# CIViC: Evidence Gates for Audio-Conditioned Visual Prediction

**Anonymous authors**
Paper under double-blind review

## Abstract

Removing sound from a visual predictor may change its output, but that test leaves the model's use of the intended audio information uncertain. We call this protocol Conditional Input and Visual Interpretation Checks (CIViC). It links a specific interpretation to the intervention, task comparison, or target check needed to support it. We apply the protocol to an actor-disjoint face-image task, a synthetic feature-mixture task, and a fresh source-video-disjoint SpeakerVid proxy task. The static face-image test detects a small response to replacing audio with another actor's; a separate temporal test does not establish an audio benefit on its endpoint. In a post-confirmation comparison on a second actor-disjoint raster target, four constant-audio fits score better than their audio-trained counterparts, despite a large response when audio is removed at test time. In the synthetic feature-mixture task, the added condition changes a 128-coordinate visual proxy but raises withheld-composition error from 0.1707 for audio alone to 0.1807 at 1,000 steps; the gap shrinks with longer fitting. In the source-video-disjoint SpeakerVid task, a same-target audio-zero arm has lower motion-proxy RMSE than aligned audio in all four held-out conditions and all five paired seeds on four accelerator platforms using one source pack. The same direction appears within each of the four held-out source-video groups, but a secondary long-form drift proxy favors aligned audio in all four conditions. Audio-only, dummy-condition, and shuffled-condition arms expose distinct endpoint behavior, showing why an output response alone cannot justify a conditional interpretation.

## 1 Introduction

Audio-conditioned video and avatar models often expose paths for speech, speaker identity, timing, and structured scene information (Corona et al., 2025; Kong et al., 2025; Zhong et al., 2025; Peng et al., 2025; Chu et al., 2026). A common intervention removes or perturbs one path and checks whether the output changes. That test is useful, but it leaves several interpretations open. A model can react to audio removal while ignoring the identity of replacement audio. An auxiliary condition can alter predictions while worsening the held-out task. A proxy ordering can appear favorable even when a condition-only comparator predicts its target better than the audio-informed model.

This paper asks a narrow question: which checks distinguish input activation from a supported audio-conditioned visual prediction statement in these controlled tasks? CIViC answers this question with four gates. Each gate targets one alternative explanation. We fix the split, intervention, score, aggregation rule, and decision rule before held-out scoring. CIViC is not a video generator. It is a diagnostic framework that makes a conditional interpretation falsifiable before it is used to motivate a broader audiovisual statement. It does not establish that these four checks are necessary or sufficient outside the tested settings.

Prior audio-driven avatar systems evaluate generation, synchronization, or multi-person control (Corona et al., 2025; Kong et al., 2025; Zhong et al., 2025; Peng et al., 2025; Chu et al., 2026); other work studies disentangled audio representations or disrupts audio control through portrait perturbations (Mittal & Wang, 2020; Gan et al., 2025). MultiTalk targets audio–person binding, and DualTalk studies speaker–listener role transitions. CIViC does not rank those generators or assess their perceptual outputs. Its contribution is a mapping from a specified conditional interpretation to the matched intervention, aggregation unit, and endpoint needed to evaluate it; the case studies introduce no new architecture.



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.

The closest intervention protocol is the lip-leakage evaluation of Yaman et al. (2026), which tests silent input, mismatched audio, matched audio, and reference-frame selection on talking-face generators using LRS2. These tests already show why matched synthesis alone cannot identify audio control; our removal and replacement actions are not new. CIViC adds a held-out task comparison and a condition-only endpoint comparator within a registered interpretation rule. It does not evaluate the generators or perceptual endpoints in that work. Weng et al. separate speech, effects, and music into temporal controls (Weng et al., 2025); our feature-space study instead tests a withheld combination of synthetic perturbations and a factor vector against a compact proxy. Across these papers, the tasks and evidence differ; CIViC is a bounded interpretation-to-control mapping, not a new identification method or a general evaluation standard.

Table 1 makes the distinction from familiar evaluation actions explicit. Audio removal, replacement, held-out scoring, and shortcut controls each answer a useful but narrower question. CIViC does not replace any of them; it registers the interpretation for which each is needed and blocks that interpretation when its required evidence fails.

Table 1: From common evaluation actions to bounded interpretations. The entries are complementary, not a ranking of prior work.

| **Evaluation action** | **What it can establish** | **What CIViC adds before the stated interpretation** |
|---|---|---|
| Audio removal or ablation | The registered path affects this output | A matched replacement for a value-specific statement |
| Input replacement | The stated alternative changes this endpoint | Held-out task evidence when benefit is asserted |
| Held-out task score | An ordering on the registered endpoint | A shortcut control for a transfer or proxy statement |
| Perturbation or defense test | Robustness to the tested disruption | A gate-specific stop rule and retained failed gate |

For anonymous review, we provide the portable evaluation toolkit and lightweight checks through a 4open.science Git mirror.[1]

We apply CIViC to actor-disjoint face-raster checks, a synthetic feature-mixture task, and a fresh source-video-disjoint SpeakerVid task (Zhang et al., 2026). The studies retain a small static detection, an early proxy disadvantage that shrinks with fitting, and a source-video-disjoint endpoint warning. They do not establish general audio-conditioning benefit or failure. Disjointness is a split property: repeated seeds and hardware runs do not add independent actors or source videos.

**Contributions.** The paper makes four contributions:

1. CIViC, a four-question diagnostic framework that separates input activation, substitution specificity, held-out benefit, and endpoint validity;
2. actor-disjoint CREMA-D raster checks: the static protocol satisfies its registered detection rule but has a very small cross-actor audio effect, while a separate temporal renderer does not establish an audio benefit on its endpoint;
3. a synthetic audio-feature study showing an active condition branch, a shrinking error gap with training, policy-dependent controls, and a proxy endpoint shortcut; and
4. a fresh source-video-disjoint SpeakerVid replication with a five-arm, five-seed control matrix, four-platform execution replication, and a post hoc check across its four held-out source-video groups.

