, the minimum over mapped AST target classes with input probability above 0.001. The SAM Audio replacement branch additionally requires $\text{RMS}(A_{\text{out}})/\,\text{RMS}(A_{\text{in}}) < 0.3$, an amplitude ratio that can also reflect over-removal. The MUSIC-AVQA replacement branch is excluded following identified audio–visual inconsistency risks.

No human audit of teacher outputs. The screens target suppression, not preservation of other sources. No systematic human audit of the teacher outputs was run, so their quality is known only through the automatic screens. Recorded box verification, mask admission and TV-AudioRemover duration checks pass for all joined records, but there is no record that the optional compositing, semantic-judge and background-QC stages ran on every record.

B Training compositions and source exclusions

Composites. The original 600 composites pair speech or instrument donors with an instrument donor, and teacher outputs of the donors form both requested edits through Equation (2). Layout, audio gains and any peak-control factor are common to the input and both target outputs. Donor categories come from metadata and are not isolated-source ground truth.

Training inventory. The original 600 composites yield 1,200 query-target examples and cover 13 instruments and voice; voice occupies 300 of their 1,200 source slots. The executor inventory totals 1,836 examples: these 1,200 examples, 440 from the 220 additional pairs, and 196 from natural clips.

Continuation data. Executor continuation adds 220 pairs: 80 piano, 60 pipa, 30 drum, 25 erhu, and 25 suona. Each has one full teacher output and one exact audio remix whose video is copied with a zero loss mask, giving 220 video-masked entries in the 1,836-entry inventory. The 196 natural examples supply one teacher direction each. The selector's 1,739 clips from 1,063 source videos use associated person and instrument boxes, excluding ambiguous overlap.

Source exclusions. The clean subsets exclude every base video used in task-LoRA training, the final candidate pool, selector training or validation, the original composite donors and earlier development pools; Composite-N donors are drawn from the confirmed Composite-2 donors. The 30 Natural-2, 27 Natural-3 and 20 Composite-2 donor recordings share no recording with the 600 composites (208 donor recordings), the 8 held-out composites (12 recordings) or the 196 natural clips (162 recordings).

Scope. Identifier checks are not near-duplicate audits; the public benchmarks were not checked, and the backbone's pretraining data are unknown. Earlier development evaluations are kept separate; some of their windows and donors overlapped task-adaptation data, and early mono-to-stereo handling attenuated some audio.

One ill-posed query. Listening to one Composite-2 input admitted as a flute–ukulele pair found no audible ukulele, as its free-text note indicated. A ruling recorded at 2026-09-21T18:00:56Z, before any result of the same review round existed, excludes the ukulele-removal query from the blind endpoints for every system, leaving 9 inputs for P and 19 queries for J. Table 9 scores all 20 queries; the excluded one is the only query with invalid TSSR, so no removal count changes.

Automatic subset scores. In Table 9, rem. and kept follow Table 2 with exact donor labels, PROVE (Li et al., 2026) masks the target slot, PSNR, SSIM and LPIPS use the retained slots, and SAJ is the judge's Overall field.

Table 9: AV-SwapTarget composite and three-source results, scored as in Table 2. Audio and visual metrics are separated for readability. Bold/underline indicate best/second best among the four editors in each subset; the unadapted backbone is a no-op reference. –: not computed. Visual metrics are unavailable for Natural-3.

Composite-2: 10 inputs, 20 queries						
Method	rem.	kept	TSSR ↑	SAJ ↑	PQ ↑	IB-AV ↑
LTX-2.5, no LoRA	0/20	17/20	0.048	2.61	6.30	0.268
Kiwi cascade	**17/20**	8/20	**0.999**	1.80	3.97	0.052
InstructAV2AV	<u>13/20</u>	<u>11/20</u>	0.876	<u>2.93</u>	5.67	0.165
Task-Adapted LTX-2.5	10/20	9/20	0.829	2.71	**7.51**	<u>0.204</u>
AV-IRE (ours)	12/20	**17/20**	<u>0.897</u>	**3.62**	<u>6.30</u>	**0.205**
Method	RC-S ↑	RC-T ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
LTX-2.5, no LoRA	0.165	0.028	27.96	0.903	0.041	
Kiwi cascade	**0.176**	0.239	22.04	0.717	0.166	
InstructAV2AV	0.154	0.056	<u>24.70</u>	<u>0.831</u>	<u>0.159</u>	
Task-Adapted LTX-2.5	<u>0.159</u>	<u>0.024</u>	20.50	0.758	0.200	
AV-IRE (ours)	0.158	**0.020**	**25.43**	**0.866**	**0.091**	

Continued on next page

Table 9 continued

Composite-N: 20 inputs, 70 queries; 178 retained sources

Method	rem.	kept	TSSR ↑	SAJ ↑	PQ ↑	IB-AV ↑
LTX-2.5, no LoRA	–	–	–	–	–	–
Kiwi cascade	**65/70**	141/178	**0.995**	2.01	3.95	0.051
InstructAV2AV	25/70	161/178	0.439	2.42	5.27	0.160
Task-Adapted LTX-2.5	36/70	140/178	0.818	**3.38**	**7.23**	**0.217**
AV-IRE (ours)	41/70	**167/178**	0.908	2.95	6.39	0.185

Method	RC-S ↑	RC-T ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LTX-2.5, no LoRA	–	–	–	–	–
Kiwi cascade	0.151	0.317	18.91	0.686	0.157
InstructAV2AV	0.144	0.104	**23.32**	**0.840**	0.114
Task-Adapted LTX-2.5	**0.159**	0.051	19.95	0.774	0.124
AV-IRE (ours)	0.157	**0.031**	22.97	0.833	**0.099**

Natural-3: 27 inputs, 81 queries; 158 retained sources

Method	rem.	kept	TSSR ↑	SAJ ↑	PQ ↑	IB-AV ↑
LTX-2.5, no LoRA	–	–	–	–	–	–
Kiwi cascade	**66/81**	117/158	**0.971**	2.29	4.40	0.081
InstructAV2AV	31/81	**129/158**	0.474	2.86	5.63	0.165
Task-Adapted LTX-2.5	49/81	117/158	0.849	2.80	**7.06**	0.198
AV-IRE (ours)	38/81	125/158	0.671	**3.18**	6.85	**0.308**

C Implementation and checkpoint lineage

Hardware. Task-LoRA and executor runs each used one NVIDIA H100 80GB HBM3, and the offline data pipeline 24 NVIDIA H20 143 GB devices. The task LoRA was trained from the base model in one continuous 6,000-step run, rank and alpha 64, checkpointing every 500 steps, with peak memory 42.79 GB. The retrained replicate R recorded in the development analysis differs from the final training configuration by execution nondeterminism alone: no deterministic-algorithm switch was set, and training used bf16 mixed precision with int8 weight quantization.

