BEYOND TARGET MATCHING: MEASURING AND MITIGATING NON-TARGET LEAKAGE IN SELECTIVE VIDEO-TO-AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective video-to-audio (V2A) generation aims to synthesize sounds only for user-specified sources while suppressing other visible sound sources in the scene. Generated audio may contain the requested sound along with sounds from other visible sources. We call this **non-target leakage** and evaluate source selectivity through both target retention and competing-source suppression. To measure this behavior, we introduce **VGG-SourcePair**, a paired evaluation benchmark for selective generation in dual-source scenes. Given the same silent video containing two sound sources, we alternately designate each source as the target and jointly evaluate target preservation and non-target leakage, providing a more direct measure of source selection ability. We further propose **FocusV2A**, which uses the target prompt to select local visual evidence relevant to the requested source before audio generation. A Focus Selector estimates the target relevance of visual features and uses it to modulate spatial attention, yielding a target-focused visual condition. Training proceeds in two stages: the first learns target-aware visual selection, and the second adapts the generator to the resulting visual condition. At inference time, only a silent video and a target prompt are required. Experiments show that FocusV2A achieves stronger source selection performance on VGG-SourcePair while maintaining competitive audio quality and audiovisual synchronization on VGG-MonoAudio.

1 INTRODUCTION

In film, short-form video, and game post-production, sounds are typically created and edited as separate tracks, allowing their volume, timing, and mixing to be controlled independently. Accordingly, when a video contains multiple potential sound sources, a user may want to generate audio for only one specified source. For example, in a scene containing both a dog and a cat, a request such as "dog growling" should generate the sound of the dog while preventing cat sounds from entering the target track. Even if the target sound itself is accurate and natural, the resulting track is still contaminated if sounds from other visible sources are also generated. We refer to this phenomenon as **non-target leakage**.

Recent V2A models have improved audio quality, semantic alignment, and audiovisual synchronization, as demonstrated by MMAudio, HunyuanVideo-Foley, and Omni2Sound (Cheng et al., 2025; Shan et al., 2025; Dai et al., 2026). More recent work extends these capabilities toward finer control over sound production. MultiFoley and ControlFoley combine multimodal conditions to control sound content and acoustic properties (Chen et al., 2025; Yang et al., 2026), while AC-Foley transfers acoustic characteristics from reference audio (Fang et al., 2026). Temporal control provides a complementary capability: FoleyDirector uses structured temporal scripts to guide when sound events occur (Li et al., 2026b). Source selection concerns which visible source should be heard. SelVA addresses this problem by modifying intermediate video representations through text-guided cross-attention, with learnable supplementary tokens that help suppress irrelevant activations (Lee et al., 2026). The resulting target-conditioned features guide the audio generator toward the requested source. Evaluating this selectivity requires examining both correspondence to the target and the presence of competing sources in the generated track.

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.

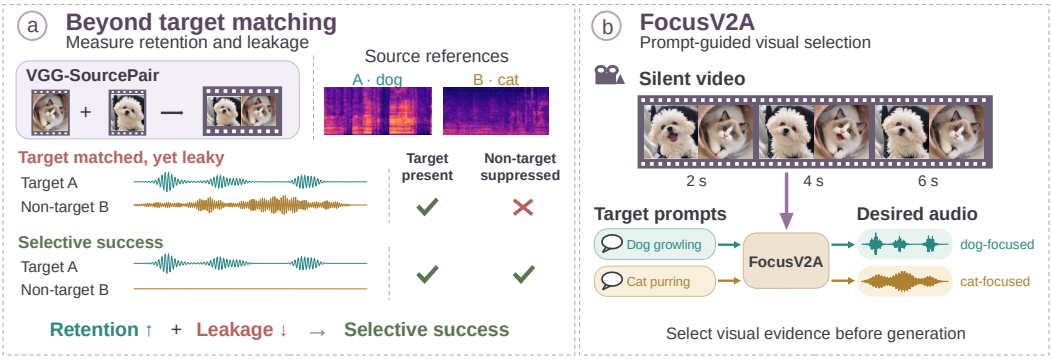


Figure 1: Overview of the problem and approach. (a) VGG-SourcePair distinguishes target retention from non-target leakage using paired component references. (b) FocusV2A selects visual evidence according to the target prompt. Waveforms schematically illustrate source activity.