Table 2 records the decision rule. A detection excludes only its named alternative; broader or practically important interpretations also need a task-scale threshold or equivalence region registered with its source, unit, direction, and failure action. No threshold was registered for the tiny CREMA-D

[1]Anonymous repository and documentation: 4open.science project page; source code: download the code-only ZIP, also provided as a separate supplementary file. The archive includes deterministic fixtures, aggregate receipts, and a pinned lightweight environment, but not restricted study inputs or assets needed to replay the full experiments.

cross-actor contrast, so it remains practically unclassified. The three-seed Study II result is a fixed-run direction without a held-out-case interval or population inference.

Table 2: Interpretation rules. A detection result excludes only the named alternative; practical use needs a task-scale rule.

| Gate | Required evidence | What a detection does not establish |
|---|---|---|
| Activation | Output changes after the registered removal | Which input values the model uses |
| Substitution | Matched replacement changes the stated endpoint | Task benefit or general audio use |
| Benefit | Matched held-out error comparison | A mechanism for any observed gap |
| Endpoint validity | Control excludes the named shortcut | Validity outside the checked endpoint |

## 2 CIViC maps interpretations to four evidence questions

Let $f(x, a, z)$ map a fixed visual context $x$, audio $a$, and optional auxiliary condition $z$ to a motion target $y$. A nonzero difference between $f(x, a, z)$ and $f(x, 0, z)$ only shows that removing audio changes the output. Figure 1 maps four CIViC gates to the alternative explanation each gate excludes. The gates are not a universal checklist. A study runs the gates required by its intended interpretation. For example, a speaker-specific statement needs a registered speaker-audio substitution, a mixed-condition benefit needs a matched task comparison, and a proxy-transfer statement needs an endpoint control that excludes non-audio shortcuts.

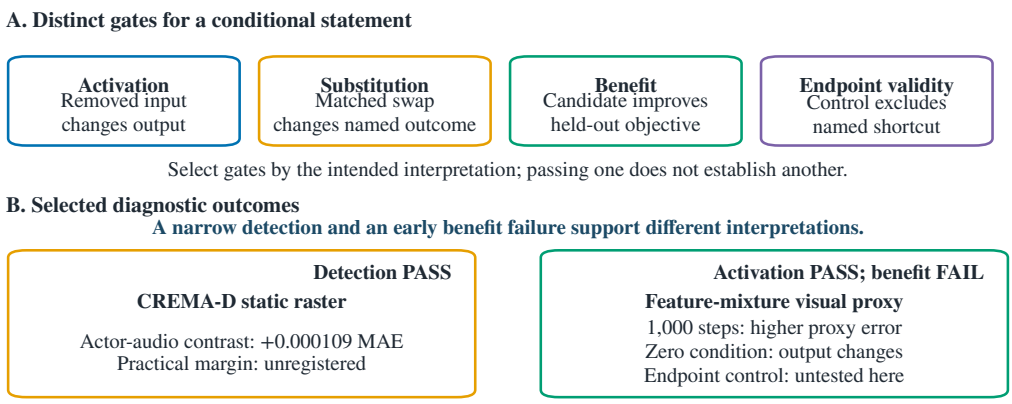



Figure 1: Four interpretation gates (A) and three bounded outcomes (B). CREMA-D yields a tiny static detection. The feature-mixture condition changes its 128-coordinate proxy without early benefit; the gap nearly closes with training. The original feature-mixture endpoint has no matched no-audio comparator, whereas the fresh SpeakerVid proxy does. Color and text distinguish the gates.

The gates are selected for the intended interpretation and fix the split, intervention, score, aggregation unit, and decision rule before held-out scoring. An endpoint warning excludes the named interpretation but does not identify its shortcut; a new control must exclude the competing endpoint result before that interpretation is restored. Repeated fits measure run variation, not added population coverage. For a practically useful actor-matched raster statement, the protocol card therefore requires an audio replacement, a held-out score, a task-scale threshold, and an endpoint control when the target is a proxy. Figure 1 contrasts the narrow CREMA-D detection with the failed early-benefit check; Table 3 gives the paired raster results.

# 3 Study I: static and temporal face-raster checks

## 3.1 CREMA-D meets a narrow fixed-pair detection rule

The first raster application uses CREMA-D speech-video recordings (Cao et al., 2014). A fixed protocol assigns 73 actors to development, 9 to validation, and 9 to sealed confirmation. The renderer receives an identity raster decoded at 0.25 seconds and a waveform-derived log-mel window from 0.75 to 1.25 seconds. It predicts the $64 \times 64$ target raster at 1.0 seconds. The audio window extends 0.25 seconds beyond the target time; this is a centered-window raster reconstruction. Inputs exclude emotion, intensity, sentence, and demographic labels. For this raster application, mean absolute error (MAE) averages absolute RGB intensity differences over all $64\times64\times3$ values after division by 255, so it is measured on a $[0, 1]$ channel scale. We sealed the confirmation protocol, pairings, and bootstrap rule before opening the 732 confirmation clips. For actor-audio replacement, actors follow a fixed cycle; a protocol hash and clip identifiers select a clip from the next actor. Utterance, emotion, and intensity are not matched, so this control changes speaker and speech content together.

Table 3 reports the fixed paired contrasts. Every registered lower interval bound is positive. The cross-actor-audio effect is 0.000109 MAE, however. This meets the registered sign-based detection rule, but no task-scale threshold was registered, so its practical relevance is unclassified. It does not establish robust speaker transfer, temporal generation, natural multi-person behavior, active-speaker accuracy, or human-perceptual quality.

Table 3: CREMA-D static confirmation on 732 fixed pairs from an actor-disjoint split. Differences are control MAE minus aligned MAE; aligned MAE is 0.02907. Paired intervals use 10,000 clip-level resamples with seed 1701 for the fixed checkpoint and pair set; they do not quantify actor-population uncertainty. Positive differences favor aligned input.

| **Control** | **Control MAE − aligned MAE ↑** | **Paired 95% interval** |
|---|---|---|
| Zero audio | +0.131475 | [0.127163, 0.135640] |
| Actor-audio replacement | +0.000109 | [0.000034, 0.000186] |
| Identity-raster replacement | +0.064860 | [0.062795, 0.066897] |

## 3.2 A separate temporal renderer fails its audio controls

The static test does not answer whether audio improves a sequence renderer. A separately registered renderer was fitted on 5,038 development sequences, selected on 689 validation sequences, and evaluated on 708 sealed sequences. Aligned MAE was 0.030449; shifting audio raised it by 0.0000176, while repeating the identity raster lowered it by 0.010592. Without an interval or practical margin for the shift, this check does not establish a sequence-level audio benefit or measure perceptual video quality.