Selector. The selector uses 256-dimensional projections, a 512-wide hidden layer, $m = 2$, $\lambda_{\text{swap}} = 1$ and $\lambda_{\text{bg}} = 0.1$, and is trained for 1,500 updates with batch size 32, learning rate 3×10^{-4}, weight decay 0.01 and seed 31.

Executor schedule. The executor trains for 1,000 updates on the original complementary composites, followed by 1,000 updates on the full 1,836-entry inventory. In both stages, an original-composite microbatch uses route intervention with probability 0.5 and flow matching otherwise; all other microbatches use flow matching. The two objectives are not summed within a microbatch, and gradients accumulate over four microbatches. The stage-two plan specifies microbatch size one, learning rate 3×10^{-4}, cosine scheduling, and sampling weights 0.5, 0.3 and 0.2 for original composites, added pairs and natural examples. Route steps use $\lambda_{\text{route}} = 1$, $\lambda_{\text{mid}} = 0.1$, $w_v = 0.5$, $w_a = 3$ and $\sigma \in \{0.5, 0.7, 0.9\}$.

Map curriculum. The conditioning map is the oracle for the first 40% of each stage's updates, predicted out-of-fold or oracle with probability 0.5 per microbatch up to 70%, and predicted thereafter. Entries without an out-of-fold prediction fall back to the oracle, counted in the run logs. This curriculum introduces predicted localization errors during training.

Ablation ladder. The four executor configurations in Table 4 use the same two-stage schedule and trainable parameterization: 52,998,400 parameters in 120 tensors. They share the frozen backbone, supervision data, seed, selector routing-map file and task LoRA. Each configuration is a single training run. Task-Adapted LTX-2.5 uses no executor.

Configuration settings. Rows two to four put the phrase in the prompt with Task-Adapted LTX-2.5's template, feed the audio code's phrase projection to the video code as well, and use no map, loading the routing-map file without using it: a code switch removes the video gate and pooling is uniform. The fixed-target configuration fixes one direction per paired input, exactly half of each pair type each way, with the same assignment in both stages; the intervention-loss configuration uses diagonal pairs only. Table 10 lists the resulting configuration of each row.

Table 10: Configuration of each row of Table 4.

Configuration	Exec.	Prompt	Video code	Spatial map	Targets	Route loss
Task-Adapted LTX-2.5	–	phrase	–	–	–	–
+ executor, fixed target	✓	phrase	phrase	–	one per pair	–
+ complementary targets	✓	phrase	phrase	–	complementary	–
+ route-intervention loss	✓	phrase	phrase	–	complementary	diagonal
+ full routing (AV-IRE, ours)	✓	neutral	map pooling	✓	complementary	diagonal, crossed

Variability. Adjacent graded-score steps have raw $p = 0.449$, 0.756, 0.129 and 0.0185, the last Holm-adjusted to 0.0739. A development rerun with identical nominal settings changed a count of bidirectional successes by 4 of 60, showing execution variability, not an uncertainty estimate for the reported results.

Checkpoints. The stage-two executor continues the stage-one executor. AV-IRE+head used the stage-two executor with predicted routing for video and the audio head for audio; its oracle-map variant differed only in the routing map. The job log lists 43 training jobs from 2026-09-15T23:10:18Z onward, after task-LoRA training, 7 of them incomplete.

Final system. AV-IRE contains no separate audio head. It loads the frozen LTX-2.5 backbone with its text encoder and both VAEs, the frozen step-6,000 task LoRA, and the final stage-two executor. The executor comprises six route blocks at transformer depths 0, 8, 16, 24, 32 and 40 with 48,838,656 parameters, plus route-code projections with 4,159,744 parameters, totaling 52,998,400 parameters or 53.0M. The frozen selector is loaded separately. The oracle-map variant changes only the precomputed routing maps; checkpoint identifiers remain in the release manifest.

Inference. Inference uses two feature-extraction steps followed by 30 joint sampling steps, with default outputs of 121 frames at 512×320 and 30 frames per second. The route is fixed throughout generation, which requires no teacher outputs, user masks, oracle maps, or retained-source descriptions. Reference audio hidden states have 101 time positions, and the feature-extraction run exposes the state at noise level 1 read by the selector, so inference cost includes this run. Generation takes about 71 s per query on one H100, measured as job wall-clock time including model loading and feature extraction (80 queries in 95 minutes).

D METHOD DETAILS

Reference features. Audio features $a = \{a_i\}_{i=1}^{T_a}$ come from block 32 of the feature-extraction run, as the video features do, with $\bar{a} = T_a^{-1} \sum_i a_i$; these pooled features differ from the depth-aligned reference-token states ρ_v, ρ_a read by the route blocks. The target phrase is pooled before the text connector to obtain t.

Spatial selector. In Equation (3), $[\cdot;\cdot]$ denotes concatenation, P_v, P_t are layer-normalization–projection maps and S is an MLP. The selector is trained on cached features of natural two-person instrument clips with box-derived source regions and then frozen. For instruction-side logits s_n^A, s_n^B,

let $\Delta_n = s_n^A - s_n^B$. With source regions Ω_A, Ω_B and background Ω_0,

$$\mathcal{L}_{\text{sel}} = \mathcal{L}_{\text{bce}} + \lambda_{\text{swap}} \left[\operatorname*{mean}_{n \in \Omega_A} \operatorname{softplus}(m - \Delta_n) + \operatorname*{mean}_{n \in \Omega_B} \operatorname{softplus}(m + \Delta_n) \right] + \lambda_{\text{bg}} \operatorname*{mean}_{n \in \Omega_0} |\Delta_n|. \tag{10}$$

A position is labelled A if box A covers at least half of it and box B none, B symmetrically, Ω_0 if neither box covers it, and ignored otherwise; pairs with an empty A or B region are dropped. The region-balanced cross-entropy $\mathcal{L}_{\text{bce}}$ averages within regions: for query A, Ω_A is positive and Ω_B, Ω_0 negative, with background weight $w_{\text{bg}} = 0.5$; query B is symmetric, and the two are averaged.