Figure 1(a) illustrates the distinction: both outputs contain the target, but only the selective output suppresses the competing source. We introduce VGG-SourcePair to measure these two aspects together. Each source in a two-source video is requested in turn, and each output is evaluated against both component recordings. Reversing the requests tests whether the model can retain either source while excluding the other.

FocusV2A addresses this choice by using the target prompt to guide visual aggregation before audio generation (Figure 1b). The relevance of a local visual cue depends on the request: evidence associated with one source should be emphasized when that source is requested and downweighted when the request shifts to another. This motivates a competitive allocation of each local feature between the target and its complement. A Focus Selector predicts these assignments from the target prompt and local visual tokens, then converts the target assignments into log-biases for pretrained spatial attention. This changes how visual evidence is aggregated while preserving the original tokens and keeping the visual backbone and aggregation parameters frozen. We first train the selector to match component-level teacher representations, then freeze it and adapt MMAudio to the resulting target-focused visual condition.

Experiments on VGG-SourcePair show improved target retention and non-target suppression over the evaluated video-conditioned baselines. Evaluation on VGG-MonoAudio further demonstrates competitive audio quality, semantic alignment, and audiovisual synchronization.

Our main contributions are as follows:

- **A more complete evaluation perspective for selective V2A.** We argue that generating the requested source is not sufficient for successful source selection; other visible sources must also be prevented from leaking into the generated track.
- **VGG-SourcePair.** We introduce a paired benchmark that requests each source in the same two-source scene in turn and evaluates both target retention and non-target leakage, providing a more direct source-level assessment of selective V2A generation.
- **FocusV2A.** We propose a selective V2A framework that uses the target prompt to select source-relevant visual evidence before audio generation and adopts a two-stage training strategy to adapt the generator to this target-focused visual condition.

2 Related Work

2.1 Video-to-Audio Generation and Control

Recent V2A systems combine video and text to generate semantically aligned and synchronized audio, including MMAudio, HunyuanVideo-Foley, and Omni2Sound (Cheng et al., 2025; Shan et al., 2025; Dai et al., 2026). ReWaS uses video-derived energy to control text-to-audio generation (Jeong et al., 2025), while VinTAGe jointly conditions generation on video and text to produce holistic

soundtracks (Kushwaha & Tian, 2025). Controllable synthesis extends these capabilities to specific sound content and acoustic properties. MultiFoley and ControlFoley use multimodal conditions to guide generation (Chen et al., 2025; Yang et al., 2026), while AC-Foley transfers acoustic characteristics from reference audio to video-conditioned outputs (Fang et al., 2026).

Other methods guide generation through explicit reasoning. ThinkSound uses chain-of-thought reasoning for audio generation and editing (Liu et al., 2026a), and PrismAudio combines decomposed reasoning with rewards for semantics, timing, aesthetics, and spatial accuracy (Liu et al., 2026b).

Source selection uses the target prompt to distinguish the requested sound from competing visible sources. SelVA introduces text-guided cross-attention and supplementary tokens into the video encoder, learning selective visual representations through self-supervised teacher–student distillation with video mixing (Lee et al., 2026). By contrast, FocusV2A expresses selection as a competitive target–complement allocation over local tokens. These assignments modulate frozen spatial attention, changing the contribution of each token while retaining the pretrained visual features and aggregation parameters.

2.2 Evaluation of Selective Video-to-Audio Generation

V2A evaluation commonly measures audio quality, semantic correspondence, and temporal alignment using FAD, CLAP, and synchronization models (Kilgour et al., 2019; Wu et al., 2023; Iashin et al., 2024). VABench and FlatSounds extend evaluation to complementary generation capabilities (Hua et al., 2026; Li et al., 2026a). For selective generation, VGG-MonoAudio provides single-source references to assess correspondence to the requested source (Lee et al., 2026). VGG-SourcePair complements this protocol by evaluating each output against both component recordings and measuring target and competing-source activity separately. Its paired requests assess retention and leakage as the two sources exchange roles.