## 3.3 A separate confirmation fails the audio-replacement gate

A second actor-disjoint raster protocol uses the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018). The old recap scored 1.0 seconds, not the registered 2.0-second target, so we replayed the frozen 240-pair protocol and checkpoint at 2.0 seconds. Aligned MAE is 0.08990. Zero audio raises it by 0.20931 (paired 95% interval $[0.20469, 0.21358]$), while another actor's audio changes it by $-0.00045$ ($[-0.00129, 0.00039]$); the registered replacement rule fails. Exploratory filename-matched cross-actor and same-actor statement swaps change MAE by $-0.00102$ ($[-0.00491, 0.00141]$, four actors) and $+0.00004$ ($[-0.00017, 0.00025]$), respectively.

A same-architecture constant-audio fit scores 0.08652 on those targets, versus 0.08990 for the audio-trained fit (difference $-0.00338$, descriptive four-actor bootstrap 95% interval $[-0.00512, -0.00164]$). Three matched-seed fits also favor constant audio; all four differences range from $-0.00338$ to $-0.00216$. A fresh four-seed, 80-epoch repeat favors constant audio in all four pairs, but trajectories diverge before epoch 40, so it does not isolate extra fitting. These retrospective comparisons do not create a registered benefit gate. Removal is activation-sensitive, but replacement fails and neither swap isolates speaker identity because speech content also changes. Appendix B gives the replay and within-trajectory details; the old recap is superseded.

### 3.4 An incompatible external audit tests a separate boundary

We apply the frozen renderer to 33 AVA Active Speaker validation recordings (Roth et al., 2020). The shared soundtrack has no per-face audio, so this is an external boundary check only. Aligned, shifted, and zero-audio AUROC are 0.63678, 0.63670, and 0.63672; the aligned-minus-shifted interval is $[-0.00017, 0.00028]$. This check supplies no speaker-attribution evidence.

## 4 Study II: an active condition offers no stable benefit on a withheld feature mixture

### 4.1 Controlled composition protocol

Study II uses 819 indexed synthetic feature mixtures from a private, provenance-approved source family: 615 train, 96 validation, and 108 test. These counts do not verify unique source clips or cross-split disjointness and are not treated as independent samples. Each mixture starts with five speech summaries and adds the same nuisance sequence to both speakers. Environment-like features are seeded normalized noise; music-like features are seeded sinusoids, with either slow/transient noise and an optional beat envelope. They are five-coordinate feature arrays, not waveforms, scaled to 0.25 and 0.20 times speech-feature standard deviation. The 128-coordinate visual proxy is copied from the source pack. RMSE is computed over these dimensionless coordinates without per-condition rescaling; it is not a physical motion or video-quality measure.

Both arms receive five audio summaries, a 256-coordinate speaker embedding, and six dialogue coordinates. The added condition is the seven-bit vector $[s, e, m, e_0, e_1, m_0, m_1]$, encoding speech, environment, music, and two type indicators per nuisance stream. The shared past-only backbone uses a 96-frame context, width 256, and four attention heads. Adding the condition increases input width; the joint mixed condition is withheld from training. The primary comparison is the 1,000-step checkpoint, with the same three seeds checked at 3,000 and 10,000 steps. Separate analyses test training with all conditions present, 24-seed training-support policies, five-seed parameter-matched controls, and source-disjoint scalar endpoint shortcuts; their estimands are not pooled. The registered fixed-run rule requires positive structured-minus- audio gaps for all three seeds to indicate no benefit at 1,000 steps. This is a directional result without a held-out-case interval or population inference.

### 4.2 The early error gap shrinks with training

Table 5 reports three checkpoints from the same runs. At 1,000 steps, all three paired mixed-condition gaps are positive; by 10,000 steps the mean is $0.0019 \pm 0.0032$ and one seed gap is negative. The early disadvantage therefore shrinks with fitting. Figure 2 shows that zeroing the condition changes predictions by $0.2123 \pm 0.0070$ and replacing the mixed vector with the clean vector changes them by 0.1473 on average; neither displacement measures task benefit. Clean, environment-only, and music-only test gaps are −0.000284, 0.001008, and 0.000909, versus +0.010032 on the withheld mixture. Relative to clean, mixture error rises by +0.0095 for audio only and 0.0198 with the condition. The 24-seed support audit has a positive structured-minus-audio gap for every policy (Table 4). With all conditions in training, a separate three-seed control gives mixed RMSE $0.1659 \pm 0.0027$ for audio only and $0.1684 \pm 0.0047$ with the condition; it does not test withheld composition. Together, these results show activation without a demonstrated held-out gain in these settings, not a training-independent penalty.

Table 4: Twenty-four-seed, 1,000-step training-support audit on the held-out mixed condition (mean ± population standard deviation). The gap is added-condition minus audio-only RMSE; positive gaps favor audio only. Each policy changes training support and is not pooled with the three-seed result.

| **Training support** | **Audio-only RMSE ↓** | **Added-condition RMSE ↓** | **Gap ↓** |
|---|---|---|---|
| Single clean | 0.4249 ± 0.0097 | 0.6921 ± 0.0218 | +0.2672 ± 0.0276 |
| Single environment | 0.1638 ± 0.0026 | 0.4192 ± 0.0239 | +0.2554 ± 0.0240 |
| Single music | 0.1681 ± 0.0025 | 0.4169 ± 0.0247 | +0.2487 ± 0.0252 |
| Clean plus environment | 0.1685 ± 0.0027 | 0.4121 ± 0.0225 | +0.2436 ± 0.0231 |

Table 5: Withheld mixed-composition error at three checkpoints of the same three-seed runs (mean ± population standard deviation; lower RMSE is better). Audio only omits the seven-bit condition; both arms retain speaker and dialogue inputs. The gap is added-condition minus audio-only RMSE; positive favors audio only.

| Steps | Audio-only RMSE ↓ | Added-condition RMSE ↓ | Gap ↓ |
|---|---|---|---|
| **1,000** | 0.1707 ± 0.0024 | 0.1807 ± 0.0059 | +0.0100 ± 0.0040 |
| 3,000 | 0.1338 ± 0.0004 | 0.1366 ± 0.0023 | +0.0028 ± 0.0019 |
| 10,000 | 0.1237 ± 0.0014 | 0.1256 ± 0.0042 | +0.0019 ± 0.0032 |