Route interface. The route is the tuple $r = (g_{\text{gate}}, c_v, c_a)$, with two 256-dimensional codes. We write g_c for the pooling map in Equation (4) and g_{gate} for the execution map; at inference $g_{\text{gate}} = g_c = g$. Codes are broadcast over tokens and the execution map over latent frames and channels. The spatial map and both codes remain fixed throughout generation. The output projections of E_v, E_a, R_v, R_a are zero-initialized.

Flow matching toward the target. For an input–target pair (X, Y), let $z = (z_v, z_a)$ encode the input and $y = (y_v, y_a)$ the edited target. With $\varepsilon_m \sim \mathcal{N}(0, I)$ and noise level σ_m,

$$x_{\sigma_m, m} = (1 - \sigma_m) y_m + \sigma_m \varepsilon_m, \qquad v_m^\star = \varepsilon_m - y_m, \quad m \in \{v, a\}, \tag{11}$$

so zero is clean data and one is noise. With $v_{\theta,m}$ the velocity prediction for modality m, conditioned on both noisy modalities, the reference latents, the constant prompt b and the route r,

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\sum_{m \in \{v,a\}} \alpha_m \left\langle \left(v_{\theta,m}(x_\sigma, \boldsymbol{\sigma}; z, b, r) - v_m^\star \right)^2 \right\rangle_{\mathcal{M}_m} \right], \tag{12}$$

where $\alpha_v = \alpha_a = 1$ for ordinary executor flow matching, and $\langle \cdot \rangle_{\mathcal{M}_m}$ is the masked mean over supervised tokens and channels, zero for an empty mask. Task adaptation instead uses each pair's instruction and no route. Audio-only remix entries have an empty video mask and so apply no direct video target loss.

Route assignment. For evaluation i of an intervention pair with map-side source $s_g^{(i)}$ and code-side source $s_c^{(i)}$, g_{gate} follows $s_g^{(i)}$, both codes follow $s_c^{(i)}$ with g_c and t taken from it, and the targets are $y_v^{\star(i)} = y_{v, s_g^{(i)}}$ and $y_a^{\star(i)} = y_{a, s_c^{(i)}}$. The two evaluations share the reference, prompt, noise level and per-modality noise draws. Table 11 lists the four assignments.

Table 11: Route assignments in the paired evaluations. Codes $c_S = (c_{v,S}, c_{a,S})$ are computed from source S's phrase and pooling map.

Evaluation	Execution map	Codes	Video target	Audio target
Diagonal A	g_A	c_A	remove A	remove A
Diagonal B	g_B	c_B	remove B	remove B
Crossed A/B	g_A	c_B	remove A	remove B
Crossed B/A	g_B	c_A	remove B	remove A

Normalizers. In Equation (9), $\langle \cdot \rangle_m$ averages over the union of the two annotated source regions in video and over all output tokens and channels in audio, and

$$D_m = \operatorname{sg}\big(\langle (d_m^\star)^2 \rangle_m\big) + \epsilon, \qquad B_m = \max\big(\operatorname{sg}(\langle (\mu_m^\star - z_m)^2 \rangle_m),\, 10^{-4} \operatorname{sg}(\langle (d_m^\star)^2 \rangle_m)\big) + \epsilon,$$

with stop-gradient sg and $\epsilon = 10^{-12}$; pairs with zero target-difference energy are rejected. The weights w_m are distinct from α_m.

Differential and common modes. For one modality, write $e^{(i)} = \hat{y}^{(i)} - y^{\star(i)}$, $e_- = (e^{(1)} - e^{(2)})/\sqrt{2}$ and $e_+ = (e^{(1)} + e^{(2)})/\sqrt{2}$. Since $\hat{d} - d^\star = e^{(1)} - e^{(2)}$ and $\hat{\mu} - \mu^\star = \frac{1}{2}(e^{(1)} + e^{(2)})$, the bracket of Equation (9) equals

$$\frac{2\langle e_-^2 \rangle}{D} + \frac{\lambda_{\text{mid}} \langle e_+^2 \rangle}{2B},$$

so the two modes are weighted in the ratio $4B/(\lambda_{\text{mid}} D)$, which depends on the per-sample normalizers as well as on λ_{mid}. Unless the two mode weights are equal, a cross term $\langle e^{(1)} e^{(2)} \rangle$ remains in the two evaluations' errors, so the loss is not a per-evaluation reweighting of target errors.

Relation to flow matching. Asking the one-step estimate $\hat{y}^{(i)}$ from the input-built state ξ_σ to equal y corresponds to the velocity $(\xi_\sigma - y)/\sigma = \varepsilon - y + \frac{1-\sigma}{\sigma}(z - y)$, not to the flow-matching target $\varepsilon - y$ of Equation (11).

Training stages. Table 12 lists the training stages; Appendix C gives their schedules and hyperparameters.

Table 12: Training stages. FM: flow matching toward the target by Equation (12); route: route-intervention step by Equation (9).

Stage	Trained	Frozen	Data	Objective and length
Task adaptation	task LoRA (rank 64)	backbone	54,705 teacher pairs	FM, 6,000 steps
Selector	selector projections and MLP	backbone and task LoRA (cached features)	1,739 natural two-person clips with boxes	Equation (10), 1,500 updates
Executor stage 1	code networks and route blocks	backbone, LoRA, selector	600 composites	FM or route, each with probability 0.5; 1,000 updates
Executor stage 2	code networks and route blocks	backbone, LoRA, selector	1,836 entries, sampled 0.5/0.3/0.2	FM; route with probability 0.5 on original composites only; 1,000 updates

E Evaluation and statistical reporting

E.1 Blind review of the final system

Numerical reporting. Rates and percentage-point differences are computed before rounding; Table 1 retains the underlying counts. Reported J counts queries passing all five criteria; P counts recordings on which both requests pass. The graded score is the mean number of criteria passed per recording across its two requests, out of 10. For the AV-IRE–Task-Adapted LTX-2.5 McNemar comparisons, b counts units passed by AV-IRE alone and c those passed by Task-Adapted LTX-2.5 alone; the unit is a recording for P and a single-target query for InsAVE J.

Blind judgments. Every rater is independent: not an author and never involved in the project. Raters in the original review rounds had not previously seen the models or their outputs or worked on this task. The second-reviewer coverage follow-up extends the same reviewer's initial overlap ratings. The review rounds and rater assignments are documented in the analysis records. The reviewer sees the exact input window and requested target but not the system, and the five criteria of Section 3.1 are retained separately. Partial or unsure judgments are counted but are not passes. An unchanged output fails removal, a silent output fails preservation when another source is present, and editing only one modality fails joint success.

