SELFOLEY: TURNING AUDIO-CONDITIONED VIDEO-TO-AUDIO MODELS INTO TEXT-GUIDED SOURCE SELECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-guided selective video-to-audio (V2A) asks a model to generate only a requested sound from a silent video that contains several plausible sources. Existing systems learn source selection and soundtrack generation together from large, automatically constructed data. We take a different route. Audio-conditioned V2A (AC-V2A) models normally transfer the content and acoustic character of a reference sound to a new video. In the open-weight ControlFoley model, we find that an isolated reference does more: it suppresses other visible events, and a controlled pathway intervention localizes this choice to an internal vector that represents *what sound* the reference contains. We call this the *source-content condition*. The **Clean Source Isolation Operator (CSIO)** is a 1.31M-parameter adapter that predicts this condition from target text while the generator remains frozen. CSIO alone improves selection but weakens audiovisual timing; **SelFoley** combines CSIO with early visual-synchronization guidance to recover alignment. We train the adapter on **StemFoleyDB-3.3K**, a dataset of 3,357 clips and 5,383 human-audited source stems. Its training split contains 3,287 clips (about 9.0 clip-hours); 3,274 clips with complete frozen features supply 300 adapter updates. Relative to SelVA's reported 179k-video, roughly 500-hour task corpus, SelFoley uses 98.2% fewer task-training clips and 96.8% fewer trainable parameters. On a fixed 50-clip test set, SelFoley reaches 78.76% target selection versus 51.33% for SelVA, while its Synchformer error is 0.384 versus 0.413. Eighteen listeners also rate non-target suppression at 3.30 versus 1.98 on a five-point scale. Separately trained CSIO adapters improve target selection by 7.4–23.6 points within three open-weight AC-V2A backbones, showing that generator-specific native reference-content interfaces can be reused instead of relearning separation end to end. Dataset examples and model outputs are available at `https://sel-foley.github.io/`.

1 INTRODUCTION

Video-to-audio (V2A) generation usually produces one soundtrack for an entire scene (Cheng et al., 2025). In practice, a user may want only one visible source: footsteps without speech, a creature call without its movement, or fire crackling without the surrounding sea. Text-guided selective V2A asks for exactly this behavior: given a silent video and a target caption, generate the requested source while suppressing other plausible sources in the scene.

SelVA first addressed this task by fine-tuning a V2A generator on 179k automatically constructed videos (about 500 hours) (Lee et al., 2026). Its training construction places two videos with one dominant sound side by side, mixes their audio, and uses text to request one of the two sounds. This creates an explicit two-way selection target, but differs from a natural multi-source scene in which target and competing events share the same camera view and context. On 20 such held-out scenes, 18 listeners rate SelVA's non-target suppression at 1.98 out of 5, compared with 3.30 for SelFoley and 4.26 for the human-audited target stem. Appendix Figure 4 contrasts these constructions, and representative outputs are available at `https://sel-foley.github.io/`.

We hypothesize that audio-reference transfer models already contain a useful form of source control. Audio-conditioned V2A (AC-V2A) generates a soundtrack for a new video while following a

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.

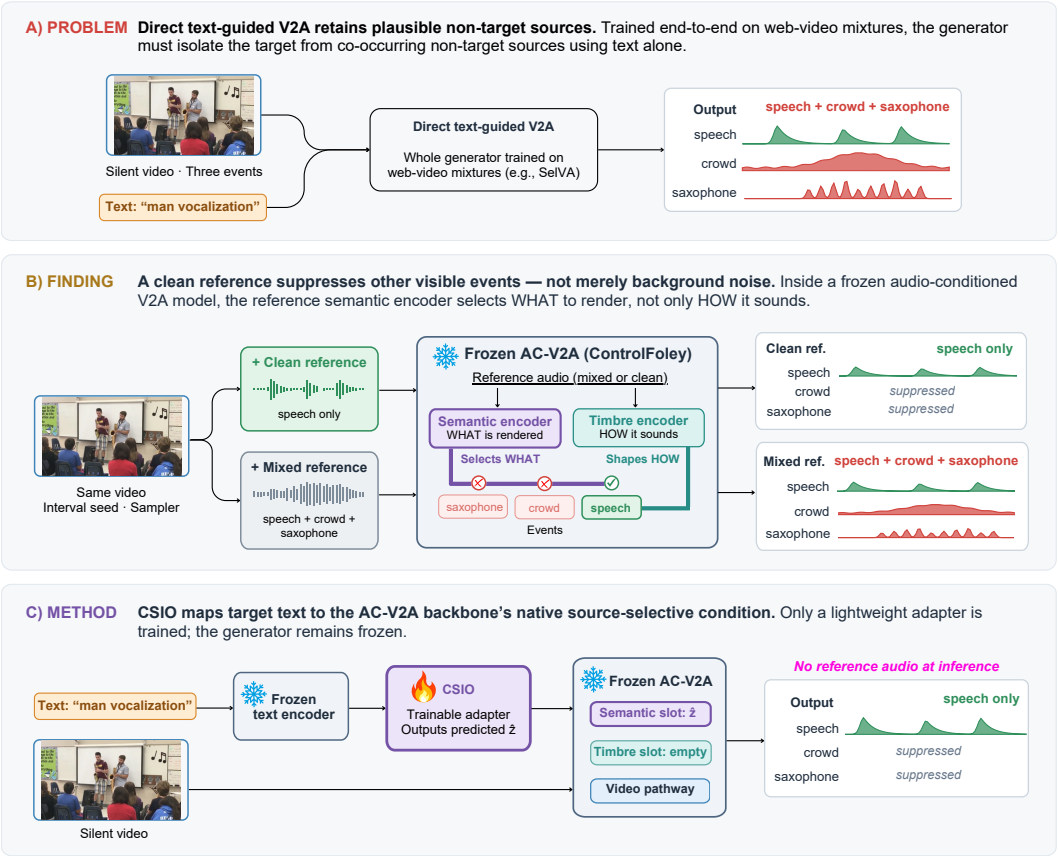


Figure 1: SelFoley in one view. Direct text-guided V2A must learn source separation from imperfect mixtures. We instead reuse a native reference-content interface in audio-conditioned V2A and train the Clean Source Isolation Operator (CSIO), a lightweight adapter that maps text to the generator's internal source-content condition. The frozen generator then renders the selected source in the video.

reference sound (Du et al., 2023; Chen et al., 2025; Fang et al., 2026; Yang et al., 2026). ControlFoley is an open-weight AC-V2A model with separate semantic and timbre encoders for its reference audio (Yang et al., 2026). Unexpectedly, giving it an isolated requested source—the *target stem*—does not merely change timbre: it suppresses sounds from other visible events. Controlled listening and encoder-swap experiments localize this selection to its semantic reference pathway.

This finding changes how we approach selective V2A. The **Clean Source Isolation Operator (CSIO)** is a small adapter that maps requested text to the source-content condition that a frozen AC-V2A generator already understands. Inference therefore needs only silent video and text. Because strong source-content conditioning can damage the backbone's timing, the complete **SelFoley** system combines CSIO with an early-sampling synchronization correction.

Training this interface requires reliable examples of each clean source. We therefore build **StemFoleyDB-3.3K**: 3,357 clips paired with 5,383 isolated audio tracks (stems) and captions. Automated tools propose candidates, but fifteen human curators decide whether each source is audible, isolatable, and correctly named. Multi-source clips are especially useful because targets and competing sources share the same video context.

Our training split contains 3,287 clips (about 9.0 clip-hours); 3,274 clips have every cached condition required to update the 1.31M-parameter adapter. Relative to SelVA, this is 98.2% fewer task-training clips and 96.8% fewer trainable parameters; the released V2A generator is never updated. The comparison concerns task adaptation, not the different backbones' original pretraining.

Language-queried separation isolates a described source from an existing mixture (Liu et al., 2022; 2023); applied after V2A, it inherits whatever the generator has already mixed, whereas CSIO selects

the source before waveform synthesis. Predicting a frozen generator's internal condition from text follows textual inversion and adapter-based control in image generation (Gal et al., 2023; Ye et al., 2023).

Our contributions are threefold:

- Across three open-weight AC-V2A backbones—ControlFoley, AC-Foley, and Video-Foley—we show that each model's native reference-content interface can be repurposed to improve selection of a requested visible source. This supports a latent source-selection capability across the evaluated AC-V2A models, without claiming that they share one identical mechanism.
- We analyze ControlFoley in depth, localize its source control to the semantic reference pathway, and develop CSIO to translate target text into that pathway's source-content condition. Synchronization-aware inference then turns the frozen backbone and CSIO into the complete SelFoley system.
- We introduce StemFoleyDB-3.3K, a human-audited dataset comprising 1,620 single-stem clips and 1,737 multi-stem clips. The multi-stem clips provide verified same-scene alternatives and enable our proposed TargetSel metric, whose underlying target-versus-strongest-non-target margin agrees with listener judgments in 86.9% of decisive comparisons.

2 From an audio reference to text-guided selection

2.1 ControlFoley's semantic reference pathway

We use ControlFoley (Yang et al., 2026) as the main generator because its released reference-audio pathway has two explicit parts, labeled *Semantic Encoder* and *Timbre Encoder* in Figure 1(b). The semantic encoder uses CLAP features to describe what the reference contains; the MusicGen-based timbre encoder summarizes how it sounds. We denote them by E_{sem} and E_{tim}. ControlFoley's native text condition is fixed to its learned empty value throughout. For a reference waveform R,

$$z^{\mathrm{sem}} = E_{\mathrm{sem}}(R), \qquad z^{\mathrm{tim}} = E_{\mathrm{tim}}(R). \tag{1}$$

We call z^{sem} the *source-content condition*: the controlled tests below show that it changes *which source* is rendered, whereas z^{tim} mainly changes *how that source sounds*. Throughout the paper, *semantic reference pathway* means the complete ControlFoley branch from reference waveform through E_{sem} to the generator; *source-content condition* means only its output z^{sem}. For other backbones, whose implementations differ, we use the generic term *native reference-content interface*. These are our descriptive terms, not terminology from the original ControlFoley paper. A separate visual pathway provides frame-level synchronization. Unless stated otherwise, all three pathways and ControlFoley's 2.82-billion-parameter generator remain frozen.

2.2 Does the semantic reference pathway select sources?

We test whether ControlFoley's semantic reference pathway carries a genuine source-isolation capability rather than merely changing the acoustic character of the output. We first localize the effect quantitatively, then check that listeners hear the same source-selection change. Section 5.2 subsequently analyzes what information the pathway represents.