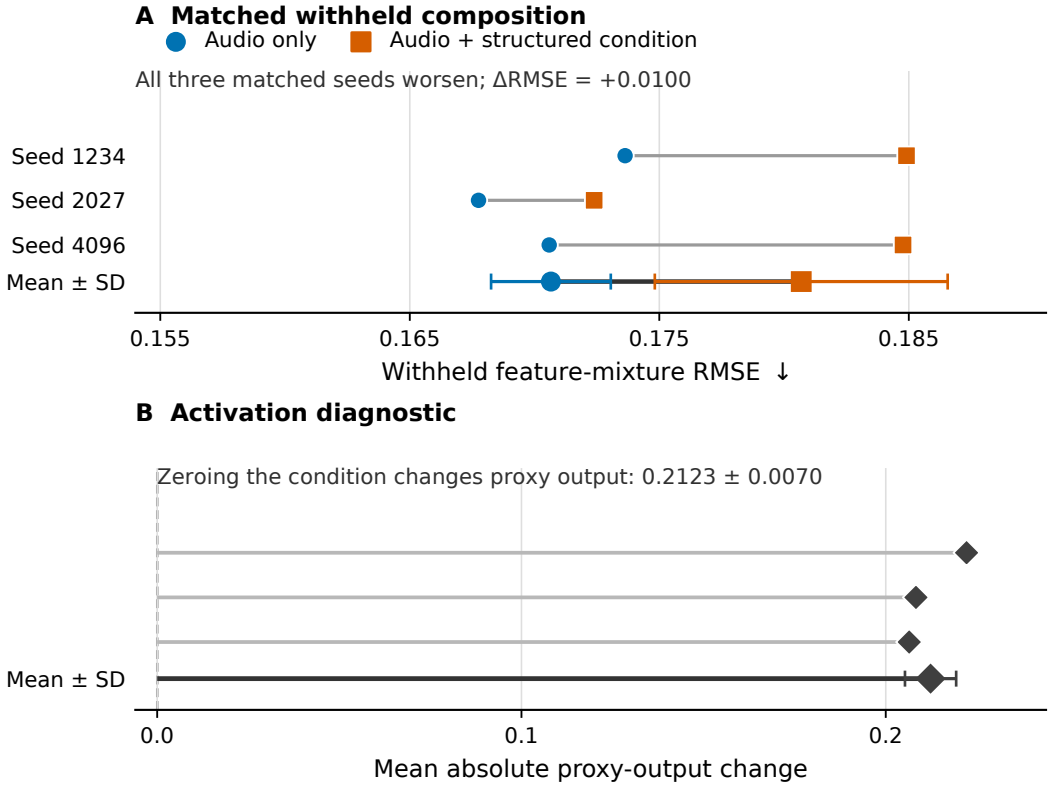



Figure 2: Registered 1,000-step feature-mixture evidence. (A) Adding the condition raises withheld-composition RMSE in all three matched seeds; the target is a dimensionless stored proxy. (B) Zeroing the condition changes proxy predictions in the same runs. Points are seed values; markers are mean ± standard deviation. Activation is not a task-error comparison.

### 4.3 A PARAMETER-MATCHED AUDIT EXPOSES A TRAINING-SUPPORT MISMATCH

The primary comparison changes input width, so we separately tested five parameter-matched controls under clean-plus-environment support. Each uses the seven-coordinate interface and 4,908,416 parameters (Table 6). The dummy-condition model scores 0.1822 ± 0.0025; training with shuffled conditions but aligned evaluation scores 0.4248 ± 0.0217; and audio-zero with aligned conditions scores 0.5003 ± 0.0278. These policies change multiple inputs or train–evaluation support, so their gaps isolate neither audio nor condition alignment. The same-checkpoint intervention in Figure 2 supplies the condition-activity test. This audit neither makes the primary three-seed comparison parameter matched nor shows benefit. An earlier reported train-by-evaluation matrix lacks source reports and is excluded.

Table 6: Five-seed parameter-matched controls on the held-out mixture under clean plus environment training support (mean ± population standard deviation; seeds 1234, 2027, 4096, 8191, 16384). All entries use seven condition coordinates and 4,908,416 parameters. Shuffled values are used in training but aligned values in evaluation; this is a policy-specific diagnostic.

| Control | Audio | Condition relation | Mixed RMSE ↓ |
|---|---|---|---|
| Dummy condition | Observed | Deterministic dummy | 0.1822 ± 0.0025 |
| Shuffled condition | Observed | Train shuffled; test aligned | 0.4248 ± 0.0217 |
| Audio-zeroed | Zeroed | Aligned | 0.5003 ± 0.0278 |

### 4.4 A condition-only comparator raises an endpoint warning

A source-file-disjoint CREMA-D ridge check (71 clips, three held-out actors, five mixtures) gives RMSE 0.00190136 for the interaction model versus 0.00189705 for audio only; no paired interval or practical margin was retained. A frozen RAVDESS ridge check over 1,441 clips and 24 source groups gives structured-minus-audio RMSE $-0.00000924$ with interval $[-0.00002350, 0.00000148]$; metadata-only RMSE (0.00881499) is below both audio-informed variants (0.00926477 and 0.00927402). Neither check isolates audio use in the scalar proxy.

A separate source-disjoint neural follow-up fits fixed width-32 ReLU regressors on CREMA-D and evaluates four held-out GRID talkers. Its structured-minus-audio contrast is $-0.0023303$ with talker-clustered interval $[-0.0032104, -0.0006693]$, but a condition-only comparator reaches RMSE near 0.00109, below the structured (0.00865–0.01341) and audio-only (0.01231–0.01592) models. These descriptive scalar-proxy checks show that an apparently favorable ordering cannot support the intended audio-conditioning interpretation; they do not identify the shortcut or validate Study II's 128-coordinate target.