3 Benchmark

3.1 Task Definition

Given a silent video v and a target prompt p_t, selective video-to-audio generation produces a waveform

$$\hat{a} = G(v, p_t; \xi), \tag{1}$$

where ξ is the sampling seed. Each two-source video is evaluated with one request per source. The requested component recording serves as the target reference and the other as the non-target reference; their roles reverse with the request (Figure 1). The benchmark component recordings are used only for scoring and are not provided to the generator. The objective is to retain the requested source while suppressing its competitor.

3.2 Paired Benchmark Construction

VGG-SourcePair contains 512 clips from distinct parent videos in the official VGGSound test split, spanning 48 event categories and six superclasses (Chen et al., 2020). Clips pass automatic checks for energy, silence, clipping, duration, and resolution, and are prioritized within categories by frame–label similarity.

We pair clips with different event labels across or within superclasses, using each clip once. The first eight seconds of each view are resized and padded with aspect ratio preserved, then concatenated horizontally into a silent video. The corresponding original recordings, including residual background sounds, form the evaluation references. Prompts are derived from event labels without spatial cues or descriptions of the competing source.

The resulting 256 videos yield 512 directed tasks. Test clips and their parent videos are excluded from our training pool. Appendix A specifies the screening criteria and pairing strata.

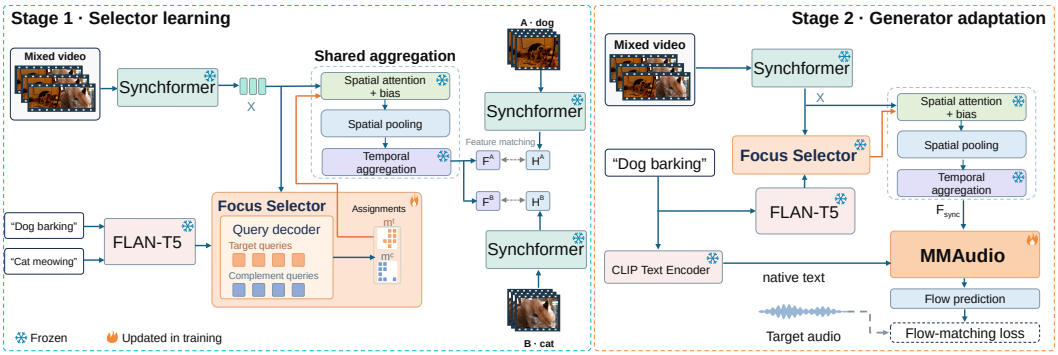


Figure 2: Two-stage training of FocusV2A. Stage 1 learns visual selection by matching mixed-scene representations to unmixed-component teachers. Stage 2 freezes the selector and adapts the generator to the selected visual sequence and native text condition. Icons indicate frozen and updated modules.

4 METHOD

FocusV2A uses the target prompt in two complementary paths. FLAN-T5 conditions the Focus Selector, which assigns relevance to local features from a frozen Synchformer (Iashin et al., 2024) and biases their spatial aggregation. Frozen temporal aggregation forms the synchronization sequence used by MMAudio (Cheng et al., 2025). In parallel, the native CLIP text encoder (Radford et al., 2021) supplies the generator's text condition. Figure 2 shows how we first learn the visual selection path from component representations, then adapt the generator while keeping this path fixed.

4.1 PROMPT-GUIDED VISUAL SELECTION

For each video, we arrange the local Synchformer features as $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times d_v}$, where $x_i \in \mathbb{R}^{d_v}$ and $N = ST_vHW$. Here, S is the number of segments, T_v the number of temporal positions per segment, $H \times W$ the spatial token grid, and d_v the feature dimension. The index i enumerates these spatiotemporal positions. Before spatial pooling, the Focus Selector jointly assigns each token to the target and its complement.