Quantitative evaluation. We send either a clean target stem or its same-interval scene mixture independently to the semantic and timbre encoders, forming a 2×2 intervention with every other generation input fixed. We measure selection with TargetSel, defined in Section 4.3 and perceptually checked below. TargetSel is 60.18% with mixture audio in both encoders and 61.95% when only timbre is clean. It rises to 94.69% when only the semantic input is clean, matching the result when both inputs are clean. Thus changing the semantic input is sufficient for the observed selection gain, whereas changing timbre alone is not.

Listening evaluation. We also conduct a controlled A/B test in which only the source-content condition or only the timbre condition changes. Across the 30 intervention pairs, the TargetSel margin direction agrees with 86.9% of decisive listener judgments, and the timbre-only negative controls yield no clear difference in 168 of 180 judgments. The study therefore confirms that changing the

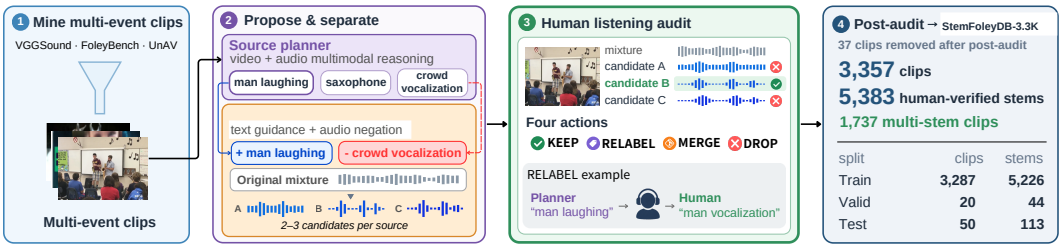


Figure 2: StemFoleyDB-3.3K construction. Automated prefiltering and multimodal reasoning propose sources; SAM-Audio separates candidates; human curators keep, relabel, merge, or drop them; and post-filtering removes silence, duplicates, and residual interference.

source-content condition changes audible source selection and that TargetSel tracks its perceived direction. Section 4.4 and Appendix D give the design and full statistics. Figure 3 and Section 5.2 analyze how the semantic reference pathway represents source identity.

2.3 Clean Source Isolation Operator

The **Clean Source Isolation Operator (CSIO)** predicts this source-content condition from the target caption. Let T be the caption, E_T the frozen M2D-CLAP text encoder (Niizumi et al., 2024), A_ϕ the 1.31-million-parameter adapter, and $z^{\text{tim}}_{\varnothing}$ ControlFoley's default empty timbre condition:

$$\hat{z}^{\text{sem}} = A_\phi(E_T(T)), \qquad \hat{X} = C\big(V, \hat{z}^{\text{sem}}, z^{\text{tim}}_{\varnothing}; \epsilon\big). \tag{2}$$

We do not also send the caption through ControlFoley's original text branch. Target stems supervise CSIO only during training; inference uses silent video and text, with no reference waveform.

CSIO learns to match the target stem's source-content condition, prefer it over the conditions of same-scene stems and the mixture, and preserve differences between source pairs. These objectives teach source identity rather than broad scene class. Appendix F gives their definitions, weights, and ablations. CSIO neither synthesizes an intermediate reference nor updates the generator.

2.4 From CSIO to SelFoley: Restoring Synchronization

CSIO exposes a trade-off: its strong source-content condition improves selection but can override the visual pathway that supplies timing. We diagnose when each signal acts by activating source-content and visual conditions over different portions of ControlFoley's 25 denoising steps. Source identity is established mainly early, and visual guidance is most effective there. SelFoley therefore uses

$$v = v_{\text{v2a}} + \beta\big(v_{\text{csio}} - v_{\text{v2a}}\big) + \gamma(t)\big(v_{\text{csio}} - v_{\text{nosync}}\big), \tag{3}$$

where v_{v2a} is the backbone's video-conditioned update, v_{csio} additionally includes the predicted source-content condition, and v_{nosync} removes frame-level synchronization. We use $\beta = 2.20$ throughout sampling and $\gamma(t) = 10.5$ only during the first quarter. This is the complete SelFoley system. Section 5.3 reports the selection–synchronization trade-off; Appendix F gives the search and timing diagnostics.

2.5 Other AC-V2A Backbones

CSIO is an interface-learning recipe, not one shared checkpoint. AC-Foley exposes one global reference-audio vector (Fang et al., 2026); Video-Foley separates a CLAP content condition from its Video2RMS timing module (Du et al., 2023). We train a separate adapter for each native reference-content interface while freezing the generator and leaving its visual sequence or timing module unchanged. Table 5 gives matched within-model comparisons; Appendix G gives implementation details.

3 StemFoleyDB-3.3K: human-audited source stems

CSIO needs the source-content condition of an isolated target stem and verified alternatives from the same scene. Existing web-video datasets rarely provide both. Figure 2 summarizes how we construct StemFoleyDB-3.3K.

We mine VGGSound (Chen et al., 2020), FoleyBench (Dixit et al., 2026), and UnAV (Geng et al., 2023) for clips likely to contain several events. A multimodal reasoner examines sampled frames and an audio visualization to propose source names, intervals, and separation prompts. SAM-Audio (Shi et al., 2025) then produces several candidates per source. These automatic stages propose candidates but never determine a released label or stem.

Fifteen undergraduate and graduate curators, including participants with prior audio-listening-evaluation experience, make every release decision. They listen to the mixture and candidates, then keep, relabel, merge, or drop each source. Captions describe what is audible in the retained stem rather than every visible attribute. Every released stem passes listening review and a final duplicate and leakage audit.

The resulting dataset contains 3,357 clips and 5,383 stems, including 1,737 multi-stem clips. Its train, validation, and test splits contain 3,287, 20, and 50 clips, with no original video shared across splits. Appendix B reports the prompt, representative proposals, full stage counts, human edits, preprocessing cost, and split statistics.

4 Experimental setup

4.1 Evaluation data

Our main quantitative evaluation uses the fixed StemFoleyDB-3.3K test split: 50 clips containing 113 requested sources. A disjoint 20-clip validation split is used for adapter development. ControlFoley generates eight seconds, so each test clip is assigned one common video window before any model is run. The default is 0–8 s; 13 clips containing a retained event that begins after 8 s use 2–10 s instead. Every target and system receives the same window. The 20 videos used for human ratings are independently curated UnAV clips excluded from all StemFoleyDB-3.3K splits.

We also evaluate SelVA and SelFoley on VGG-MonoAudio, SelVA's native VGGSound-derived benchmark (Lee et al., 2026). It reproduces SelVA's training construction: two clips curated to contain one dominant sound are placed side by side, their audio is mixed, and target text requests one sound. The clips and pairs are held out, but the input structure is matched to SelVA and differs from the natural multi-source scenes in StemFoleyDB-3.3K. VGG-MonoAudio's intra-class split pairs sources from the same broad category and is more confusable; its inter-class split pairs different categories. Appendix Figure 4 illustrates the constructions, and Appendices A and C.2 give the window rule and full metrics.

4.2 Compared systems

Table 1 compares official SelVA (Lee et al., 2026), default ControlFoley, SelFoley, a StemFoleyDB-trained SelVA LoRA, and a privileged target-stem-reference diagnostic. Table 4 isolates generator training, CSIO, and SelFoley's sampling rule. Checkpoints are selected on validation; Appendices A and F give implementation and search details.

4.3 Quantitative metrics

Target selection. Each test request provides a generated waveform $\hat{X}_i$, its human-audited target stem X_i^+, and verified non-target stems X_{ij}^- from the same video. A fixed pretrained audio encoder E converts each waveform into a vector. We compute the output's similarity to the target and subtract its largest similarity to any non-target:

$$m_i = \cos\Big(E(\hat{X}_i), E(X_i^+)\Big) - \max_j \cos\Big(E(\hat{X}_i), E(X_{ij}^-)\Big), \quad \text{TargetSel} = \frac{1}{N}\sum_i \mathbf{1}[m_i > 0]. \tag{4}$$

Table 1: Text-guided selective V2A on the fixed StemFoleyDB-3.3K test set (50 clips/113 sources). TargetSel is computed independently with frozen M2D-CLAP and LAION-CLAP audio encoders; neither evaluator is updated with CSIO. CSIO uses only the predicted source-content condition without SelFoley's synchronization repair. The target-stem row requires target audio and is a privileged diagnostic, not a deployable system. Bold and underline mark the best and second-best deployable results. This table reports the paper's fixed generation seed; Table 5 averages three matched seeds.

Method	TargetSel M2D ↑	TargetSel LAION ↑	DeSync ↓	FD-VGGish ↓	FD-PANN ↓	KL-PANN ↓	IS-PANN ↑
SelVA (Lee et al., 2026)	51.33	47.79	<u>.413</u>	25.65	67.28	2.868	**4.972**
SelVA + StemFoleyDB-3.3K LoRA	76.11	<u>74.34</u>	1.180	6.14	39.01	2.457	<u>4.846</u>
ControlFoley (default)	58.41	60.18	.630	13.52	47.16	2.635	3.746
ControlFoley + CSIO (no sync repair)	**82.30**	**79.65**	.779	<u>5.30</u>	<u>35.31</u>	<u>1.943</u>	4.559
SelFoley (ours)	<u>78.76</u>	**79.65**	**.384**	**4.71**	**33.73**	**1.860**	4.485
ControlFoley + target-stem reference	94.69	93.81	.882	3.89	26.83	1.128	4.970

Thus $m_i > 0$ means the output resembles the target more than its strongest same-scene alternative; TargetSel is the percentage of requests satisfying this test. Our primary evaluator is the frozen audio encoder of M2D-CLAP (Niizumi et al., 2024). Although CSIO receives text features from M2D-CLAP, the evaluator is never trained with CSIO and never compares generated audio with text: it ranks generated audio against held-out target and non-target stems. Shared M2D pretraining could still favor M2D-aligned outputs, so we repeat the entire audio-to-audio ranking with independently trained LAION-CLAP (Wu et al., 2023), which is never used by CSIO. Table 1 reports both and preserves the ordering. A CSIO adapter trained on LAION-CLAP text features reaches 81.42% TargetSel (Appendix Table 21), so the gain is not tied to M2D text features. Sections 2.2 and 4.4 provide perceptual validation. TargetSel measures relative source identity, not complete removal of residual sound.

Synchronization and audio distribution. Synchformer DeSync measures broad audio–video alignment, not target-onset error (Iashin et al., 2024). We compare generated and clean-target distributions using Fréchet distance (FD) in VGGish (Hershey et al., 2017) and PANN (Kong et al., 2020) spaces (Heusel et al., 2017; Kilgour et al., 2019), plus PANN KL divergence (Kullback & Leibler, 1951) and Inception Score (Salimans et al., 2016). Lower is better for DeSync, FD, and KL; higher is better for IS. These are distribution rather than perceptual-quality measures. Confidence intervals resample videos because one clip can contain several targets.