## 5 Study III: a source-video-disjoint SpeakerVid replication

Study II's primary target also has 128 coordinates, but its separate shortcut follow-ups use a scalar proxy. We run a new source-video-disjoint study on SpeakerVid-5M (Zhang et al., 2026). The frozen selection has 865 decoded two-speaker windows from 31 videos (649/108/108 train/validation/test records across 23/4/4 videos). Source-clip, source-video, and exact-target overlap are zero between splits. One unverified video was excluded; this is a new source pack, not a replay of Study II. It does not recover Study II's missing original source-ID/split intersection or supply a matched comparator for that study.

Each record has five speech features, speaker and dialogue inputs, and a 128-coordinate visual proxy. An internal deterministic feature-mixture generator (`mixtext-v2`) provides clean, environment-like, and music-like training conditions plus a held-out mixture; its seven-dimensional vector encodes synthetic presence/type, not a waveform or natural-language prompt. Five trained arms are aligned speech and condition, zeroed speech with aligned condition, observed speech with the condition branch omitted, observed speech with a dummy condition, and observed speech with batch-shuffled conditions. They share the data, targets, split, 1,000-step schedule, and five seeds (1234, 2027, 4096, 8191, 16384).

We repeated the full matrix on four accelerator platforms: one H100, two A100 (labeled A100-1 and A100-2 in Table 7), and one V100. This produced 100 fits and 100 validation/test reports. They reuse one source pack and are compute replications, not new independent evidence. A post-hoc evaluation groups aligned and zero-audio predictions by the four held-out videos; it describes those groups only.

### 5.1 Audio-zero has lower test error across platforms and source groups

Table 7 reports the paired test difference in motion-proxy root mean squared error (RMSE), computed as control minus aligned audio for each seed. Negative values favor the control. Audio-zero has a negative mean in all 16 platform-by-condition cells, and all five paired seeds are negative in each cell. On validation, the same contrast is negative for all five seeds in every condition on every platform.

This ranking is specific to motion RMSE. A secondary segment-drift proxy, for which lower is better, has the opposite mean ordering for audio-zero minus aligned audio in all 16 test platform-condition cells. The four-platform ranges of five-seed means are $[0.0057, 0.0064]$ for clean, $[0.0076, 0.0079]$

for environment-like, $[0.0069, 0.0070]$ for music-like, and $[0.0043, 0.0050]$ for mixed conditions. These are repeated fits on one source pack, not intervals over new videos. The evaluator flattens the predicted feature windows, divides that array into 10-second segments, and compares later segment means with the first; it does not track drift in a continuous rendered clip. Appendix A, Table 8, reports the contrasting proxy alongside motion RMSE and its seed-level signs. Neither proxy establishes rendered-video quality; the drift result prevents an unqualified claim that audio-zero is better.

Table 7: Paired test motion-RMSE differences for the audio-zero control, computed as audio-zero minus aligned audio. Entries are mean ± sample standard deviation over five paired seed differences. Negative values favor audio-zero. The four platform runs share one source pack and are hardware or execution replications, not independent data. The secondary segment-drift proxy favors aligned audio in every condition; Table 8 reports that opposing result.

| **Condition** | **H100 ΔRMSE↓** | **A100-1 ΔRMSE↓** | **A100-2 ΔRMSE↓** | **V100 ΔRMSE↓** |
|---|---|---|---|---|
| Clean | −0.0142 ± 0.0036 | −0.0143 ± 0.0035 | −0.0143 ± 0.0035 | −0.0142 ± 0.0036 |
| Environment-like | −0.0214 ± 0.0022 | −0.0215 ± 0.0022 | −0.0215 ± 0.0022 | −0.0216 ± 0.0024 |
| Music-like | −0.0210 ± 0.0019 | −0.0211 ± 0.0019 | −0.0211 ± 0.0019 | −0.0213 ± 0.0020 |
| Mixed | −0.0206 ± 0.0024 | −0.0208 ± 0.0024 | −0.0208 ± 0.0024 | −0.0212 ± 0.0027 |

The audio-only model, which keeps observed speech but omits the condition branch, has lower mean test RMSE in all 16 platform-condition cells. All five seed differences are negative except music, where three platforms have four negative seeds and V100 has five. Because input dimensionality changes, this is a branch-omission comparison, not an isolated audio effect. The dummy-condition model has higher RMSE in every cell and seed. Shuffled conditions have small positive test means (0.0006–0.0022) but variable seed signs. On validation, audio-only means are lower in all cells (four of five negative seeds for clean, all five for the other conditions); dummy-condition means are higher in every cell and seed. Shuffled-condition means are positive in all 16 validation cells, but seed signs vary. Validation remains separate from the test result. Appendix A, Table 9, shows the full test-condition contrast: the audio-only arm favors omission of the condition branch, whereas the dummy and shuffled controls do not on motion RMSE. Their secondary segment-drift means have a different ordering, detailed in Appendix A; this is not a single-metric ranking of the complete system.

The post-hoc grouping has 108 unique test windows from four videos, each scored in four conditions (432 video-condition rows). Every video-condition mean and all five seed differences are negative on each platform; means range from −0.0334 to −0.0123. These four groups are the complete held-out video subset, so this is descriptive consistency, not a population interval. Appendix A, Table 10, separates the four source videos; no one video reverses the audio-zero ordering in any condition, but these are not four independent test populations.

A raw per-run audit covers 15 numeric evaluator fields and the nested condition-intervention outputs. On test, replacing the aligned vector with zero moves predictions by 0.2127–0.2129 (seed SD 0.0115–0.0117) in aligned fits and 0.2614–0.2619 (SD 0.0104–0.0105) in audio-zero fits. Replacing it with the clean vector moves them by 0.1570 (SD 0.0105) and 0.1338–0.1341 (SD 0.0080), respectively. These are displacements, not condition correctness or task benefit; text fields are absent for audio-only because it has no condition branch.

Motion RMSE is the direct endpoint. Active/inactive RMSE are components; causal selectivity, target fidelity, and listener responsiveness are derived; non-target leakage duplicates inactive RMSE; and speaker-order delta is a raw perturbation diagnostic. Identity stability measures input-order invariance, not identity fidelity, while the coarse CIViC score is uncalibrated. For zero-audio fits, audio-video sync and modality consistency use a zero-energy fallback of 0.5, and lag is at its −400 ms search boundary; none is interpreted as synchronization performance. Turn-timing error remains a secondary proxy output. The bounded result is an endpoint warning for this compact target: aligned audio changes predictions, yet audio-zero has lower motion RMSE and higher segment-drift proxy scores on these held-out records. It establishes no natural-video quality or validity for Study II's separately sourced target.