A frozen FLAN-T5 encoder (Chung et al., 2024) maps p_t to token features. We project and pool its non-padding features into a prompt representation that conditions two learned query banks, one for the target and one for its complement. Each bank contains four queries and attends to the same projected visual tokens through two transformer decoder layers. Averaging within each bank gives a target state and a complement state. We add a shared prompt-dependent projection to the target state and subtract it from the complement state. Scaled dot products with the local features then produce logits ℓ_i^t and ℓ_i^c, which are normalized jointly:

$$[m_i^t, m_i^c] = \mathrm{softmax}\left([\ell_i^t, \ell_i^c]/\tau\right), \qquad m_i^t + m_i^c = 1, \tag{2}$$

Here, τ is the assignment temperature. Joint normalization allocates each feature between the target and its complement, so a larger target assignment leaves a smaller complement assignment. The complement collects evidence not selected by the current prompt and does not require a separate source description. Only m_i^t biases the visual condition used by the generator.

4.2 SELECTION-BIASED VISUAL AGGREGATION

The selector guides spatial aggregation through an additive key bias. At each temporal position, we add the bias derived from the target assignments to the pretrained query–key scores before softmax:

$$\widetilde{m}_i^t = \mathrm{clamp}(m_i^t, \varepsilon_m, 1), \qquad b_i^t = \beta \log \widetilde{m}_i^t,$$
$$A^t = \mathrm{softmax}_{\mathrm{key}}\left(\frac{QK^\top}{\sqrt{d_h}} + \mathbf{1}(b^t)^\top\right). \tag{3}$$

Here, $\beta > 0$ controls selection strength and $0 < \varepsilon_m < 1$ keeps the logarithm finite. Q and K are computed after input normalization, and d_h is the head dimension. The bias is shared across queries and heads, with zero bias for the aggregation-token key. For spatial key i, attention is proportional to $\exp(a_{qi})(\widetilde{m}_i^t)^\beta$, where $a_{qi} = Q_q K_i^\top / \sqrt{d_h}$. Selection therefore reweights pretrained attention while preserving its content dependence.

The pretrained value projection, residual pathway, and spatial/temporal aggregation parameters remain frozen. Spatial aggregation-token outputs pass through the temporal aggregator, and segment outputs are concatenated into F_{sync}. MMAudio projects this sequence into its synchronization-conditioning modules. The target prompt separately enters the generator's native CLIP text encoder, while the CLIP image branch receives an empty condition. Thus, F_{sync} supplies the video-dependent condition.

4.3 Two-Stage Training

Mixed training examples concatenate two preprocessed VGGSound source views into a square frame, with each source occupying half the width. Both source requests provide supervision, and unmixed examples are included to maintain alignment on single-component inputs.

Stage 1: Learning visual selection. Only the Focus Selector is updated; the text encoder, visual backbone, and spatial/temporal aggregators remain frozen. For a component pair, let F^A and F^B be the selected synchronization sequences under the two requests, and H^A and H^B the teacher sequences extracted from the corresponding unmixed videos. Gradients are backpropagated through the frozen aggregators, allowing the selector to recover the requested component representation from the mixed scene.

We train the selector with component-feature alignment and two regularization terms. Writing $s(U, V)$ for mean token-wise cosine similarity, $\mathcal{L}_{\mathrm{align}}$ averages $2 - s(F^A, H^A) - s(F^B, H^B)$ over training pairs. The relative-similarity term $\mathcal{L}_{\mathrm{sep}}$ uses a margin μ between similarities to the requested and competing teachers. The cross-request regularizer $\mathcal{R}_{\mathrm{pair}}$ combines assignment complementarity and centered anti-correlation:

$$\mathcal{L}_{\mathrm{stage1}} = \mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{sep}}\mathcal{L}_{\mathrm{sep}} + \lambda_{\mathrm{pair}}\mathcal{R}_{\mathrm{pair}}. \tag{4}$$

Here, λ_{sep} and λ_{pair} weight the two regularizers. Mixed examples use all terms, whereas unmixed examples use only $\mathcal{L}_{\mathrm{align}}$. The paired regularizer links the two requests, complementing the within-request allocation in Equation 2.