4.4 Human evaluation

Eighteen listeners complete two blinded evaluations. First, on 20 independently curated videos outside every StemFoleyDB split, they rate SelVA, SelFoley, and the human-audited target stem for non-target suppression, caption match, and synchronization.

The second evaluation tests the semantic-reference-pathway finding and TargetSel. Each of 30 A/B pairs keeps the video, initial noise, sampler, and other generation settings fixed. Ten pairs replace the target source-content condition with that of another verified same-scene source, ten replace the clean target input to the semantic encoder with the scene mixture, and ten change only timbre as a negative control. Listeners select which output makes the requested source clearer, or report no clear difference. We compare decisive choices with the M2D TargetSel-margin direction. Section 2.2 summarizes the result; Appendix D gives both protocols, screening rules, and tie analysis.

5 Results

5.1 Source selection, synchronization, and listener judgments

Under the M2D-CLAP evaluator defined in Section 4.3, SelFoley selects the requested source in 78.76% of the 113 cases: 27.43 points above official SelVA (51.33%; video-bootstrap 95% CI $[20.18, 34.91]$) and 20.35 above default ControlFoley (58.41%). CSIO alone reaches 82.30% but weakens synchronization; SelFoley gives up 3.54 points to reduce DeSync from 0.779 to 0.384. A

Table 2: Blinded human evaluation by 18 listeners. (A) 1–5 MOS on 20 independent targets. Bold marks the better of the two generated systems; the human-curated stem is an unranked reference condition. (B) Perceptual validation of TargetSel: on 30 controlled interventions, we compare which output listeners judge more source-selective with the direction of the M2D target-minus-strongest-non-target margin. Timbre-only pairs are negative controls. Appendix D gives the protocol and tie analysis.

A. Human ratings of selective generation

System	Non-target suppression	Caption match	Audio–video sync
SelVA	1.98	**3.37**	**3.40**
SelFoley	**3.30**	3.20	2.94
Human-audited target stem (reference)	4.26	4.13	4.03

B. Human validation of the target-selection margin

Evaluator	Response agree. (%)	Item agree. (%)	Spearman ρ	Tie (%)
M2D TargetSel margin	86.9	88.5	0.908	49.3

Table 3: VGG-MonoAudio evaluation in SelVA's benchmark domain. Source Top-1 uses the independent MS-CLAP evaluator (Elizalde et al., 2023). Full alignment and distribution metrics are in Appendix C.2.

Split	Method	Source Top-1 ↑	DeSync ↓	FD-PANN ↓	KL-PANN ↓
Inter	SelVA	78.04	**.801**	**79.45**	**2.81**
	SelFoley	**85.00**	.856	82.60	3.28
Intra	SelVA	**72.41**	.760	66.15	**1.83**
	SelFoley	72.02	**.673**	**64.15**	2.24

StemFoleyDB-trained SelVA LoRA reaches 76.11%, but its DeSync of 1.180 indicates severe timing degradation from fine-tuning the generator; SelFoley keeps the generator frozen. The independent LAION-CLAP evaluation in Table 1 gives the same ordering for the three principal systems: 79.65% for SelFoley, 74.34% for the SelVA LoRA, and 47.79% for official SelVA. Consistent with stronger isolation, SelFoley's outputs lie closer to the target stem than to the scene mixture (FD-VGGish 4.71 vs. 10.28), whereas SelVA's lie closer to the mixture (25.65 vs. 11.47; Appendix Table 13). Generating a full soundtrack with Foley-Omni (Tao et al., 2026) and then separating the captioned source with SAM-Audio reaches only 46.02% TargetSel (Appendix Table 9), consistent with the cost of separating after generation.

SelFoley also improves clean-target distribution distances and aggregate Synchformer alignment over SelVA, although these automatic measures are not perceptual-quality judgments. The 18-listener study confirms the main claim: non-target suppression rises from 1.98 to 3.30 MOS (difference $+1.32$, 95% CI $[0.69, 1.91]$), while caption match is statistically indistinguishable. Listeners prefer SelVA's synchronization (3.40 vs. 2.94), revealing a target-specific timing limitation that aggregate DeSync does not fully capture.

VGG-MonoAudio is an in-distribution check for SelVA rather than a replacement for our natural-scene test. Like SelVA training, it places two clips with one dominant sound side by side and asks for one by text; the held-out clips differ, but the construction is the same (Appendix Figure 4). On its harder intra-class split, SelFoley and SelVA are tied in source Top-1 (72.02% vs. 72.41%), while SelFoley has lower DeSync (0.673 vs. 0.760). On the inter-class split, SelFoley improves source Top-1 from 78.04% to 85.00%. SelVA retains better distribution scores on this native benchmark; Appendix C.2 reports the complete results.

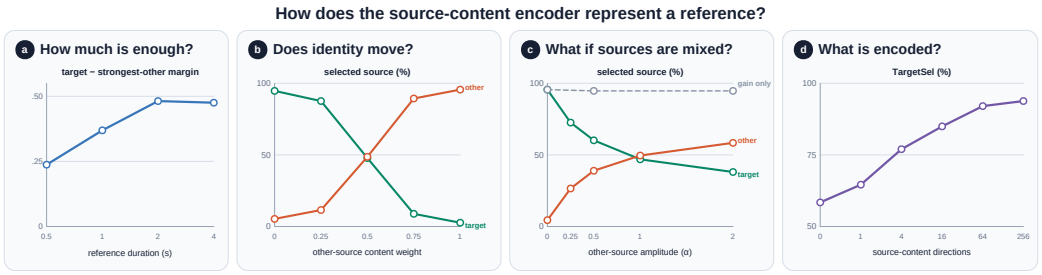


Figure 3: **How frozen ControlFoley's semantic encoder represents a reference.** (a) Target-minus-strongest-other representation margin as clean-reference duration increases. (b) Moving that condition toward another same-scene source switches the output. (c) Adding RMS-matched competing audio changes selection, whereas changing target gain alone does not. (d) Target selection recovered by an increasing number of train-derived source-content directions.

Table 4: Interface learning and synchronization repair on the fixed test set. ControlFoley full fine-tuning updates the 2.82B-parameter generator; CSIO and SelFoley train only the same 1.31M-parameter adapter. SelFoley differs from CSIO only through inference-time visual synchronization guidance. All rows use the same generation seed as Table 1.

Method	Params.	TargetSel M2D	TargetSel LAION	Margin	DeSync	FD-VGGish
ControlFoley (default)	0	58.41	60.18	.033	.630	13.52
ControlFoley (full fine-tune)	2.82B	63.72	58.41	.036	.608	12.98
CSIO adapter (no sync repair)	1.31M	**82.30**	**79.65**	**.116**	.779	5.30
SelFoley	**1.31M**	78.76	**79.65**	.111	**.384**	**4.71**

5.2 How does the semantic encoder represent a reference?

Figure 3 asks what information makes the semantic reference pathway selective. First, the source-content condition becomes more source-specific as the encoder observes more evidence: extending an RMS-matched clean reference from 0.5 to 2 s raises the target-minus-strongest-other representation margin from .237 to .482, after which 4 s provides little additional gain. Across different crop lengths, the same source is retrieved in 96.24% of off-diagonal comparisons. The encoder therefore aggregates source identity over time rather than depending on one brief acoustic cue.

Second, the source-content condition controls *which* source is generated. We continuously interpolate it from a target source toward another source visible in the same video while keeping video, timbre, sampling noise, and inference settings fixed. Target selection falls monotonically from 94.7% to 2.7%, crossing the decision boundary near the interpolation midpoint as selection of the substituted source rises. This dose–response is stronger evidence than comparing only two endpoints: the generated identity follows the source-content condition itself.

Third, the pathway responds to competing content rather than loudness. Adding an RMS-matched same-scene source to the reference moves both the source-content condition and generated output toward that source, while a gain-only control leaves target selection near 95%. Finally, one universal clean-versus-mixed direction is insufficient: 16 training-derived source-content directions recover 85.0% TargetSel and 64 recover 92.0%, approaching the full 94.7%. The semantic encoder therefore represents a structured, multidimensional source identity, which explains why CSIO predicts a source-dependent condition instead of adding one generic "clean" vector. Appendix E reports the complete sweeps, confidence intervals, representation geometry, and sampling-time diagnostics.

5.3 From CSIO to SelFoley: visual synchronization repair

Table 4 fixes the backbone, split, and test protocol. Updating all 2.82B generator parameters moves TargetSel by only a few points, and in opposite directions under the two evaluators (M2D 58.41% to 63.72%, LAION 60.18% to 58.41%), whereas the 1.31M-parameter CSIO raises it to 82.30% and

Table 5: Backbone-specific CSIO on the same 50 clips/113 sources, averaged over three seeds. M2D-CLAP and LAION-CLAP repeat the TargetSel comparison; arrows are paired default→CSIO changes within each backbone, not rankings across architectures.

	TargetSel via M2D (%)		TargetSel via LAION (%)			
Backbone	Default	CSIO	Default	CSIO	FD-PANN ↓	DeSync ↓
ControlFoley	58.70	**82.30**	57.82	**79.65**	53.64→36.72	.676→.776
AC-Foley	53.98	**61.36**	53.39	**62.24**	72.26→51.32	.362→.464
Video-Foley	73.45	**81.71**	67.55	**78.76**	41.67→37.84	1.214→1.173

79.65%. CSIO also raises DeSync from 0.630 to 0.779 by overriding visual timing. Because source identity and useful visual guidance both act early, SelFoley strengthens visual guidance only during the first sampling quarter. This lowers DeSync to 0.384 and FD-VGGish to 4.71 while retaining 78.76% TargetSel. Appendices E.4 and F report the intervention and full search.

5.4 Transfer to other AC-V2A backbones

Table 5 evaluates a separate adapter for each frozen backbone. Across three seeds, M2D-CLAP TargetSel improves by 23.60 points for ControlFoley, 7.37 for AC-Foley, and 8.26 for Video-Foley; LAION-CLAP confirms every direction, and FD-PANN falls on all three backbones. Without inference-time guidance, DeSync rises for ControlFoley and AC-Foley; calibrated guidance restores AC-Foley's DeSync to 0.378 at 58.70% TargetSel. This supports a portable interface-learning recipe, not a shared checkpoint or mechanism. Appendix G gives full results.

6 Limitations

SelFoley controls source identity more reliably than event timing. Stronger early visual guidance improves aggregate synchronization but does not provide explicit event boundaries; attempted temporal masks neither relocate events nor preserve selection reliably. SelFoley's sampling rule also evaluates three flow branches at every denoising step, so inference costs more than the frozen backbone; we have not timed it.