## 6 What the combined evidence establishes

CREMA-D meets its fixed-pair audio-replacement rule, but the cross-actor effect is small, confounds actor with utterance, and has no registered practical threshold. Study II shows an active condition branch and an early held-out disadvantage that shrinks with fitting; its scalar follow-ups expose a condition-only shortcut. In Study III, audio-zero and audio-only lower proxy error, the dummy condition raises it, and shuffled-condition results vary by seed. The audio-zero ordering repeats across hardware on the same source pack and within the same four test videos. The secondary drift proxy instead favors aligned audio in all four conditions, so the motion-error ordering is not a metric-wide win. The post-confirmation RAVDESS fits and post-hoc SpeakerVid source-group checks are descriptive; their repeated contrasts do not support a familywise inferential statement. Condition-vector interventions move predictions but do not show target benefit. Across studies, these proxy results do not establish rendered-video quality, natural-mixture performance, speaker attribution, or human judgments.

For audio-zero and dummy, all five seed signs agree with their respective negative and positive mean in every cell and platform. Audio-only has five negative seed differences except in music-like conditions, where three platforms have four and the V100 has five. Shuffled conditions have positive platform means, but only two of five paired seeds are positive for clean, three or four for environment-like, four for music-like, and four or five for mixed, depending on platform. A positive platform mean is therefore not a seed-stable advantage for the aligned condition in this control.

Table 8 makes the opposing endpoint visible. For audio-zero, the mean difference in motion RMSE favors the control, whereas the mean difference in long-form drift favors aligned audio in every condition. This drift statistic averages distances between the first and later 10-second segment means after the feature windows are flattened in evaluation order. Those windows do not form one continuous rendered video; the statistic is not a perceptual or natural-video quality score. Clean and mixed conditions have four of five positive paired drift differences on each platform; environment-like and music-like have five of five. The ranges pool neither independent videos nor new data sources.

Table 8: Audio-zero minus aligned-audio test differences for two lower-is-better proxies. The drift statistic compares segment means in a flattened window array, not a continuous video. Each interval gives the minimum and maximum of four platform-level five-seed means on the same held-out source pack; it is not a confidence interval. A negative motion-RMSE difference favors audio-zero, while a positive drift difference favors aligned audio. The final column counts positive paired drift differences out of five seeds on each platform.

| **Condition** | **Δmotion RMSE ↓** | **Δlong-form drift ↓** | **Positive drift seeds** |
|---|---|---|---|
| Clean | $[-0.0143, -0.0142]$ | $[0.0057, 0.0064]$ | 4/5 |
| Environment-like | $[-0.0216, -0.0214]$ | $[0.0076, 0.0079]$ | 5/5 |
| Music-like | $[-0.0213, -0.0210]$ | $[0.0069, 0.0070]$ | 5/5 |
| Mixed | $[-0.0212, -0.0206]$ | $[0.0043, 0.0050]$ | 4/5 |

## 7 Conclusion

CIViC separates activation, substitution specificity, held-out benefit, and endpoint validity. In a fresh source-video-disjoint SpeakerVid task, the same-target audio-zero comparison lowers motion-proxy error across four conditions, five seeds, and four execution platforms sharing one source pack, while raising the secondary long-form drift proxy; the fitted dummy and shuffled controls show why those outcomes need separate interpretation. The other studies retain their narrower diagnostic findings. Human judgments of rendered outputs remain future work.

## Reproducibility Statement

Sections 2–5 specify the protocols and bounded results. The anonymous code supplement provides deterministic checks, media-free fixtures, and aggregate receipts; full experiment replay requires approved source material absent from the archive.

## AI Use Statement

OpenAI Codex helped edit manuscript prose, prepare submission keywords and a TL;DR, and locate related papers. Gemini and Claude reviewed the wording of earlier drafts. The authors take responsibility for the final text and citations.

## References

AMI Project. The AMI meeting corpus. University of Edinburgh corpus site, 2005. URL https://groups.inf.ed.ac.uk/ami/corpus/.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4): 377–390, 2014.

Xuangeng Chu et al. Unils: End-to-end audio-driven avatars for unified listening and speaking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chu_UniLS_End-to-End_Audio-Driven_Avatars_for_Unified_Listening_and_Speaking_CVPR_2026_paper.html.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. doi: 10.1121/1.2229005.

Enric Corona et al. Vlogger: Multimodal diffusion for embodied avatar synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Corona_VLOGGER_Multimodal_Diffusion_for_Embodied_Avatar_Synthesis_CVPR_2025_paper.html.

Yuan Gan, Jiaxu Miao, Yunze Wang, and Yi Yang. Silence is golden: Leveraging adversarial examples to nullify audio control in ldm-based talking-head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13434–13444, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gan_Silence_is_Golden_Leveraging_Adversarial_Examples_to_Nullify_Audio_Control_CVPR_2025_paper.html.

Zhe Kong et al. Let them talk: Audio-driven multi-person conversational video generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/66fdfe692a360dfbad1fba32f64694da-Abstract-Conference.html.

Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018.

Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3290–3298, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Mittal_Animating_Face_using_Disentangled_Audio_Representations_WACV_2020_paper.html.

Ziqiao Peng et al. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Peng_DualTalk_Dual-Speaker_Interaction_for_3D_Talking_Head_Conversations_CVPR_2025_paper.html.

Joseph Roth et al. Ava-activespeaker: An audio-visual dataset for active speaker detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020. URL https://research.google/pubs/ava-activespeaker-an-audio-visual-dataset-for-active-speaker-detection/.

Shuchen Weng, Haojie Zheng, Zheng Chang, Si Li, Boxin Shi, and Xinlong Wang. Audio-sync video generation with multi-stream temporal control. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/133239b0506b84c802a12b0e5a764a17-Abstract-Conference.html.

Dogucan Yaman, Fevziye Irem Eyiokur, Hazim Kemal Ekenel, and Alexander Waibel. Assessing identity leakage in talking face generation: Metrics and evaluation framework. *arXiv preprint arXiv:2511.08613*, 2026. URL https://arxiv.org/abs/2511.08613. Accepted to ICASSP 2026.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=U0O4uqALWl.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, Ying Qin, Huan Li, Shuiyang Mao, Wei Liu, and Wenhan Luo. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025. URL https://arxiv.org/abs/2511.23475.