Review rounds. The final system's review round holds 396 items, each one configuration's output for one query: 360 outputs from six configurations on 60 Natural-2 queries, 20 appended JAVEd-

itBench outputs, and 16 repeats, rated in two sessions of 198. A second page carries 120 overlap items: ten Natural-2 inputs in both directions for each of the six configurations.

Package verification. Before rating, an author who had opened neither mapping verified package integrity, checked that the pages and manifests contain no system directory names, seeds or system fields, and confirmed that both mappings remain outside the page directories.

Repeated items. The 16 repeats fall unevenly across configurations by draw. The same joint verdict was reproduced on 14 of 16, and acoustic removal and audio-visual consistency on 12 of 16 each.

Inter-rater agreement. On the 120 overlap queries, verified byte-identical for both reviewers, joint success J agrees on 109, with $\kappa = 0.798$ (both pass 36, first reviewer only 9, second only 2, both fail 73). On the 60 input-configuration units they form, bidirectional success P agrees on 56, with $\kappa = 0.739$, reported descriptively since each reviewer records only 9 such successes. The table gives each reviewer's J out of 20 and P out of 10 per configuration, with rows as in Table 4.

	First reviewer		Second reviewer	
Configuration	J/20	P/10	J/20	P/10
Task-Adapted LTX-2.5	4	0	3	1
+ executor, fixed target	4	1	4	1
+ complementary targets	6	1	5	0
+ route-intervention loss	9	2	6	1
AV-IRE (ours)	10	3	10	3
AV-IRE (ours), oracle map	12	2	10	3

Ordering and criteria. Both reviewers rank AV-IRE above Task-Adapted LTX-2.5 in joint success on the overlap, although some intermediate configurations tie under one reviewer. The second reviewer credits Task-Adapted LTX-2.5 with one bidirectional success on the overlap, so its $0/30$ in Table 1 is the first reviewer's count. Per criterion, κ is 0.911 for visual removal, 0.872 for preservation of the other people, 0.717 for preservation of the other sounds, 0.671 for acoustic removal and 0.629 for audio-visual consistency.

Conservative merge. The pre-declared sensitivity analysis counts either reviewer's failure on the overlap as a failure. AV-IRE retains its advantage over Task-Adapted LTX-2.5 and all 10 joint successes, while Task-Adapted LTX-2.5 falls from 4 to 2 and the oracle-map variant from 12 to 9.

Recorded deviation. On the main set, the non-passing options of three criteria were worded differently from the rating page the registration referred to. The passing option is identical, so J, P, the graded score and every comparison built on them are unaffected, but the finer failure categories cannot be pooled with those of the earlier review round.

Wilson intervals. The Wilson 95% intervals on P are $[0.07, 0.34]$ for AV-IRE's 5 of 30 and $[0, 0.11]$ for Task-Adapted LTX-2.5's 0 of 30. They are marginal; the paired contrast is carried by the exact McNemar test. J is reported as counts only, since the two instructions of an input are not independent.

Post-unblinding corrections. After unblinding, the first rater confirmed data-entry errors on two Task-Adapted LTX-2.5 outputs that removed the wrong performer in both modalities. One originally passed all five criteria; the other already failed acoustic preservation and therefore did not count toward J. The primary analyses retain the frozen answers. Applying both corrections as a sensitivity analysis lowers J from 7 to 6 and the graded score from 4.53 to 4.30, with P unchanged. The supplementary correction ledger preserves the original and corrected responses separately.

E.2 PREREGISTERED FOLLOW-UPS

The time-stamped analysis record (xi) fixed three follow-ups before any of their outputs or ratings existed and required each to be reported whatever its direction. The second-reviewer and fresh-set comparisons use the main analysis's endpoint tests. The audio-only intervention uses the tests specified in Appendix E.3.

Second reviewer on all AV-IRE and Task-Adapted LTX-2.5 queries. The second reviewer of the final round rated the 80 remaining AV-IRE and Task-Adapted LTX-2.5 outputs on a new page with identical media and wording. With the 40 overlap items, this covers all 120 outputs for 60 queries. Eight overlap items were repeated to check drift, and J agreed on all eight. Record (ix)'s endpoints still use the first rater. In the table below, criterion and wrong-source counts are out of 60, and graded scores are out of 10.

Method	VR	AR	VP	AP	C	J/60	P/30	Graded	Wrong V/A
Task-Adapted LTX-2.5	35	40	34	13	19	5	1	4.70	24/30
AV-IRE (ours)	52	40	48	37	33	26	6	7.00	9/17

Second-reviewer tests and agreement. The primary contrast gives $b = 5$, $c = 0$ and $p = 0.0625$, so it does not reach the prespecified threshold; the graded score gives $p = 1.1 \times 10^{-4}$. The two reviewers agree on J for 113 of 120 queries ($\kappa = 0.849$). Per criterion, κ is 0.872 for visual removal, 0.813 for preservation of the other people, 0.673 for preservation of the other sounds, 0.695 for acoustic removal and 0.729 for audio-visual consistency. Counting either reviewer's failure as a failure gives J of 24 versus 4 and P of 4 versus 0 ($p = 0.125$).

Fresh test set: construction. Candidates come from the natural two-person pool that passed person detection and lie outside Natural-2. Any candidate sharing a source video with the task-LoRA training data, selector training, composite donors, any development set or the Natural-3 donors was removed, and executor training had excluded the whole pool. An input was admitted if it shows two performers of different instrument classes, both audible in the input, over a 121-frame window. An author confirmed audibility from the input alone. Candidates were shuffled with a fixed seed and the first 30 admissible inputs were kept. AV-IRE, unchanged and with routes from the same selector, and Task-Adapted LTX-2.5 were run with the protocol of the final round. A new independent rater, different from the rater of the audio-only test, rated the 120 queries and 4 repeats with the wording of the final round; J agreed on all four repeats.

Fresh test set: results. P is 8 versus 1 of 30 ($b = 8$, $c = 1$, $p = 0.039$), J is 34 versus 7 of 60, and the graded score is 8.07 versus 4.23 ($p = 4.6 \times 10^{-6}$). This set does not replace Natural-2, whose primary contrast stays at $p = 0.0625$.