Our mechanism claim is specific to ControlFoley, whose content and timbre inputs can be changed independently. AC-Foley and Video-Foley support only the narrower result that separately trained adapters can predict their model-specific conditions; they do not establish one shared mechanism or checkpoint.

TargetSel measures embedding-based source dominance, not complete isolation. We therefore pair it with LAION-CLAP, distribution metrics, and listener judgments. Listening validates its direction on screened interventions, not universal calibration, and 18 listeners still limit fine-grained conclusions. The 113 test sources also cluster in 50 clips and are human-audited separated stems rather than studio ground truth; small differences may reflect separator or clip-level uncertainty.

7 Conclusion

Text-guided selective V2A need not teach a generator to separate every source from scratch. In ControlFoley, an isolated target reference exposes a semantic reference pathway that changes which visible source is generated. CSIO predicts its source-content condition from text; SelFoley reinforces the frozen backbone's visual signal during early sampling; and StemFoleyDB-3.3K supplies human-audited targets and same-scene alternatives. Separate adapters for the native reference-content interfaces of two additional frozen AC-V2A backbones support the portability of this interface-learning formulation. Target-specific timing and broader architectural coverage remain the clearest next steps.

ETHICS STATEMENT

StemFoleyDB-3.3K is derived from existing research video corpora and is intended for academic research on controllable audio generation. Before release, we will document source licenses, remove identifying speech where required by the source terms, and publish only media for which redistribution is permitted. Listening-study participants provide informed consent. Names and email addresses are collected only for compensation and duplicate-response control, stored separately from ratings, and deleted after compensation; only de-identified aggregate results are reported. Selective audio generation can support accessibility and media editing, but can also facilitate misleading audiovisual edits; released models and demos will therefore include provenance and use restrictions inherited from their backbones.

REPRODUCIBILITY STATEMENT

Appendix A specifies the immutable split hashes, preprocessing contract, losses, inference settings, and statistical tests. Upon acceptance, we will release the SelFoley checkpoint, evaluation code, and StemFoleyDB-3.3K under the applicable source licenses.

AI USE STATEMENT

Generative AI tools assisted with code inspection, experiment bookkeeping, and language editing. All scientific claims, data decisions, experimental configurations, and reported numbers were checked against stored artifacts by the authors. Reported human-evaluation results are based exclusively on ratings collected from human listeners; generative AI neither supplied nor imputed these judgments.

REFERENCES

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*, 2020.

Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18781, 2025.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28901–28911, 2025. doi: 10.1109/CVPR52734.2025.02691.

Satvik Dixit, Koichi Saito, Zhi Zhong, Yuki Mitsufuji, and Chris Donahue. Foleybench: A benchmark for video-to-audio models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 14512–14516, 2026. doi: 10.1109/ICASSP55912.2026.11461920.

Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Pengjun Fang, Yingqing He, Yazhou Xing, Qifeng Chen, Ser-Nam Lim, and Harry Yang. Ac-foley: Reference-audio-guided video-to-audio synthesis with acoustic transfer. In *International Conference on Learning Representations*, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131–135, 2017. doi: 10.1109/ICASSP.2017.7952132.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Proceedings of Interspeech*, 2022.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Junwon Lee, Juhan Nam, and Jiyoung Lee. Hear what matters! text-conditioned selective video-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36680–36690, 2026.

Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Separate what you describe: Language-queried audio source separation. In *Proc. Interspeech*, pp. 1801–1805, 2022.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. Separate anything you describe. *arXiv preprint arXiv:2308.05037*, 2023.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Masahiro Yasuda, Shunsuke Tsubaki, and Keisuke Imoto. M2D-CLAP: Masked modeling duo meets CLAP for learning general-purpose audio-language representation. In *Proceedings of Interspeech*, pp. 57–61, 2024. doi: 10.21437/Interspeech.2024-29.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

Ye Tao, Lupeng Liu, Xuenan Xu, Jiasun Feng, Jiarui Wang, Ying Qin, Shuiyang Mao, Wei Liu, and Shuai Wang. Foley-omni: A unified multimodal generation model from task-level audio synthesis to complete video soundtrack generation. *arXiv preprint arXiv:2606.03672*, 2026.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP*, 2023.

Jianxuan Yang, Xinyue Guo, Zhi Cheng, Kai Wang, Lipan Zhang, Jinjie Hu, Qiang Ji, Yihua Cao, Yihao Meng, Zhaoyue Cui, Mengmei Liu, Meng Meng, and Jian Luan. Controlfoley: Unified and controllable video-to-audio generation with cross-modal conflict handling. *arXiv preprint arXiv:2604.15086*, 2026.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

HaeJun Yoo, Yongseop Shin, Insung Lee, Myoung-Wan Koo, and Du-Seong Chang. Omni-embed-audio: Leveraging multimodal LLMs for robust audio-text retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

A Reproducibility details

The appendix follows the paper's evidence flow. Appendix A fixes the reproducibility contract; Appendix B documents StemFoleyDB-3.3K; Appendix C expands the benchmark tables; Appendix D gives the listening protocols behind the human results; Appendices E and F report semantic-reference-pathway analyses and CSIO development; and Appendix G covers the other AC-V2A backbones.

A.1 Data splits and evaluation windows

The release uses seed 20260918. SHA-256 hashes are train `60f977a1...96e642`, validation `3d5eb55d...98bd79c`, and test `6ee1f334...13306d`. The splits contain 3,287/20/50 clips and 5,226/44/113 stems, with zero parent-video overlap.

Every model produces an 8-s soundtrack from the same preassigned 8-s video window. Clips use 0–8 s unless a retained event begins after 8 s, in which case they use 2–10 s. This gives 37 clips (83 sources) with the first window and 13 clips (30 sources) with the second. The crop is fixed per clip, never per caption or system; models receive neither event boundaries nor activity masks. Metric references are cropped to the identical window.

A.2 Adapter training and task resources

The frozen M2D-CLAP audio–text model supplies the caption representation. It is passed through layer normalization, a 1,024-unit linear layer, a Gaussian Error Linear Unit (GELU) activation, dropout 0.1, and a 512-unit output layer followed by ℓ_2 normalization. The adapter has 1.31 million parameters. Following the main text, E_{sem} denotes ControlFoley's released CLAP-based Semantic Encoder. Let $z_i^m = E_{\text{sem}}(R_i^m)$ and $z_i^+ = E_{\text{sem}}(R_i^c)$ denote the source-content conditions from the scene mixture and the human-audited target stem. For ordered same-video source pairs (i, j), the two auxiliary losses used by the final adapter are

$$\mathcal{L}_{\text{iso}} = \mathbb{E}_i \left[\max(0, m + \cos(\hat{z}_i, z_i^m) - \cos(\hat{z}_i, z_i^+)) \right], \tag{5}$$

$$\mathcal{L}_{\Delta} = \mathbb{E}_{i,j} \left[1 - \cos(\hat{z}_i - \hat{z}_j, z_i^+ - z_j^+) \right]. \tag{6}$$

The pointwise and same-scene ranking terms are

$$\mathcal{L}_{\text{fit}} = 1 - \cos(\hat{z}_i, z_i^+), \tag{7}$$

$$\mathcal{L}_{\text{rank}} = \max\left(0, m - \cos(\hat{z}_i, z_i^+) + \max_{j \neq i} \cos(\hat{z}_i, z_j^+)\right). \tag{8}$$

The complete objective is $\mathcal{L}_{\text{fit}} + \lambda_r \mathcal{L}_{\text{rank}} + \lambda_i \mathcal{L}_{\text{iso}} + \lambda_\Delta \mathcal{L}_\Delta$. We set $m = 0.15$ and $\lambda_r = \lambda_i = \lambda_\Delta = 0.5$. Training uses AdamW for 300 epochs, with all training embeddings included in one batch per

Table 6: Task-adaptation resources for selective V2A. Released backbone pretraining is excluded for both methods. SelVA does not report wall-clock training time, so we compare documented data, optimization, and trainable parameters rather than estimating a speedup.

	SelVA	SelFoley (CSIO)
Released task-training data	179k videos (∼500 h)	3,287 clips / 5,226 stems
Cached encoder features	not applicable	3,274 clips / 5,191 stems
Trainable parameters	41.0M	**1.31M**
Updated V2A-generator parameters	22.0M	**0**
Optimization	50k + 25k steps	300 full-batch steps
Selected checkpoint	final two-stage model	epoch 75 (validation)
Reported training hardware	1×RTX 4090; 1×A6000	1×A100-SXM4-80GB
Adapter optimization time	not reported	13.0 s[†]
Frozen V2A backbone	No	**Yes**
Sampling	25-step Euler	25-step Euler
Inference input	silent video + text	silent video + text

SelVA trains 19M video-encoder parameters in stage 1 and 22M generator parameters in stage 2. The reported CSIO checkpoint uses cached encoder features after omitting 35 stems from 13 clips that lack one or more required embeddings. [†]Measured for 300 updates from cached features; frozen feature extraction is excluded. Full deployed parameter totals are not compared because the two systems use different frozen encoders and decoders.

epoch, learning rate 2×10^{-4}, weight decay 0.01, gradient clipping at 5.0, and seed 20260920. The adapter checkpoint is selected using validation data.

Table 6 compares the documented task-specific resources. SelVA reports 179k videos drawn from a roughly 500-hour corpus, 41M task-trained parameters, and two stages totaling 75k updates (Lee et al., 2026). StemFoleyDB-3.3K provides 3,287 training clips (9.03 clip-hours); 3,274 clips with complete cached conditions supply CSIO's 300-update, 1.31M-parameter training run. We do not claim a wall-clock speedup because SelVA does not report comparable end-to-end training time and CSIO's timing excludes frozen feature extraction.

A.3 Inference configuration

At inference, the target caption is consumed only by the frozen text encoder and CSIO adapter. ControlFoley's default text and timbre-reference conditions are left empty; the generator receives the silent video and CSIO's predicted source-content condition. The human-audited target stem is used only to construct the training target and post-generation evaluation reference. No reference waveform or target timing is available to the deployed system.

For the non-deployable target-stem reference diagnostics, the semantic encoder receives the audible crop at its original duration. The timbre encoder receives the same crop, padded only to its minimum supported duration and capped at 4 s. We do not append silence to an 8-s semantic input, tile the crop, or peak-normalize it. SelFoley uses 25 Euler steps, classifier-free guidance 1.0, source-content coefficient $\beta = 2.20$, and synchronization coefficient 10.5 during the first quarter of sampling. Paired ControlFoley interventions reuse the same initial noise. Comparisons involving a different released backbone retain that model's released preprocessing and inference settings unless stated otherwise.