## A SpeakerVid test contrasts by condition and source video

The test endpoint here is root mean squared error over the dimensionless 128-coordinate visual proxy. Each difference subtracts the aligned-audio arm's error from the control arm's error on the same held-out records and paired training seed. Negative differences favor the control. The four accelerator runs reuse one source pack and the same five seeds; their spread is descriptive hardware sensitivity, not a confidence interval over videos.

Table 9 separates four controls that change different things. Zeroing speech during fitting has lower error in every condition, and removing the condition branch also lowers the mean error. The dummy condition raises error for every seed. Shuffling condition vectors raises the mean error slightly, but the seed-level direction varies. These comparisons cannot be collapsed into one estimate of audio or condition use: the controls differ in fitted inputs, branch capacity, or condition pairing. The secondary segment-drift proxy adds a further qualification. Relative to aligned audio, audio-only lowers both mean motion RMSE and mean drift in every platform-condition cell. Dummy and shuffled conditions raise mean motion RMSE but lower mean drift in every such cell. For example, across the four platforms in the mixed condition, dummy-minus-aligned drift ranges from $-0.0017$ to $-0.0015$ and shuffled-minus-aligned drift from $-0.0044$ to $-0.0041$. These are descriptive five-seed means on the same four-video test subset, not independent-video intervals. Drift is computed over flattened feature windows, so the opposing directions do not establish improved continuous-video quality.

Table 9: Test motion-proxy RMSE difference (control minus aligned audio) by condition. Each cell is the minimum–maximum of four platform-level means, each computed from five paired seed differences. These ranges are not confidence intervals; platforms share the same held-out videos and seeds. Negative values favor the named control.

| **Condition** | **Audio-zero ΔRMSE ↓** | **Audio-only ΔRMSE ↓** | **Dummy ΔRMSE ↓** | **Shuffled ΔRMSE ↓** |
|---|---|---|---|---|
| Clean | [−0.0143, −0.0142] | [−0.0040, −0.0037] | [0.0165, 0.0167] | [0.0006, 0.0007] |
| Environment-like | [−0.0216, −0.0214] | [−0.0045, −0.0042] | [0.0165, 0.0167] | [0.0011, 0.0012] |
| Music-like | [−0.0213, −0.0210] | [−0.0046, −0.0042] | [0.0163, 0.0166] | [0.0011, 0.0012] |
| Mixed | [−0.0212, −0.0206] | [−0.0106, −0.0101] | [0.0108, 0.0111] | [0.0019, 0.0022] |

Table 10 asks whether the audio-zero result is concentrated in one held-out video. The 108 unique windows come from four videos, with 26, 20, 48, and 14 windows, respectively; each window is scored under all four conditions. Every video-by-condition cell has a negative mean on all four platforms, and all five paired seed differences are negative on each platform. The stronger mixed-condition difference for Unit 2 is a descriptive within-source observation, not an independently replicated effect.

Table 10: Audio-zero minus aligned-audio test motion-proxy RMSE within each held-out source video. Unit numbers are labels for the four videos, not new sampling units; $n$ counts unique windows per video, each reused across four conditions. Cells give the minimum–maximum of four five-seed platform means, not population intervals. Negative values favor audio-zero.

| **Unit** | $n$ | **Clean ΔRMSE ↓** | **Environment ΔRMSE ↓** | **Music ΔRMSE ↓** | **Mixed ΔRMSE ↓** |
|---|---|---|---|---|---|
| 1 | 26 | [−0.0124, −0.0123] | [−0.0175, −0.0172] | [−0.0189, −0.0188] | [−0.0173, −0.0169] |
| 2 | 20 | [−0.0135, −0.0127] | [−0.0267, −0.0258] | [−0.0243, −0.0237] | [−0.0334, −0.0326] |
| 3 | 48 | [−0.0161, −0.0158] | [−0.0232, −0.0231] | [−0.0225, −0.0222] | [−0.0204, −0.0199] |
| 4 | 14 | [−0.0134, −0.0132] | [−0.0192, −0.0191] | [−0.0199, −0.0195] | [−0.0185, −0.0177] |

Neither table estimates uncertainty over new videos. The video-level comparison covers the complete four-video held-out subset; repeated windows, conditions, seeds, and platforms do not increase its independent-video count. The result warns that output sensitivity to audio or the condition vector does not establish a benefit on this compact visual target.

## B RAVDESS raster confirmation replayed on the registered target

The registered RAVDESS speech-video protocol (Livingstone & Russo, 2018) uses an identity raster at 0.25 seconds, mono 16 kHz audio from 1.0 to 1.5 seconds, and a $64\times64$ target raster at 2.0 seconds. Sixteen actors are assigned to development, four to validation, and four to confirmation. The compact renderer uses a visual encoder, an audio multilayer perceptron, audio-derived scale and bias, and a visual decoder; inputs exclude emotion, intensity, statement, and actor identifiers. The 240 fixed-pair test compares aligned audio with zero audio, audio from another confirmation actor, and a swapped identity raster. Partner actors cycle through the confirmation actors; a hash of the frozen protocol and clip identifiers selects the partner clip. Statement and emotion are not matched, so the audio replacement changes actor and utterance together.

The corrected replay of the frozen protocol and checkpoint scores the registered 2.0-second target. Aligned MAE is 0.08990. Control-minus-aligned contrasts are +0.20931 for zero audio, −0.00045 for actor-audio replacement, and +0.04654 for identity-raster replacement. Their paired 95% bootstrap intervals are $[0.20469, 0.21358]$, $[-0.00129, 0.00039]$, and $[0.04362, 0.04926]$. The registered rule requires a strictly positive lower bound for every control; it therefore fails. No fitting or threshold selection used confirmation actors. The actor-audio swap also changes utterance, so it does not identify a speaker effect.