Fresh test set: label errors. The rater left the seven items of five queries unrated, noting that their target phrases do not match the recordings; the authors confirmed these as admission errors, one query on each of five inputs. The mapping was opened once to check completeness, no result was computed, and the rater, who did not see it, then rated those items on the same page. The primary analysis keeps all 60 queries. As a sensitivity analysis, excluding the five queries gives J of 33 versus 5 of 55, and excluding their five inputs gives P of 8 versus 0 of 25 ($p = 0.0078$) and graded scores of 8.24 versus 4.08.

E.3 ROUTING-MAP DIAGNOSTICS

These diagnostics separate spatial localization from the contribution of pooled visual features to the audio code.

Predicted localization. The selector's map peaks on the target side for 55/60 queries and both requests on 25/30 inputs, with 0.56 of its mass inside the target region.

Box-map localization diagnostic. The oracle-map variant uses the same AV-IRE weights with box-derived routing maps in place of predicted maps. In the first reviewer's full Natural-2 evaluation, it achieves $J = 30/60$, $P = 5/30$, and a graded score of 7.67. The counts passing VR, AR, VP, AP, and C are 58, 40, 56, 38, and 38 out of 60; wrong-source removal occurs on 1 video query and 10 audio queries. These are diagnostic results with supplied target locations, distinct from the predicted-map inference of Table 1. The second reviewer assessed this variant on the initial overlap only, rather than the full Natural-2 set; the fresh-set evaluation includes AV-IRE and Task-Adapted LTX-2.5 without this variant.

Effect of supplied locations. With the executor fixed, the oracle-map substitution raises first-reviewer joint success from 25/60 to 30/60 and the graded score from 6.83 to 7.67 (exploratory, $p = 0.015$). Visual preservation improves on 9 queries across 6 inputs (exact sign test over inputs, $p = 0.031$), while bidirectional success remains 5/30. This comparison measures the effect of supplied target locations without retraining the generator.

Audio-only intervention. Three configurations regenerate the 60 Natural-2 queries with the same inference code, job batch and seed. The native configuration is unchanged; the mean configuration replaces p in the audio code of Equation (5) with the spatial mean of f; the swap configuration replaces it with the feature pooled under the other performer's map. The visual code, the spatial routing map and the phrase are unchanged. Run-time assertions confirmed that these are identical across the three configurations and that the audio code differs from the native one in the other two. An independent rater heard only the input and output audio, with the target phrase, for 180 items and 9 repeats in random order. Audio success requires the target sound to be fully absent and the other sounds intact.

Configuration	Audio success/60	Wrong-source audio/60	Inputs with 0/1/2 successes
Native	32	24	5/18/7
Mean feature	32	22	4/20/6
Swapped feature	32	19	4/20/6

Intervention tests. The primary test compares the per-input success counts of the native and swap configurations. They differ on four inputs, two in each direction, so the exact sign test gives $p = 1.0$; the native and mean configurations also differ on four inputs, two in each direction ($p = 1.0$). Wrong-source audio removal occurs in the swap configuration alone on 3 queries and in the native configuration alone on 8 (exact McNemar, $p = 0.23$). All nine repeats reproduced the rater's judgments. These interventions detect no contribution of pooled visual feature p to audio success under the preregistered test. They isolate this input to the audio code; visual conditioning through other pathways in the joint generator remains untested. In each configuration, one output has an RMS level below -60 dBFS, none is digitally silent, and none was excluded.

E.4 Earlier review round

Design and rater. The earlier review round rated four systems, including the retired AV-IRE+head, on all 60 Natural-2 queries, 240 items plus 10 repeats. One independent rater answered all 250 items in one session, and the blind-code mapping was opened only after the answers were complete.

Post-hoc frame check. After all results were known, one author aware of the hypothesis checked, unblinded, two to three frames from each of the 240 main items against the visual ratings. Every item rated a full success for AV-IRE+head (13), its oracle variant (20) and InstructAV2AV (1) was consistent with its ratings. Of Task-Adapted LTX-2.5's 8, 3 show a different performer removed than requested, each one instruction of one of its three bidirectional successes.

Effect of the check. Treating only these three as failures would lower Task-Adapted LTX-2.5's P to 0, which leaves the primary contrast non-significant ($P = 2/30$ against $0/30$, exact McNemar $p = 0.50$), and would change J to $13/60$ against $5/60$. Because the check was post hoc and every

correction it proposes falls on Task-Adapted LTX-2.5, we revise no reported number and report the check as evidence that the rater's wrong-source detection is imperfect.

Unconfirmed observation. The check also noted, without confirmation by a second viewer, that Task-Adapted LTX-2.5 appears to remove the same performer under both instructions on at least 10 of 30 inputs, against 2 for AV-IRE+head; it is flagged for independent review rather than reported as a result.

Task-Adapted LTX-2.5 rated in two review rounds. Task-Adapted LTX-2.5's outputs are identical in this round and in the final system's round, where a different rater, who did not know of this check, rated them. That rater's criterion-level tallies are about two to one stricter, and two of the three inputs that lost bidirectional success fail there in the wrong-performer pattern the check had flagged. The joint-success flips between the rounds are balanced (4 against 3); those three inputs lie outside the overlap and have no second rating in the later round.

E.5 Automatic metrics

Classifier criteria. For AST (Gong et al., 2021) probabilities $P_{\text{in}}, P_{\text{out}}$ and target class r,

$$\text{TSSR} = \frac{P_{\text{in}}(r) - P_{\text{out}}(r)}{P_{\text{in}}(r)}. \tag{13}$$

For retained class k, let $\delta = P_{\text{out}}(k) - P_{\text{in}}(k)$ and $\text{PSF} = 1 + \delta$ if $\delta < 0$, otherwise $1 + \min(\delta, 0.5)$. Removal passes at TSSR at least 0.7 with output target probability below 0.3, preservation at PSF at least 0.85, and each is valid only if its input-class probability exceeds 0.001. We run the reference implementation unmodified except for four path constants.

Validity and medians. TSSR and PSF are each valid on 57 of the 60 Natural-2 queries for every system: in three inputs the classifier finds one source absent, invalidating TSSR on the query naming it and PSF on the complementary query. Those queries are excluded from the medians of Table 2; pass counts treat an invalid TSSR as a removal failure and an invalid PSF as a preservation pass. We report median TSSR because the mean is unstable: Task-Adapted LTX-2.5's mean over its valid Natural-2 queries is -10.0.