For completeness, let v_{v2a} be ControlFoley's released video-conditioned denoising update, v_{csio} the update after adding the predicted source-content condition, and v_{nosync} the same CSIO run with frame-level synchronization features removed. SelFoley uses

$$v = v_{\mathrm{v2a}} + \beta(v_{\mathrm{csio}} - v_{\mathrm{v2a}}) + \gamma(t)(v_{\mathrm{csio}} - v_{\mathrm{nosync}}), \tag{9}$$

where $\gamma(t) = 10.5$ for the first quarter of sampling and zero afterward.

Table 7: StemFoleyDB-3.3K split statistics. Validation and test contain only multi-source clips; all splits are parent-video disjoint.

Split	Clips	Stems	Parent videos	Single/Multi clips
Train	3,287	5,226	3,146	1,620 / 1,667
Validation	20	44	20	0 / 20
Test	50	113	50	0 / 50
Total	3,357	5,383	3,216	1,620 / 1,737

B StemFoleyDB-3.3K curation details

B.1 Candidate proposal and separation

The automatic proposal step receives a short video, its original mixed soundtrack, and weak dataset metadata. We use the archived GPT-5.6 Sol model at high reasoning effort. It receives a 2-fps contact sheet, a shared-axis waveform and log-spectrogram, clip-level root-mean-square energy, strong onset times, and low/mid/high spectral-energy ratios. Researchers refined these inputs and the instruction while listening to proposals until the model produced a useful high-recall list of possible audible sources. In a separate audit of 179 non-speech UnAV clips containing 273 annotated event families, GPT-5.6 Sol with frames and the waveform–spectrogram image achieved 0.608 event-family recall and 0.706 union temporal intersection-over-union; frames alone achieved 0.520 and 0.684. Event-family recall does not penalize extra proposals because UnAV annotations are not exhaustive, and temporal overlap is computed only for semantically recovered families. These numbers therefore measure high-recall proposal coverage and localization, not final-label accuracy. Human listening determines every released source decision. Prompts and intermediate outputs will be archived with the data release, while representative final-stem quality can be inspected on the anonymous project page at `https://sel-foley.github.io/`.

The fixed instruction to the multimodal model was:

> "Analyze the attached timestamped video contact sheet and mixed-audio waveform/spectrogram. The dataset caption is a weak prior, not ground truth. Propose a variable-count, high-recall inventory of causal sources for subsequent SAM-Audio extraction. Every candidate must be jointly supported by visible objects/actions and acoustic timing. Give precise activity intervals. Use onset for compact events, span for sustained actions, and ambience only for a genuine continuous bed. Do not split one action into speculative timbral fragments, invent speech or music, or force a count. Prefer short concrete labels. Return schema-valid JSON only."

Per-clip metadata and the required output fields—label, description, role, activity intervals, visual evidence, acoustic evidence, and confidence—were appended to this fixed instruction.

Two archived examples illustrate the intended granularity and the role of human review. For `unav_-9Y076KDt2w_w01`, continuous guitar motion and onset-dense harmonic energy led the model to propose *Acoustic guitar* and *Singing voice* over 0–10 s; a curator listened to the separated candidates and retained both captions. For `unav_5FJOEsNJVkk_w01`, a scene cut led it to propose *Audience applause* (0–5.72 s) and *Male podium speech* (5.72–10 s); the final human-curated record retains only *Audience applause*. Automatic suggestions are candidates for human review, not released labels.

For each proposal, SAM-Audio receives three prompt forms—the label, label plus description, and a normalized label—and produces up to eight attempts per prompt. Difficult cases additionally use an accepted same-scene source as a negative audio example—a sound SAM-Audio should suppress—and curators inspect the residual audio. Automatic signal checks rank the separated candidates and remove decoding failures and near-silence, but do not turn a proposal into ground truth.

Recorded cost. Archived billing records cover automatic filtering and proposal calls, not the full curation process. The initial 20,000-clip VGGSound/FoleyBench pre-filter records $12.22; FoleyBench candidate filtering and refinement records $41.70; and UnAV candidate filtering records

Table 8: Recorded stages of StemFoleyDB-3.3K construction. Intermediate review records overlap because uncertain clips can be checked again; each row therefore retains its own decision unit and denominator rather than reporting a misleading pooled rejection rate.

Stage	Input	Retained/revised	Rejected	Unit
Initial automatic filtering	20,000	5,553	14,447	clips
Source-level human review	1,983	979 keep + 453 relabel	551	sources
Focused review of uncertain cases	508	262 keep + 74 relabel	172	decisions
Final duplicate/leakage removal	5,420	5,383	37	stems

$194.77, for $248.69 of documented calls. This excludes local preprocessing, SAM-Audio GPU inference, human labor, and development pilots. Later multimodal source-proposal calls preserved prompts, outputs, and token totals but not provider billable input/output splits, so we do not infer a dollar total for them.

B.2 Human curation and final data

Fifteen undergraduate and graduate participants were recruited through the university for the listening-based curation. Several had prior experience in recurring audio-evaluation studies. Each curator received approximately KRW 30,000. Curators heard the original mixture and separated candidates and assigned one of the four actions used in Figure 2: keep, relabel, merge, or drop. A retained caption describes audible acoustic content; unsupported visual, spatial, ordinal, or temporal attributes are removed.

Intermediate review records overlap because later reviews deliberately revisit uncertain clips; Table 8 therefore retains the proper denominator for each stage rather than pooling decisions into one rejection rate. In the 1,983-source human review, rejected proposals were unrelated or had the wrong identity (36.5%), were effectively silent (26.1%), duplicated an acoustically indistinguishable source (18.5%), contained separation artifacts (13.8%), or failed for other reasons (5.1%). Among 442 retained edits with complete before-and-after records, 290 changed source identity, 99 simplified the label to acoustically verifiable content, 21 removed visual attributes, seven removed unsupported temporal or ordinal wording, and 25 clarified an underspecified label. Explicit relabeling was most common for animal and creature, object and material, vehicle and machine, body movement and impact, and ambience, weather, and water sounds. Because complete edit records are concentrated in FoleyBench and VGGSound, these are observed edit patterns rather than population error rates.

The 20,000-to-5,553 row describes the initial VGGSound/FoleyBench candidate cohort only. The final 3,357-clip release also includes separately prioritized UnAV cohorts, so 3,357/20,000 is not a valid end-to-end retention rate. The recorded figures are therefore kept separate: the initial VGGSound/FoleyBench pre-filter retained 27.8% of its own cohort, and the union of all human-curated cohorts contains 3,357 released clips and 5,383 stems.

For each accepted stem, frame-wise energy locates an audible 2–4 s reference crop. We do not amplify near-silence through peak normalization. To avoid recomputing frozen encoders, we cache each training source's caption, target-stem, mixture, and same-scene-source embeddings. The reported adapter uses 5,191 training sources from 3,274 clips; 35 stems missing at least one required embedding are omitted. Multi-source supervision is used only when at least two cached sources remain in a clip.

C Complete benchmark results

C.1 StemFoleyDB-3.3K

Tables 9 and 10 expand the main comparison with target similarity, target-minus-strongest-other margin, and Patchout faSt Spectrogram Transformer (PaSST) distribution metrics (Koutini et al., 2022). They also include Foley-Omni (Tao et al., 2026) followed by SAM-Audio separation (Shi et al., 2025). Every row uses the same 50 clips, 113 source requests, and fixed windows.

Table 9: Full source-selection and synchronization results on the fixed StemFoleyDB-3.3K test set. Target and margin denote target-stem cosine similarity and target-minus-strongest-same-scene-non-target similarity. The Foley-Omni + SAM-Audio baseline first generates a complete scene soundtrack, then separates the captioned source. ControlFoley (default) uses video and text without an audio reference. The final row conditions ControlFoley on target audio and is not deployable.

Method	TargetSel (%) ↑	Target ↑	Margin ↑	DeSync ↓
SelVA	51.33	0.677	0.011	0.413
SelVA + StemFoleyDB-3.3K LoRA	76.11	0.712	0.093	1.180
Foley-Omni + SAM-Audio	46.02	0.529	−0.011	1.321
ControlFoley (default)	58.41	0.675	0.033	0.630
ControlFoley (full fine-tune)	63.72	0.678	0.036	0.608
SelFoley (ours)	**78.76**	**0.732**	**0.111**	**0.384**
ControlFoley + target-stem reference	94.69	0.820	0.211	0.882

Table 10: Full audio-distribution metrics on the same fixed StemFoleyDB-3.3K test outputs. PANN and Patchout faSt Spectrogram Transformer (PaSST) are independent audio-classifier feature spaces. FD and KL are lower-is-better distances; IS is a higher-is-better Inception Score. The final row conditions ControlFoley on target audio and is not deployable. Bold and underline mark the best and second-best deployable results.

Method	FD-VGGish ↓	FD-PANN ↓	FD-PaSST ↓	KL-PANN ↓	KL-PaSST ↓	IS-PANN ↑	IS-PaSST ↑
SelVA	25.65	67.28	614.39	2.868	2.362	**4.972**	**4.284**
SelVA + StemFoleyDB-3.3K LoRA	<u>6.14</u>	<u>39.01</u>	<u>554.16</u>	<u>2.457</u>	<u>2.006</u>	<u>4.846</u>	<u>4.102</u>
Foley-Omni + SAM-Audio	16.84	88.76	1225.40	3.805	5.555	1.124	1.035
ControlFoley (default)	13.52	47.16	669.50	2.635	2.163	3.746	3.198
ControlFoley (full fine-tune)	12.98	46.98	662.57	2.628	2.145	3.860	3.214
SelFoley (ours)	**4.71**	**33.73**	**473.72**	**1.860**	**1.666**	4.485	4.002
ControlFoley + target-stem reference	3.89	26.83	390.82	1.128	1.059	4.970	4.336

StemFoleyDB-adapted SelVA. This baseline starts from official SelVA and trains rank-8 low-rank adaptation modules on its selector and generator for 1,000 updates. It uses the StemFoleyDB training split plus same-scene ranking and output-separation losses. It is distinct from both official SelVA and full-parameter training.

C.2 VGG-MonoAudio

Figure 4 makes the two benchmark constructions explicit. SelVA is trained by horizontally concatenating two VGGSound videos, mixing their soundtracks, and using target text to request one source (Lee et al., 2026). VGG-MonoAudio applies the same construction rule to held-out clips curated to contain one dominant source. It is therefore SelVA's native, structurally matched benchmark, although it does not reuse the same training clips or pairs. StemFoleyDB-3.3K asks a different question: its target and alternatives co-occur in one natural scene and are represented by human-audited stems.