After that replay, we froze an exploratory six-condition analysis on the same 240 unique targets and the same sealed checkpoint. The cross-actor control uses the registered partner actor but matches modality, vocal channel, emotion, intensity, statement, and repetition in the RAVDESS filename. Another control keeps the actor and changes only the statement code. Neither code match guarantees synchronized phonemes in the 1.0–1.5-second audio window. The new analysis retains each target's six MAEs and groups uncertainty both by pair and by the four confirmation actors. The metadata-matched cross-actor delta is −0.00102, with pair-bootstrap interval $[-0.00176, -0.00029]$ and actor-cluster interval $[-0.00491, 0.00141]$. The actor-specific deltas range from −0.00674 to +0.00176; omitting actor 21 changes the pooled delta to +0.00089. The same-actor statement delta is +0.00004, with pair-bootstrap interval $[-0.00043, 0.00052]$ and actor-cluster interval $[-0.00017, 0.00025]$. The new aligned, zero-audio, original actor-swap, and identity-swap means agree with the registered replay within 0.000001 MAE. These post-replay controls are exploratory; neither their pair-level interval nor a retrospective threshold changes the registered decision, and four actor clusters do not support population inference.

To check whether the removal response corresponds to a useful fitted audio signal on this raster endpoint, we separately fit the same AudioFiLM architecture with a constant zero audio tensor for every example. We reused the development and validation actor split, $64 \times 64$ target, seed 1701, optimizer, batch size, 40-epoch maximum, patience, and validation-MAE checkpoint selection. The comparator selected epoch 35; the original selected epoch 18. Both checkpoints score the same 240 confirmation targets. The constant-audio renderer obtains MAE 0.08652 versus 0.08990 for the original. Its paired difference is −0.00338 MAE; the pair-bootstrap interval is $[-0.00401, -0.00275]$ and the four-actor bootstrap interval is $[-0.00512, -0.00164]$. For seed 1701, all four actor-specific mean differences are negative, from −0.00614 to −0.00143. The original comparator was planned only after the confirmation actors had been opened. Three further matched-seed comparisons followed a separate frozen plan on that same opened set. Table 11 reports every seed; the constant-audio fit has lower MAE in all four. The new training runs use seeds 1702–1704, with all fit and selection settings held fixed except the seed and audio input. They check the observed endpoint but cannot replace a prospectively registered benefit test or imply perceptual video quality. One actor-specific mean is positive in seed 1702 and two are positive in seed 1704; the latter constant-audio fit selected the 40-epoch cap. Actor-level and run-to-run uncertainty therefore remain.

We then repeated all four matched seed pairs with an 80-epoch maximum and the same patience-six validation rule. Every constant-audio fit again has lower MAE on the same targets, with differences ranging from −0.00307 to −0.00070. For seed 1704, the constant-audio checkpoint now selects epoch 48, so this repeat's negative difference is not capped at epoch 40. None of the new fits selects the 80-epoch cap. The new training histories differ from the earlier histories before epoch 40, however; differences between runs cannot be assigned solely to the longer budget. Table 11 shows both sets without treating seeds or reruns as new test populations.

To separate continued fitting from differences between fresh runs, a further frozen check retained the best validation checkpoint available by epoch 40 within each new 80-epoch trajectory, then scored that

checkpoint and the final selected checkpoint on the same 240 opened targets. Both arms selected the same checkpoint at or before epoch 40 in seeds 1701–1703: aligned-audio epochs 22, 15, and 19, and constant-audio epochs 34, 29, and 31. In seed 1704, aligned audio stayed at epoch 24, while constant audio selected epoch 38 by the old cap and epoch 48 after continuation. Its MAE fell from 0.08598 to 0.08538, an after-minus-before change of −0.00060. The constant-minus-aligned difference moved from −0.00298 to −0.00358 in that same trajectory. Actor-specific constant-audio changes range from −0.00310 to +0.00077; this single retrospective fit does not establish an actor-population effect. The check uses eight newly fitted trajectories and the same opened confirmation clips. It does not change the registered replacement result or convert the earlier independent 40-versus-80 fits into a controlled budget comparison.

Table 11: RAVDESS same-target sensitivity on 240 fixed targets. Maximum epochs is the stopping cap; validation selection can stop earlier. Δ is constant-audio-fit MAE minus aligned-audio-fit MAE, so negative favors the constant-audio fit. The 80-epoch runs are fresh post-confirmation fits, not exact continuations of the 40-epoch trajectories.

| Seed | Max. epochs | Aligned MAE ↓ | Constant MAE ↓ | Δ MAE ↓ |
|---|---|---|---|---|
| 1701 | 40 | 0.08990 | 0.08652 | −0.00338 |
| 1701 | 80 | 0.08803 | 0.08599 | −0.00204 |
| 1702 | 40 | 0.08979 | 0.08749 | −0.00230 |
| 1702 | 80 | 0.08874 | 0.08740 | −0.00133 |
| 1703 | 40 | 0.08984 | 0.08719 | −0.00265 |
| 1703 | 80 | 0.08841 | 0.08770 | −0.00070 |
| 1704 | 40 | 0.08853 | 0.08637 | −0.00216 |
| 1704 | 80 | 0.08873 | 0.08565 | −0.00307 |

The earlier outcome recap (aligned MAE 0.07404) is retained only in the private audit record. Its executed target time was not established by a pair-level report or code receipt, and its numbers are not used as paper findings. The new 2.0-second replay resolves the target-time integrity issue for the current result but does not expand the task beyond held-out actor raster reconstruction.

## C Development checks that stopped visible-speaker routes

These checks explain why the paper reports controlled raster and visual-proxy outcomes instead of a positive visible-speaker result. None opened a sealed confirmation set after its development or validation rule failed.

On AMI meeting recordings (AMI Project, 2005), a registered cross-session word-timing positive control failed. A replacement test of participant headset audio against close-up video found four eligible participants but only two visual-timing positives in one opened session; a second session had only two eligible participants. On HDTF clips (Zhang et al., 2021), a fixed audio-shift timing test passed for two of four development clips and neither of two validation clips. These outcomes stopped the respective source routes before renderer fitting or confirmation; they are endpoint-eligibility findings, not negative held-out renderer results.

On opened GRID development clips (Cooke et al., 2006), an AVA-style evaluator favored aligned over shifted input for two clips and reversed for two; an independent SyncNet checkpoint favored aligned input for one of four. A source-local fixed-crop predictor then passed its directional check for four of six validation talkers, below the preregistered five; its talker-bootstrap interval $[-0.01798, 0.06214]$ crossed zero. The reserved confirmation talkers remained unopened. These development checks do not supply a compatible external evaluator for a visible-speaker statement.