Other scores. SAJ is the SAM Audio Judge's Overall score and its Faithfulness, Recall and Precision fields, PQ production quality, IB-AV ImageBind audio-video similarity and DeSync synchronization. The judge receives a description of retained sound for evaluation only; that description does not enter an instruction-only baseline's inference. On composites, PROVE uses the target slot as the removal mask, and slot-masked PSNR, SSIM and LPIPS compare unedited regions against the input.

No-op reference. The unadapted backbone leads the preservation criterion, the judge's faithfulness and recall, and IB-AV by editing almost nothing, so it diagnoses metric interpretation and is not a competitive baseline.

Agreement with blind judgment. In multi-source scenes the classifier criteria agree poorly with blind judgment at the level of a query. On 20 development queries the classifier removal criterion agreed with blind judgment at $\kappa = 0.175$ (95% interval $[-0.25, 0.59]$) and the preservation criterion at $\kappa = -0.071$, and in a 30-query review round the classifier passed 15 queries that no reviewer passed. The single-speaker comparison is reported with the public-benchmark results.

Short clips. Wrappers use two frames for clips shorter than four seconds and repeat aligned audio and video for synchronization below 4.8 seconds, so these are adapted short-clip metrics. DeSync is computed on Natural-2 but not reported, since its 4.03 s clips take the short-clip branch; the released per-item file contains it.

F AUDIBILITY AUDITS AND ELIGIBILITY RULINGS

Procedure. A query whose declared target is inaudible in the input is ill-posed: acoustic removal passes trivially for any editor that adds no sound, and preservation then rewards whichever editor changes least. We re-checked every declared target in the earlier natural-subset review pools and in JAVEditBench, and one composite flagged by its admission note, listening to the input window alone with no system output present, under rules fixed and time-stamped before listening:

- the pre-registered denominators remain primary, and the audited denominators are reported as a sensitivity analysis;
- an ill-posed query is removed from J, and its whole input from P and the graded score, since that input cannot contribute a bidirectional outcome;
- the rule applies identically to every system.

Table 13: Declared targets found inaudible on input-only re-listening, and the effect on the retired AV-IRE+head against Task-Adapted LTX-2.5 in the earlier review round.

Subset	Inaudible	When checked	Effect
Natural-2 (two sources)	2/60 (3.3%)	after unblinding	AV-IRE+head J 13→11, j_2 19→17; Task-Adapted LTX-2.5 J and P unchanged, j_2 42→41
Natural-3, earlier 45-query pool	4/45 (8.9%)	one during rating, three after	excluded from the AV-IRE+head rating
Composite-2	1 flagged input	before any result of the same round	development P over 9 inputs
JAVEditBench	3/25 (12%)	before any output was viewed	rem. over the 22 audible clips

Natural-2. Table 13 summarizes the four audits. The Natural-2 audit was performed after the earlier review round was unblinded; its predicted direction, that it would lower AV-IRE+head more than Task-Adapted LTX-2.5, was recorded before listening and held. The exclusion table was fixed before the final system's review round was built. The earlier Natural-3 audit covered 45 queries, distinct from the 81-query automatic evaluation in Table 6. There, both excluded queries are five-criterion passes for AV-IRE and failures for Task-Adapted LTX-2.5.

Audited contrasts. Over the audited denominators AV-IRE's J falls from 25 of 60 to 23 of 58 and Task-Adapted LTX-2.5's stays at 7 of 58. P becomes 4 of 28 against 0 of 28, not significant as on the full subset (exact McNemar, $p = 0.125$), and the graded score 6.71 against 4.57, still significant (paired Wilcoxon over inputs, $p = 0.0013$). The oracle-map variant reaches 28 of 58 on J and 5 of 28 on P.

Admission notes and window length. In three flagged cases a free-text admission note contradicted the structured admission answer and was correct; in the others the structured answers were wrong, so structured admission alone does not certify audibility. The JAVEditBench check listened to the 4.03-second windows of the earlier review round, while the reported runs use the native 4.835-second windows.

G PREDECESSOR SYSTEM AND HISTORICAL EVALUATION

Predecessor. An earlier version, which we write AV-IRE+head, paired the routed generator's video stream with a separate instruction-conditioned audio head that edited the recorded waveform by conservative subtraction. The project specification requires a single joint generative model, and AV-IRE replaced it after its pre-registered primary contrast on Natural-2 was found null; the time-stamped analysis records document the decision and its timing. Its video stream is identical to AV-IRE's, so the earlier review round rated frames that AV-IRE also produces.

Table 14: Blind five-criterion review on Natural-2 (30 natural two-source inputs, 60 queries, one review round). P is input-level bidirectional success; ≥ 1 counts inputs on which at least one instruction succeeds; J is query-level success; J_v and J_a require removal and preservation in video and audio respectively; graded is the mean number of criteria passed per input (0 to 10). The audited sensitivity columns follow the audibility audit: J is over the 58 individually eligible queries, whereas P and graded are over the 28 inputs whose both instructions are eligible (56 queries); the eligible counterpart of each excluded query remains in J only. The primary contrast is AV-IRE+head against Task-Adapted LTX-2.5 on P ($b = 2$, $c = 3$, $p = 1.00$); the pre-registered secondary endpoint is graded ($p = 0.0017$; audited $p = 0.0049$). The oracle row replaces only the routing map and shares AV-IRE+head's audio bitwise.

	Preregistered						Audited		
Method	P	≥ 1	J	J_v	J_a	graded	P /28	J /58	graded /28
InstructAV2AV	0	1	1	2	16	3.67	0	1	3.64
Task-Adapted LTX-2.5	3	5	8	20	14	5.13	3	8	5.29
AV-IRE+head (ours)	2	11	13	42	17	7.13	2	11	7.04
AV-IRE+head (ours), oracle map	5	15	20	55	22	8.17	5	18	8.14

G.1 Results in the earlier review round

Table 14 reports the preregistered earlier review round, not the final AV-IRE system. The primary contrast detects no significant difference in bidirectional success, $P = 2/30$ for AV-IRE+head against $3/30$ for Task-Adapted LTX-2.5 (exact McNemar, $p = 1.00$), while the graded secondary endpoint is 7.13 versus 5.13 ($p = 0.0017$). The oracle-map variant has identical audio yet receives more acoustic-removal passes, 24 versus 19, showing that these joint audio-video judgments need not depend on the waveform alone. InstructAV2AV's Natural-2 results in Table 1, the frame check, and Task-Adapted LTX-2.5's earlier rating in Appendix E come from this round.