Stage 2: Adapting audio generation. We freeze the trained selector and all visual/text encoders, updating only the synchronization-input projection, positional embedding, and selected latent adaptive-normalization modules of MMAudio-S-16k.

Following MMAudio (Cheng et al., 2025), conditional flow matching supervises the target audio latent z_1. Given Gaussian noise z_0, flow time $u \in [0, 1]$, interpolation $z_u = (1-u)z_0 + uz_1$, and condition $c = (F_{\mathrm{sync}}, E_{\mathrm{CLIP}}(p_t))$, we optimize

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{u, z_0, (z_1, c)}\left[\mathrm{MSE}\left(v_\theta(z_u, u, c),\, z_1 - z_0\right)\right], \tag{5}$$

where the mean squared error is averaged over latent coordinates and $\mathcal{L}_{\mathrm{stage2}} = \mathcal{L}_{\mathrm{CFM}}$.

Inference. Given a silent video and one target prompt, FocusV2A computes the selected visual sequence and native text condition. MMAudio generates an audio latent from these conditions and decodes it to a waveform. Appendix C provides the architecture settings and full training objectives.

Table 1: VGG-SourcePair results on 512 requests from 256 videos. † denotes text-only controls; bold and underlined values indicate the best and second-best video-conditioned scores.

	Source Identity		Temporal Selection			Quality	Sync
Method	TextSep↑	RefSep↑	Leak@Dur↓	Target@Dur↑	Temp Success↑	FAD↓	DeSync↓
AudioLDM2†	0.2606	0.2822	0.1940	0.9207	0.7408	40.48	1.2604
Tango†	0.0335	0.0355	0.1900	0.3167	0.2320	96.46	1.1438
MMAudio	0.0849	0.1138	0.3747	0.6424	0.4471	36.30	0.7158
Hunyuan-Foley	0.1895	0.2043	0.3040	0.8083	0.5784	28.85	0.6627
Omni2Sound	0.1267	0.1497	0.3530	0.6896	0.4997	26.12	**0.4523**
ControlFoley	<u>0.2084</u>	<u>0.2353</u>	0.2826	<u>0.8219</u>	<u>0.6212</u>	**23.35**	0.5426
ThinkSound-Light	0.1538	0.1834	0.2895	0.7510	0.5382	30.45	<u>0.4738</u>
Mel-QCD	0.1262	0.1464	0.2880	0.6462	0.4617	41.30	1.2143
SelVA	0.1838	0.2177	<u>0.2449</u>	0.8154	0.6130	37.70	0.8504
FocusV2A	**0.2757**	**0.3441**	**0.1839**	**0.9418**	**0.7635**	<u>24.75</u>	0.6422

5 Experiments

5.1 Experimental Setup

Training and evaluation. We train on 176,289 VGGSound clips (Chen et al., 2020), using 75% mixed and 25% unmixed pairs. The two stages run for 25,000 and 12,500 updates, respectively. Inference generates eight-second audio with 25 Euler steps and classifier-free guidance of 3.5. We evaluate on VGG-SourcePair and the inter-class and intra-class splits of VGG-MonoAudio (Lee et al., 2026). Appendix C lists model hyperparameters, optimization settings, and trainable parameter counts.

Our comparisons include text-only controls, the pretrained backbone, general V2A systems, and a selective-generation model. AudioLDM2 and Tango assess source correspondence from text alone (Liu et al., 2024; Ghosal et al., 2023); MMAudio is the backbone baseline (Cheng et al., 2025); and SelVA is the closest task-specific comparison (Lee et al., 2026). Hunyuan-Foley, Omni2Sound, and ControlFoley broaden the comparison across generation quality and synchronization (Shan et al., 2025; Dai et al., 2026; Yang et al., 2026). All video-conditioned methods receive the silent video and target prompt. For the main VGG-SourcePair comparisons, frames are resized with their aspect ratio preserved and padded to each model's input resolution, retaining both source views. VGG-MonoAudio baseline scores are taken from SelVA, and FocusV2A uses the official evaluator.