Because the constructions differ, we report VGG-MonoAudio target-source Top-1 separately from StemFoleyDB-3.3K TargetSel. The intra-class split pairs sources from the same broad category and is more confusable; the inter-class split pairs different broad categories. ImageBind audio–video similarity is included as a broad alignment diagnostic (Girdhar et al., 2023).

C.3 Effect of the metric reference

Table 13 recomputes distribution distance for the same generated waveforms using either clean target stems or original scene mixtures as references. SelVA is closer to the mixtures; SelFoley and StemFoleyDB-adapted SelVA are closer to clean stems. Thus clean-stem distance reflects residual scene content as well as generic fidelity.

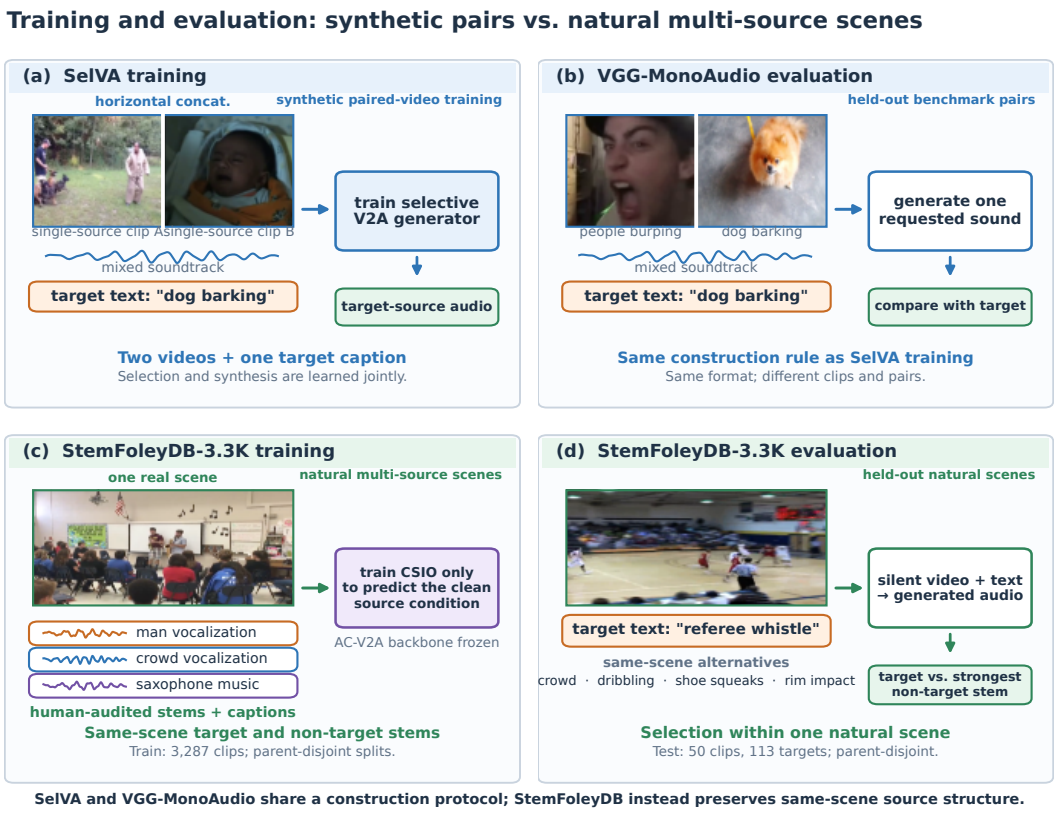


Figure 4: **Training and evaluation constructions.** (a) SelVA learns selective generation from synthetically paired VGGSound videos: two single-source clips are placed side by side, their audio is mixed, and target text names the requested source. (b) VGG-MonoAudio evaluates the same input construction on held-out curated clips. It is thus a native benchmark for SelVA, not a reuse of its training examples. (c) StemFoleyDB-3.3K instead starts from one naturally multi-source scene and supplies human-audited target and same-scene non-target stems to train CSIO while the AC-V2A generator remains frozen. (d) Its parent-disjoint test set provides a silent natural scene and target text; the generated audio is compared with the requested stem and the strongest same-scene alternative.

Table 11: Full VGG-MonoAudio target-selection and audiovisual-alignment results. Target source Top-1 and margin compare each generated waveform with the requested and paired non-target waveforms. IB is ImageBind audio–video similarity. CLAP-L and CLAP-MS are text–audio similarities from LAION-CLAP and MS-CLAP, respectively. MS-CLAP is independent of the representation used to train CSIO. Within each split, bold and underline mark the best and second-best results.

Split	Method	LAION-CLAP source ranking		MS-CLAP source ranking		IB ↑	DeSync ↓	CLAP-L ↑	CLAP-MS ↑
		Top-1 ↑	Margin ↑	Top-1 ↑	Margin ↑				
Inter-class	SelVA	<u>77.14</u>	**0.258**	<u>78.04</u>	**0.242**	**0.279**	**0.801**	**0.214**	**0.314**
	SelFoley	**85.89**	<u>0.223</u>	**85.00**	<u>0.201</u>	<u>0.189</u>	<u>0.856</u>	<u>0.146</u>	<u>0.239</u>
Intra-class	SelVA	**72.80**	**0.169**	**72.41**	**0.153**	**0.326**	<u>0.760</u>	**0.228**	**0.363**
	SelFoley	<u>72.41</u>	<u>0.117</u>	<u>72.02</u>	<u>0.118</u>	<u>0.245</u>	**0.673**	<u>0.168</u>	<u>0.290</u>

Table 12: Full VGG-MonoAudio audio-distribution and quality metrics. All values are recomputed from the same generated waveforms. Lower is better for FD and KL; higher is better for IS. Within each split, bold and underline mark the best and second-best results, respectively.

Split	Method	FD-VGGish ↓	FD-PANN ↓	FD-PaSST ↓	KL-PANN ↓	KL-PaSST ↓	IS-PANN ↑	IS-PaSST ↑
Inter-class	SelVA	**12.62**	**79.45**	**548.95**	**2.81**	**1.95**	**9.66**	**8.14**
	SelFoley	<u>33.91</u>	<u>82.60</u>	<u>700.56</u>	<u>3.28</u>	<u>2.89</u>	<u>6.67</u>	<u>5.51</u>
Intra-class	SelVA	**6.67**	<u>66.15</u>	**503.72**	**1.83**	**1.37**	**8.25**	**6.65**
	SelFoley	<u>13.10</u>	**64.15**	<u>640.97</u>	<u>2.24</u>	<u>2.24</u>	<u>6.25</u>	<u>5.46</u>

Table 13: Effect of evaluating the same 113 generated files against either the human-audited target stem (Target) or the original scene mixture (Mix). Lower is better; bold marks the closer reference within each method and feature space.

	FD-VGGish ↓		FD-PANN ↓	
Method	Target	Mix	Target	Mix
SelVA	25.65	**11.47**	67.28	**43.97**
SelVA + StemFoleyDB-3.3K LoRA	**6.14**	8.86	**39.01**	54.64
ControlFoley (default)	13.52	**6.98**	47.16	**41.46**
SelFoley (ours)	**4.71**	10.28	**33.73**	65.35
ControlFoley + target-stem reference	**3.89**	7.85	**26.83**	56.93

D HUMAN-EVALUATION PROTOCOL AND FULL RESULTS

Generation and selectivity ratings. We use 20 targets selected from a separately curated 90-video UnAV review pool. Exact identifiers, original-video identifiers, and media hashes are checked against StemFoleyDB-3.3K; these 20 videos are excluded from all dataset splits. Each trial shows one silent video, a bilingual target caption, and anonymized audio from official SelVA, SelFoley, and the human-curated target stem. The A/B/C mapping varies across items and is hidden from listeners. It is fixed at the item level rather than re-randomized for each listener, so the confidence intervals resample both listeners and items. The stem is a reference condition rather than a generated system output. We report three 1–5 ratings: (i) whether only the explicitly captioned sound is generated, (ii) whether the captioned sound is present, and (iii) audiovisual synchronization. There is no forced overall winner. The interface also collected perceived naturalness, but we exclude that response from analysis because a realistic full-scene soundtrack can score well by retaining plausible non-target sounds and thereby conflict with the selective-generation objective.

Target-selection-margin validation. To test whether TargetSel agrees with human perception, we use 30 distinct multi-source clips from the fixed test split. Candidate pairs were first constructed from human-curated captions and distinct-source relations. Before data collection, we performed an audibility and semantic-ambiguity screen: corrupted or inaudible outputs, mislabeled media, and pairs whose target and non-target descriptions could not be perceptually distinguished were replaced. The automatic M2D-CLAP scores and the metric-preferred A/B direction were hidden during this screen. Consequently, the evaluation validates the metric on audible, semantically distinguishable interventions; it is not a random sample of every generated output.

Ten pairs replace the target source-content condition with that of another verified source from the same video while keeping target timbre fixed. Ten pairs replace the target-stem input to the semantic encoder with the scene mixture, again keeping target timbre fixed. The remaining ten replace only the timbre input while keeping the target source-content condition fixed, providing a negative control. Within every pair, the clip-level 8-s window, initial noise, sampler, 25 Euler steps, classifier-free guidance, and output duration are identical. A and B are balanced within each intervention type, and condition names are hidden.

Each trial shows the bilingual target caption, the other verified source captions from the same scene, a silent video, and two anonymized outputs. The listener answers which output makes the target clearer relative to the other scene sounds, with A, B, and "no clear difference" as the three choices. For each output, the automatic margin is target similarity minus the highest same-scene non-target similarity under M2D-CLAP. The metric predicts the member with the larger margin. We report agreement on decisive A/B responses, agreement with the majority-human direction on decisive items, Spearman correlation between the signed A–B margin difference and mean signed human preference, and the tie rate. A two-way bootstrap resamples both listeners and items.

Across 540 responses, listeners report no clear difference on 266 (49.3%). Among the remaining 274 decisive responses, the automatic margin agrees with the listener on 238 (86.9%; two-way bootstrap 95% CI $[77.4, 93.7]$). Agreement is 140/160 (87.5%) when the target source-content condition is replaced with that of another source from the same video and 90/102 (88.2%) when the target-stem input is replaced with the scene mixture. The timbre intervention behaves as the intended negative control: 168 of its 180 responses report no clear difference, leaving only twelve decisive responses. At the item level, the metric agrees with the listener-majority direction for 23 of 26 decisive items (88.5%); four items tie. The signed metric difference correlates with mean human preference ($\rho = 0.908$). These findings support the audible consequence of the controlled source-content intervention and the perceptual relevance of the associated margin direction. They do not by themselves identify the internal conditioning path—that conclusion additionally requires the 2×2 experiment that swaps the source-content and timbre conditions separately—or establish M2D-CLAP as a universal model of source perception.