H Time-stamped analysis records

Table 15 lists when each analysis choice was fixed and what information was available then: mappings, aggregate results, per-system results or output media. Prospective registrations are distinguished from interim-informed and post-unblinding records. The original records retain their timestamps and amendment history.

Table 15: Time-stamped analysis records and information available when written. Times are UTC.

Record	Date and time	Status	Recorded choices
(i)	2026-09-21 16:48:49; modified 2026-09-22 01:11	Prospective: before any subset result	Blind-review primary contrast, criteria, failure policy and annotation budget
(ii)	2026-09-21 18:00:56	Before any score from that round	Eligibility of one Composite-2 input
(iii)	2026-09-21 19:04:41	Interim-informed: after analysis of the first ten inputs	Cross-modal feature comparison extended to eight natural inputs, reported separately; the effect did not replicate

Continued on next page

Table 15 continued

Record	Date and time	Status	Recorded choices
(iv)	2026-09-22 03:10:40	Independently blind, not prospective: 71 seconds after another session opened the mapping	Two hypotheses about the blind aggregate and their decision rules
(v)	2026-09-22 03:47:37; 03:52:22	After unblinding, before re-listening	Natural-2 audibility rules and predicted direction
(vi)	2026-09-22 06:33:45	Prospective: before any benchmark output	JAVEditBench protocol and random selection of ten blind-review clips
(vii)	2026-09-22 14:50:37	Before any new-system blind result; one exploratory measurement at 07:50:31 preceded it	Replacement of AV-IRE+head by a single joint generator, with rationale and unresolved questions
(viii)	2026-09-22 15:13:12; amended 15:18:57, 15:23:30	Prospective: before training or development outputs	Audio-branch configurations, development sets, floors, margin and tie-breaking order
(ix)	2026-09-23 00:59	Prospective: before the page or any rating	Final-system review: 396 items including repeats, endpoints, tests and reviewer roles
(x)	2026-09-23 15:55:10	Authors knew automatic InsAVE scores; no human rating existed	InsAVE review: four editors, 196 main items, one McNemar contrast on J
(xi)	2026-09-24 09:43	Prospective: before follow-up outputs, pages or ratings	Second-reviewer coverage, audio-only intervention and fresh test set

Amendments. Record (i) was appended three times before any result existed, noting that the composite media had been generated, adding an attribution statement and ruling that the review page would carry four systems rather than three. Record (viii) was amended twice after a wiring smoke test and before any development generation. Record (x) carries seven appended revisions; a planned second review round on AV-Remove-Bench (Guo et al., 2026) was withdrawn before any of its items existed, so the remaining InsAVE contrast required no multiple-comparison adjustment under the rule fixed at freezing. Apart from these and the revisions of record (ix), the records were written once and not edited.

Revisions of record (ix). Record (ix) has ten appended revisions. Four precede rating: they split the question wording between the main and appended sets, pin the appended JAVEditBench set to the ten clips already drawn, and register the build, a corrected packing defect and each configuration's executor checkpoint. The six after unblinding record the second reviewer's status; fix, before any InsAVE output, how the InsAVE results are read and that inaudible outputs are not successes; correct the table of configurations and stereo-averaged level figures; and fix a descriptive reading of the automatic InsAVE gap after it was known, before its decomposition. None altered an endpoint, the size, a ranking rule or a judging criterion.

Outcome of record (viii). Its margin $m = \max(2, |R - C|)$ compares the final configuration's development count C with that of a retrained replicate R from the same stage-one checkpoint with the same seed, so it measures execution nondeterminism only and likely understates seed-to-seed variation. With $m = 4$ the rule required more than 32 of 60 bidirectional development successes;

the three candidate configurations reached 28, 29 and 26 against the final configuration's 28, so no revision was adopted and the final system is the unrevised routed generator.

Reviewers. Table 16 lists the human review rounds. Every round before that of record (ix) has one rater, so its repeats measure consistency within that rater only. Record (ix) adds a second reviewer on a pre-drawn 120-item overlap, who enters only as a reliability estimate and a conservative-merge sensitivity analysis; the first rater defines every endpoint. No rater of these rounds is an author or was involved in the project, none had seen the models or their outputs before rating, and each round had a different first rater. Record (xi) adds three rounds. In the first, the second reviewer of record (ix), who had rated only the overlap and seen no result by system, completes the AV-IRE and Task-Adapted LTX-2.5 queries. The audio-only round and the fresh-set round each have one rater, who is not an author and was not involved in the project, and the fresh-set rater is new and differs from the audio-only rater.

Table 16: Human review rounds. Counts are rated output items, not independent recordings. A plus sign adds repeated items; the final round's total of 396 already includes 16 repeats. All raters are independent non-authors.

Round	Systems and subset	Items rated	Raters
Earlier round	Four systems, Natural-2	240 + 10 repeats	One
JAVEditBench	Four systems, ten clips	40 + 2 repeats	One
Natural-3	AV-IRE+head, earlier subset	45	One
Final round, (ix)	Six configurations, Natural-2; appended JAVEditBench items	396 including 16 repeats	First rater
Overlap, (ix)	Subset of the same outputs	120	Second reviewer
InsAVE, (x)	Four systems, 49 queries	196 + 12 repeats	One
Coverage, (xi)	AV-IRE and Task-Adapted LTX-2.5, Natural-2	80 + 8 repeats	Same second reviewer
Audio only, (xi)	Three configurations, Natural-2	180 + 9 repeats	One
Fresh set, (xi)	AV-IRE and Task-Adapted LTX-2.5, 30 new inputs	120 + 4 repeats	One new rater

Correction of the rater record. Record (x) describes its rater as an author who knew the automatic InsAVE readings before rating. Both statements are wrong: the rater is not an author and did not know those readings; the authors knew them. The correction was made on 2026-09-24, after all ratings were complete, and no rating file, count or statistic changed. The second reviewer of record (ix) is not an author and was not involved in the project; this was first misreported as a co-author and corrected before the paper was written.

I Baseline interfaces and public subsets

Unrated comparators on AV-SwapTarget. The annotation budget, fixed in record (i) of Appendix H before any per-system result was seen, initially allocated one round over 60 Natural-2 queries for three systems and was amended to four systems before any result, as recorded in Appendix H. The composite subsets, the Kiwi cascade and the unadapted backbone therefore do not enter this round of human rating. All comparators' Natural-2 outputs predate the review page, so this is a declared budget, not a selection made after seeing scores.