Metrics. TextSep is the CLAP similarity to the target description minus that to the competing description (Wu et al., 2023). RefSep uses the corresponding component recordings:

$$\mathrm{RefSep} = \cos\big(e_a(\hat{a}), e_a(a_t^{\mathrm{ref}})\big) - \cos\big(e_a(\hat{a}), e_a(a_d^{\mathrm{ref}})\big)\,, \tag{6}$$

where e_a is the normalized CLAP audio embedding and t, d denote the target and competing sources. Temporal metrics use one-second windows with a 0.5-second hop. For binary target and competing-source detections t_w and n_w, Target@Dur and Leak@Dur report the fractions of detector-positive windows for the two categories. TempSuccess is the fraction of windows in which the target is detected and the competing category is not:

$$\mathrm{TempSuccess} = \frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}} t_w(1-n_w). \tag{7}$$

Detections compare PANN embeddings (Kong et al., 2020) with category prototypes using fixed category-specific thresholds. For each of the 48 categories, four VGGSound training clips construct the prototype and four separate clips calibrate its threshold; all 384 clips are disjoint from the test set at both clip and parent-video levels. FAD (Kilgour et al., 2019) measures distributional audio distance with PANN embeddings, and DeSync measures temporal alignment between generated audio

and the composite video. VGG-MonoAudio uses its standard metric suite. Confidence intervals use 10,000 paired bootstrap replicates over mixed videos, keeping both requests together. Appendix B specifies the detector, calibration, and evaluation settings.

5.2 Comparison with Existing Methods

Quantitative comparison. FocusV2A achieves the best source-selection scores among the evaluated video-conditioned methods on VGG-SourcePair (Table 1). Relative to SelVA, TempSuccess rises from 61.30% to 76.35% and RefSep from 0.2177 to 0.3441. The same trend appears across both metric families: CLAP-based margins favor the target more strongly, while the temporal metrics show higher target coverage and lower leakage. FocusV2A also achieves the second-lowest FAD, combining stronger source selectivity with competitive distributional audio quality.

The baseline results also show why target coverage and leakage should be considered together. ControlFoley and Omni2Sound lead in FAD and DeSync, respectively, but yield higher competing-source detection rates. Tango has relatively low leakage, but its target coverage is also limited. AudioLDM2 shows stronger source correspondence but weaker audiovisual alignment. Among the video-conditioned methods, FocusV2A achieves the highest target coverage and the lowest leakage. Low leakage alone is therefore not enough: a model should suppress the competing source while still preserving the requested one.

Table 2: VGG-MonoAudio results on 560 inter-class and 511 intra-class tasks. Baselines are reported by SelVA (Lee et al., 2026); FocusV2A uses the official evaluator. Best and second-best scores are highlighted within each split.

	Audio Quality			Semantic Alignment			Temporal Alignment
Model	FAD↓	KAD↓	IS↑	KL↓	CLAP↑	IB↑	DeSync↓
Inter-class							
ReWaS	70.4	4.937	6.23	2.57	0.200	0.2454	1.364
VinTAGe	<u>50.5</u>	1.309	11.51	**1.69**	0.283	0.2850	1.292
MMAudio-S-16k	56.7	0.874	11.54	2.07	0.270	0.3135	0.802
VOS+MMAudio	60.0	0.878	12.11	1.91	0.291	0.3010	0.991
SelVA	51.7	<u>0.676</u>	<u>13.07</u>	<u>1.85</u>	<u>0.292</u>	<u>0.3251</u>	<u>0.721</u>
FocusV2A	**49.4**	**0.561**	**13.48**	**1.69**	**0.336**	**0.3451**	**0.631**
Intra-class							
ReWaS	57.4	3.148	6.29	1.97	0.220	0.2569	1.377
VinTAGe	<u>37.0</u>	0.690	9.28	**0.88**	0.277	0.2892	1.304
MMAudio-S-16k	41.5	0.654	9.00	1.09	0.276	<u>0.3248</u>	0.670
VOS+MMAudio	43.4	0.656	8.91	1.11	**0.287**	0.3087	0.904
SelVA	<u>37.0</u>	<u>0.492</u>	<u>9.62</u>	1.04	0.280	**0.3262**	<u>0.639</u>
FocusV2A	**34.3**	**0.453**	**9.71**	<u>1.02</u>	<u>0.285</u>	0.3177	**0.346**