For both listening evaluations, condition identity is hidden, playback may be repeated as needed, and neither headphones nor complete playback is required. Name and email are collected only for compensation and duplicate-response control, stored separately from reported aggregate results, and deleted after compensation. The reported snapshot contains 18 listeners in each evaluation.

Confidence intervals for the generation ratings bootstrap crossed listener and item units; those for target-selection-margin validation use a two-way listener-and-item bootstrap.

E Additional analysis of the semantic reference pathway

The analyses below separate four questions: how much reference audio the encoder needs, how competing content changes its representation, which ControlFoley pathway causes the output change, and when that signal acts during generation.

E.1 Reference duration and competing content

For each test source, we take nested 0.5-, 1-, 2-, and 4-s clean-stem crops around the same audited event peak. Crops are RMS-matched by attenuation only, and timbre is fixed. The target-minus-strongest-other margin rises through 2 s and is nearly unchanged at 4 s. Cross-duration retrieval identifies the same source in 96.24% of the 12 ordered off-diagonal comparisons (clip-bootstrap 95% CI $[93.87, 98.42]$). This is a representation audit, not a waveform-quality comparison.

Table 14: Reference-duration audit in ControlFoley's frozen source-content space. Crops are nested around the same event peak and RMS-matched by attenuation; timbre remains fixed. Source ID retrieves the correct verified same-scene source.

Reference duration	0.5 s	1 s	2 s	4 s
Source ID Top-1 (%)	91.15	95.58	98.23	99.12
Own–other cosine margin	.237	.369	.482	.475
Cosine to 4-s condition	.529	.662	.792	1.000

Adding an RMS-matched same-scene source changes both the representation and generated identity, while changing target gain alone does not. Across 565 source–condition pairs, the injected-minus-target representation score correlates with the generated score at Spearman $\rho = 0.786$. Point estimates weight sources equally; confidence intervals resample parent videos. The test isolates competing source content and gain, not reverberation, codec artifacts, or stationary noise.

Table 15: Reference-contamination sweep. An RMS-matched same-scene non-target is added at relative amplitude α; the gain control changes target amplitude without adding a source.

α	0	.25	.50	1.0	2.0
Reference-content target Top-1	100.0	80.5	66.4	52.2	38.1
Generated target Top-1	95.6	72.6	60.2	46.9	38.1
Generated injected-source Top-1	4.4	26.5	38.9	49.6	58.4
Target-only gain control	95.6	—	94.7	—	94.7

We also interpolate directly between target and same-scene non-target source-content conditions. At non-target weights $0, .25, .5, .75$, and 1, target selection is 94.7, 87.6, 47.8, 8.8, and 2.7%, while selection of the substituted source rises in the opposite direction. This continuous switch argues against a binary clean/unclean detector.

E.2 Pathway localization and causal controls

The 2×2 intervention changes the source-content and timbre conditions independently while holding video, reference interval, initial noise, sampler, and guidance fixed. Replacing source content accounts for the selection change; replacing timbre does not.

Table 16: Semantic/timbre encoder swap experiment on the fixed test set. *Target stem* uses the separated requested source; *scene mixture* uses original audio from the same 2–4 s reference interval. TargetSel is same-scene top-1 accuracy, and margin is target similarity minus the strongest same-scene non-target similarity. Video, initial noise, sampler, and guidance remain fixed.

Semantic input	Timbre input	TargetSel (%)	Target sim.	Margin
Scene mixture	Scene mixture	60.18	0.703	0.037
Scene mixture	Target stem	61.95	0.703	0.037
Target stem	Scene mixture	**94.69**	0.819	0.209
Target stem	Target stem	**94.69**	**0.820**	**0.211**

At three matched sampling points for each of 113 sources, changing only source content produces an update change with 0.998 times the magnitude of the complete target-reference change and cosine similarity 0.996 to it. Changing only timbre gives a magnitude ratio of 0.067. These are intervention magnitudes and directions, not explained-variance scores.

The complete target-to-non-target source swap repeats over three seeds: the substituted-source-minus-target margin rises by 0.457 on average (seed standard deviation 0.0055; video-cluster 95% CI $[0.403, 0.511]$). A stronger cross-video control inserts a donor category absent from the scene's audited inventory. Donor retrieval rises from 0.9% to 93.5% under M2D-CLAP and from 1.2% to 92.9% under LAION-CLAP, while output level changes by only -0.41 dB (95% CI $[-1.18, 0.35]$). Thus the effect is not simple attenuation. Because the source inventory is not exhaustive, this control does not claim that the donor category is visually impossible.

E.3 Representation geometry

A linear classifier trained to distinguish target-stem from mixture reference embeddings is more accurate in source-content space than in timbre space (area under the receiver operating characteristic curve, AUROC: 91.65% vs. 76.08%) and remains accurate after matching root-mean-square (RMS) amplitude (90.54%). Across test sources, a larger target-stem–mixture displacement in source-content space predicts a larger improvement in the generated target-minus-non-target margin (Spearman $\rho = 0.744$, 95% confidence interval $[0.645, 0.815]$). The embedding change is therefore connected to the audible selection change rather than being only a separable reference feature.

The average target-stem-minus-mixture source-content vector from training identifies the target-stem member of 91.15% of held-out pairs, so a shared tendency exists. It is not sufficient to specify a source: under principal component analysis (PCA), the first training-derived direction captures only 10.24% of the squared target-stem-minus-mixture vector magnitude on test, while 64 directions capture 83.67%. For diagnosis only, we project each test vector onto this training-derived basis using its true coefficients; decoded TargetSel rises from 64.6% at rank 1 to 85.0% at rank 16 and 92.0% at rank 64. Because those coefficients use the test target-stem reference, this is not a deployable inference method.

The leading axes are stable when the analysis is repeated on data partitions with no shared original video, but they are only partly nameable. The first contrasts brief foreground impacts with sustained ambience; the second separates non-vocal object or percussive events from vocal events. Later axes are reproducible without simple one-phrase labels. A linear predictor from frozen caption embeddings achieves held-out $R^2 = 0.066$ for the first 16 coefficients. CSIO is therefore better viewed as predicting source-dependent coordinates in the source-content space than adding a single universal isolation direction.

Table 17: Leading train-derived source-content directions. Stability measures recurrence across video-disjoint data partitions; text ρ measures held-out caption-to-coordinate rank correlation. Principal-component signs are arbitrary.

Factor	Stability	Text ρ	Empirical interpretation from test examples
PC1	.974	.521	Brief foreground impacts and handling $\leftrightarrow$ sustained ambience; also tracks the magnitude of the target-stem-to-mixture embedding change
PC2	.946	.454	Object contacts, claps, and percussion $\leftrightarrow$ speech, singing, and vocalization
PC3	.928	.388	Weak tonal/musical $\leftrightarrow$ dry-event tendency, with substantial mixing
PC4	.907	.103	Reproducible axis without a consistent caption-level label
First 16 directions	.862	.229	Stable shared geometry with source-dependent, entangled coefficients

E.4 When the signal acts and how it enters the generator

Figure 5 separates two experiments that are easy to confuse. Panel (a) changes the source-content condition across *sampling time*: we run the iterative generator repeatedly and switch between the target and same-scene non-target conditions at 11 points in both directions, holding every other input fixed (2,486 outputs). Supplying the target source-content condition only during the first half yields 93.8% TargetSel; supplying it only during the second half yields 6.2%. Source identity is therefore committed mainly in early sampling, although later steps can still matter for timbre and quality.

Panel (b) instead follows *network depth within one sampling step*. We copy one internal representation from a target-stem run into an otherwise identical mixture-reference run, a diagnostic known as activation patching. At the input, the source-content condition carries the full intervention effect and the generator's noisy audio state carries almost none. By blocks 3–6, most of the effect has moved into that audio state; by later blocks, copying the original condition alone recovers little. This shows how the condition is written into the state that is ultimately decoded. Network depth is not sampling time, and no single block is an independent selector.

The early-commitment result motivated focusing the later inference search on the beginning of sampling. The selected SelFoley rule strengthens visual synchronization during the first quarter of the trajectory while keeping the source-content condition active throughout; it does not patch individual blocks. A separate localization test also clarifies the pathway's limitation: its frame-level event-activity area under the receiver operating characteristic curve is 0.481, near chance. The source-content condition strongly controls *what* is generated but does not directly specify *when* it should occur.

Attempted explicit timing control. We tried two ways to confine sound to a predicted interval: multiplying generation by a frame-level activity mask predicted from video and text, and applying the reference-induced update only inside that interval. Both can attenuate energy outside the interval, but reduce source selection and can retain non-target content. They do not make the frozen generator relocate an event. We therefore treat target-specific temporal control as future work rather than a positive result.

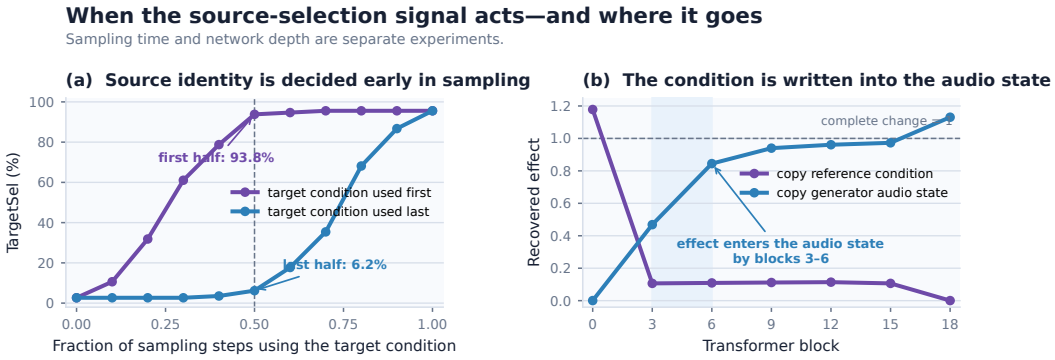


Figure 5: **When the source-selection signal acts, and where it goes.** (a) The target source-content condition has a much larger effect when supplied during early rather than late sampling steps. This finding motivated concentrating SelFoley's inference search on the early trajectory. (b) Within one sampling step, activation patching shows the source-content effect moving from the reference condition into the generator's noisy audio state across the early transformer blocks. The dashed line denotes the complete target-stem–mixture update difference. Activation patching is a diagnostic and is not used by SelFoley at inference.