Other human ratings. An earlier review round with different preservation wording passed the cascade on none of 10 items, 0 of 6 from Natural-2 and 0 of 4 from composites, counts that cannot be merged into Table 1. The unadapted backbone has never been rated, and no claim rests on a human judgment of it.

Run settings.

- *Kiwi cascade.* Kiwi-Edit with instruct-only 5B weights runs 50 steps at seed 0 with guidance 5.0 on the largest supported $4k+1$ frame count up to the input length and 121 frames, retimed to the input frame rate. CoherentAVEdit then runs once per item at seed 42 with 100 audio-generation steps and 16 kHz audio. The text protocol is instruction-only: the negative prompt is the target phrase, the positive one generic background sound, plus other sounds when the instruction says to preserve everything else.
- *InstructAV2AV.* The released removal checkpoint `removal.safetensors` runs 50 steps with seed 103 plus the manifest record index, up to 121 frames, at native 1280×704 with 16 kHz mono audio.
- *Task-Adapted LTX-2.5 and unadapted backbone.* Both use AV-IRE's 30-step, 512×320, 121-frame configuration with seed 42 and the target phrase in the prompt, differing only in whether the task LoRA is loaded.

I.1 Public single-target transfer

Benchmarks and audio window. InsAVE has 49 person-removal queries with video-only references; AV-IRE was generated on all of them at native length. JAVEditBench subject removal (25 queries, 1280×720, 121 frames at 25 fps) was run natively, with audio resampled from 16 to 48 kHz. Because the backbone's audio encoder maps the full 4.84-second window (232,320 samples) to 122 tokens, the reference audio is truncated to 232,080 samples (4.835 s, 121 tokens) and ends 5 ms before the video, an adaptation fixed before any output existed and applied to all four systems.

Scoring conventions. On JAVEditBench, rem. in Table 3 counts the 22 clips with an audible target, excluding three declared targets found inaudible on input-only listening. Over all 25 clips the criterion passes the Kiwi cascade on 22 and InstructAV2AV on 9. SyncNet finds a syncable face on only 1 to 2 clips per system after subject removal, so it is not reported.

JAVEditBench blind review round. An earlier blind review round rated the ten clips drawn at random under record (vi) for four systems, 40 main items and 2 repeats, using the eight-question Natural-2 form with single-subject preservation wording. The repeats were answered identically on 8 of 8 and 7 of 8 questions with the same J, and an unblinded frame check of all 40 items found every visual-removal rating consistent with the frames.

Inaudible Task-Adapted LTX-2.5 outputs. Of Task-Adapted LTX-2.5's outputs on this benchmark, 13 of 25 have whole-clip RMS below -60 dBFS. Of its 10 rated clips, 5 are inaudible tracks (RMS -126.5 to -126.7 dBFS, peak -79.3 to -80.0 dBFS) and a sixth is nearly so, although no output has an exact-zero sample. AV-IRE's rated clips lie between RMS -29.1 and -77.6 dBFS, none inaudible. The removal criterion, which passes Task-Adapted LTX-2.5 on 25 of 25, cannot tell an inaudible output from a removal on a single-speaker clip.

Per-criterion blind counts. The table below counts the 10 clips passing each criterion in the earlier round and in the final system's round, which rated the same clips with the same wording and gave AV-IRE and Task-Adapted LTX-2.5 no wrong-source judgment.

Round	System	VR	AR	VP	AP	C
Earlier	Task-Adapted LTX-2.5	10	10	9	8	10
	InstructAV2AV	10	3	10	10	4
	Kiwi cascade	10	7	10	0	2
Final-system round	AV-IRE	10	9	10	10	10
	Task-Adapted LTX-2.5	10	10	10	10	10

Classifier against listener. Over the 40 items of the earlier round the classifier removal criterion agrees with the human acoustic judgment on 37 ($\kappa = 0.85$), with no interval since the items are 10 clips crossed with 4 systems. All three disagreements are Kiwi cascade clips that the classifier passes

and the reviewer heard as incompletely removed, so its removal count is potentially optimistic here. This holds for single-speaker clips only, not for the multi-source scenes of Section 4.1.

InsAVE automatic readings. AV-IRE's containers carried an integer-truncated frame rate on 38 of 49 clips, so we scored a copy with only the time base corrected, verified to have identical frames and audio. Task-Adapted LTX-2.5's outputs include six inaudible tracks, all counted as removals; excluding them, the criterion credits Task-Adapted LTX-2.5 on 42 of 43 queries and AV-IRE on 26 of 43. AV-IRE fails on 20 queries Task-Adapted LTX-2.5 passes and passes none that it fails, indicating weaker suppression under this criterion. A first analysis that read the wrong field was withdrawn. The reading rule of record (ix), registered before any output here existed, returned its ordinary reading.

InsAVE blind review round. The review round of record (x) holds the 49 queries for AV-IRE, Task-Adapted LTX-2.5, InstructAV2AV and the Kiwi cascade: 196 items and 12 repeats in two sessions of 104, with the single-subject wording. Task-Adapted LTX-2.5 passes all five criteria on 47 queries, AV-IRE on 31, InstructAV2AV on 29 and the cascade on 1. Paired against Task-Adapted LTX-2.5, AV-IRE is $b = 1$, $c = 17$ (exact McNemar $p = 1.5 \times 10^{-4}$).

InsAVE failures and repeats. AV-IRE passes visual removal and both preservation criteria on all 49 queries and consistency on 45, and removes a wrong source once in each modality. Its 18 acoustic-removal failures leave the target speech partly (11) or clearly (7) audible, while Task-Adapted LTX-2.5 passes acoustic removal on all 49. The cascade removes the person from the frame on 44 queries but preserves the other sounds on 2 and removes a wrong sound on 45. Each system's three repeats received the same joint verdict, the two sessions' joint-success rates are 0.558 and 0.543, and one blank artifact answer enters no endpoint.

InsAVE rater. The planned second reviewer was cancelled, so this round has a single independent rater and no inter-rater estimate. The authors knew before rating that AV-IRE's automatic removal count on this benchmark trailed Task-Adapted LTX-2.5's, and record (x) was written after that; the rater did not know it. The automatic scores, produced independently of the ratings, point the same way as the blind verdicts.

J QUALITATIVE SELECTION

Selection rule. The input of Figure 1 is the first input in identifier order on which AV-IRE succeeds on both instructions in the blind ratings of the review round of Table 1. Because the rule is applied to ratings, the figure shows a selected case, not typical performance. The candidate lists are released with the ratings, and every frame and soundtrack shown is matched to the rated output.