F CSIO DEVELOPMENT ABLATIONS

The adapter-architecture, input-representation, and training-objective ablations use checkpoints selected by validation performance. The inference-coefficient study is a separate exploratory analysis of the source-selection–synchronization trade-off, described below; we do not characterize its final setting as an untouched, validation-only model-selection result.

F.1 PREDICTED CONDITION AND TRAINING OBJECTIVE

Table 18: Predicted-condition and objective ablation. Condition retrieval ranks each predicted representation against verified same-scene targets. Checkpoints are selected on validation.

Predicted condition and objective	Condition retrieval (%)	Target cosine	Margin
Source-content: target regression	74.34	.648	.128
Source-content: + batch contrastive	78.76	.649	**.171**
Source-content: + same-scene ranking	**80.53**	.644	.170
Timbre: target regression	53.10	**.800**	.012
Timbre: + same-scene ranking	64.60	.755	.037

Predicting the source-content condition is substantially more selective than predicting timbre. A batch-level contrastive loss, which pulls matched caption–source pairs together and pushes unmatched pairs apart, and the more targeted same-scene ranking loss both improve over fitting each example independently. Timbre prediction has high target cosine similarity but little ability to reject a same-scene non-target.

F.2 INPUTS, SUPERVISION, AND TEXT ENCODERS

Table 19: CSIO input ablation. The full objective adds same-scene ranking and difference alignment; video residual denotes a learned video-dependent correction.

Adapter input/objective	Condition retrieval (%)	Target cosine	Margin
Text + target-vs-mixture ranking	79.65	**.662**	.172
Text + full CSIO objective	**83.19**	.653	**.173**
Video+text residual + target-vs-mixture ranking	79.65	.642	.165
Video+text residual + full CSIO objective	72.57	.630	.152

Table 20: CSIO training-data ablation. Fractions are sampled deterministically by clip; checkpoints are selected on validation. Rows use target fitting and target-versus-mixture ranking unless stated otherwise.

Training data/objective	Train clips	Train stems	Condition retrieval (%)	Target cosine	Margin
All, 25%	818	1,284	74.34	.639	.151
All, 50%	1,637	2,570	76.99	.652	.164
All, 100%	3,274	5,191	79.65	**.662**	.172
Multi-stem clips only	1,651	3,568	81.42	.643	.171
Single-stem clips only	1,623	1,623	76.99	.633	.133
All, 100% + same-scene pair losses	3,274	5,191	**83.19**	.653	**.173**

Adding a video-dependent correction does not improve the text-to-source-content adapter: same-scene sources share the video and must be distinguished by the requested identity. Multi-source-only training outperforms single-source-only training despite similar clip counts, and using all sources with same-scene difference alignment is strongest at the adapter stage.

Table 21: Frozen text-representation study with separately trained CSIO adapters. All rows use the same fixed ControlFoley test protocol; the final row adds SelFoley's sampling guidance.

Text representation	Objective	TargetSel (%)	Target cosine	Margin	DeSync ↓	FD-VGGish ↓	FD-PANN ↓	KL-PANN ↓	IS-PANN ↑
LAION-CLAP	target-vs-mixture ranking	81.42	**0.733**	0.112	0.855	5.06	35.41	1.906	4.772
LAION-CLAP	all four CSIO losses	81.42	0.732	0.112	0.878	5.03	35.60	1.924	**4.786**
M2D-CLAP	all four CSIO losses	**82.30**	0.727	**0.116**	0.779	5.30	35.31	1.943	4.559
OEA	all four CSIO losses	65.49	0.673	0.067	0.873	6.28	41.23	2.467	3.480
M2D-CLAP	**all losses + SelFoley guidance**	78.76	0.732	0.111	**0.384**	**4.71**	**33.73**	**1.860**	4.485

M2D-CLAP (Niizumi et al., 2024) and LAION-CLAP (Wu et al., 2023) transfer similarly through separately trained adapters. Omni-Embed-Audio (OEA) (Yoo et al., 2026) transfers less effectively. The text embedder alone does not perform selection: every row learns the generator-specific mapping.

F.3 SYNCHRONIZATION-GUIDANCE SEARCH

The sampling-time analysis motivates one coefficient for source-content guidance and another for early visual synchronization. Table 22 contains the detailed early/late comparison moved from the main paper together with representative coefficient settings. The selected $\beta = 2.20$, $\gamma = 10.5$ rule is used unchanged in every final SelFoley result. This was an exploratory trade-off search, so its test score characterizes the selected system rather than an untouched confirmatory estimate.

Table 22: Inference-time selection–synchronization search. Validation rows use 20 clips/44 sources; β scales source content and γ scales synchronization over the stated denoising interval. The last row is the selected test configuration.

Stage	Sync schedule	β	γ	Split	TargetSel (%) ↑	DeSync ↓
CSIO only	none	1.00	0.0	valid	86.36	0.768
Late synchronization control	late 25%	1.00	1.0	valid	86.36	0.780
Early synchronization guidance	early 25%	1.00	1.0	valid	79.55	0.516
Moderate source-content scale	early 25%	1.60	4.0	valid	84.09	0.395
Stronger source-content scale	early 25%	1.80	6.5	valid	86.36	0.375
High source-content scale	early 25%	2.40	8.5	valid	**90.91**	0.366
Selected setting	early 25%	2.20	10.5	valid	86.36	**0.318**
SelFoley	**early 25%**	**2.20**	**10.5**	test	**78.76**	**0.384**

G CROSS-BACKBONE DETAILS

We evaluate ControlFoley, AC-Foley, and Video-Foley on the same source identities from the fixed 50-clip/113-source test set with seeds 240923, 240924, and 240925. Within each backbone, four conditions use the same model-specific preprocessing and matched generation seeds: its default text-guided mode, a regression adapter that independently fits each target-stem representation, CSIO with additional source-selection losses, and a non-deployable target-stem reference. Across backbones, we hold the source requests and evaluation encoders fixed but retain each released model's own video preprocessing, sample rate, conditioning interface, and inference configuration. The experiment therefore tests whether interface learning improves each frozen backbone relative to itself; it is not a direct ranking of the three architectures under identical preprocessing.

The primary M2D-CLAP ranking is repeated with LAION-CLAP. Source-selection differences weight sources equally, and confidence intervals resample original videos. *Default text-guided mode* denotes each backbone's original video–text condition. All generators remain frozen, and the target-stem reference condition is non-deployable.

AC-Foley averages its reference features into one global audio vector; its adapter predicts that vector while keeping the visual sequence fixed. Video-Foley uses a shared CLAP embedding for sound content and a separate Video2RMS module that predicts a frame-wise root-mean-square (RMS) energy envelope from video. Its adapter changes only content. No ControlFoley parameters or adapter weights are transferred to either model. Regression is already strong on AC-Foley and Video-Foley, and CSIO does not exceed simple regression on every metric. The supported conclusion is narrower: on each of the three evaluated backbones, predicting that backbone's native reference-content interface improves over its default text-guided mode under both source-ranking evaluators. The same adaptation strategy can therefore be applied to all three architectures, but this does not establish one universal checkpoint or one identical internal mechanism.

Table 23: Complete cross-backbone source-selection results on the fixed test set. TargetSel is same-scene source-ranking accuracy under M2D-CLAP or LAION-CLAP; values average the same three seeds. *Default* is each backbone's standard text-guided mode. *Regression* fits the target-stem condition independently for each example; *CSIO* additionally uses target-versus-mixture and same-scene source supervision. The target-stem reference consumes target audio and is not deployable. Switch is the fraction of same-scene source pairs for which changing the request changes the M2D-ranked output in both directions.

Backbone	Condition	M2D TargetSel	Margin	LAION TargetSel	Switch
ControlFoley	Default	58.70	.030	57.82	.259
	Regression	72.27	.077	71.39	.600
	CSIO	82.30	.120	79.65	.741
	Target-stem reference	93.51	.192	91.45	.918
AC-Foley	Default	53.98	.010	53.39	.188
	Regression	62.83	.046	61.36	.282
	CSIO	61.36	.051	62.24	.271
	Target-stem reference	81.42	.142	80.83	.647
Video-Foley	Default	73.45	.076	67.55	.588
	Regression	82.60	.108	78.47	.753
	CSIO	81.71	.108	78.76	.753
	Target-stem reference	94.99	.183	95.87	.894

Table 24: Cross-backbone audio-distribution and synchronization metrics on the fixed test set. Entries are mean ± sample standard deviation over three matched seeds. *Default* denotes each backbone's standard text-guided mode. FD-VGGish and FD-PANN use VGGish and PANN features; KL and IS use PANN. FD, KL, and DeSync are lower-is-better; IS is higher-is-better. The target-stem reference is non-deployable.

Backbone	Condition	FD-VGGish ↓	FD-PANN ↓	KL-PANN ↓	IS-PANN ↑	DeSync ↓
ControlFoley	Default	11.151±1.095	53.636±9.203	2.799±.167	4.017±.240	.676±.024
	Regression	5.797±.520	42.085±2.604	2.357±.039	4.097±.172	.833±.055
	CSIO	4.398±.683	36.718±2.874	2.107±.070	4.667±.228	.776±.019
	Target-stem reference	3.899±.907	29.466±2.040	1.482±.149	4.917±.155	.865±.046
AC-Foley	Default	29.971±1.250	72.257±5.420	3.031±.110	4.877±.169	.362±.026
	Regression	13.780±2.016	53.414±1.737	2.661±.078	4.228±.241	.493±.014
	CSIO	13.770±2.420	51.319±2.801	2.606±.055	4.325±.073	.464±.039
	Target-stem reference	9.057±1.961	42.470±4.407	1.691±.128	4.706±.216	.522±.038
Video-Foley	Default	5.512±.331	41.671±1.691	2.663±.049	3.614±.020	1.214±.015
	Regression	5.399±.799	37.036±1.017	1.960±.126	4.226±.183	1.169±.044
	CSIO	5.233±.623	37.843±1.282	1.986±.076	4.109±.158	1.173±.065
	Target-stem reference	3.962±.574	32.074±1.518	1.430±.065	4.392±.171	1.215±.047

Backbone-specific inference calibration. The tables above use each backbone's CSIO adapter without added inference-time guidance so that all comparisons remain within backbone. We additionally calibrate AC-Foley's audio- and synchronization-guidance coefficients as an architecture-specific trade-off. At audio guidance 0.75 and synchronization guidance 1.50, M2D TargetSel changes from 61.36% to 58.70% and DeSync from 0.464 ± 0.039 to 0.378 ± 0.027 over the same three seeds. This setting recovers synchronization while retaining an improvement over AC-Foley's 53.98% TargetSel in its default text-guided mode; it is not uniformly best on every metric. Video-Foley's sweep does not improve DeSync over its default adapter condition, so we report the default condition.