VGG-MonoAudio extends the comparison to an established selective-generation protocol (Table 2). On inter-class examples, FocusV2A improves both source correspondence and timing relative to SelVA: CLAP rises from 0.292 to 0.336 and DeSync falls from 0.721 to 0.631, alongside lower FAD. On intra-class examples, the clearest gains are in audio quality and synchronization, with FAD decreasing from 37.0 to 34.3 and DeSync from 0.639 to 0.346. Quality and timing improvements recur across both splits, complementing the target-retention and leakage analysis on VGG-SourcePair.

We examine the response of the source-level metrics using controlled mixtures of the component recordings (Appendix B, Figure 6). Increasing the competing source's level reduces RefSep, while extending its retained waveform interval increases Leak@Dur and lowers TempSuccess. These trends show sensitivity to changes in source balance and detected competing-source activity.

Qualitative and perceptual evaluation. Figure 3 shows how FocusV2A changes the generated sound structure with the request. For typewriter/hair dryer, its outputs contrast separated typewriter transients with a more continuous hair-dryer texture. For guitar/chainsaw, the guitar output contains changing harmonic and transient patterns, whereas the chainsaw output has sustained spectral energy. These differences follow the temporal and spectral structure of the corresponding component recordings. The paired requests make the contrast visible within each scene.

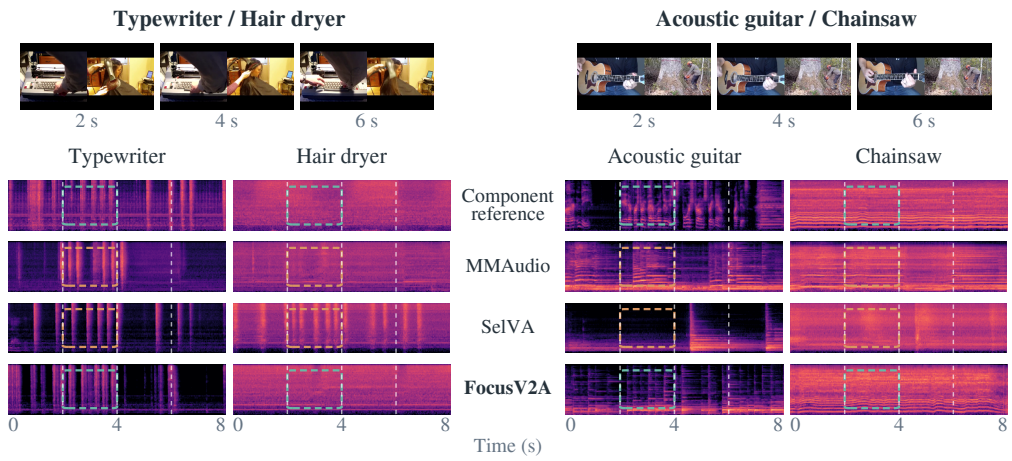


Figure 3: Paired spectrograms for typewriter/hair dryer and guitar/chainsaw, with target-component references. All use shared time and $[-90, 0]$ dB scales; boxes mark comparison regions.

Thirty participants rate 72 outputs from MMAudio, SelVA, and FocusV2A on 12 mixed videos, with 15 ratings per output (Figure 4). They assess target presence, non-target leakage, and audio quality on a 1–5 scale, plus perceived target and leakage duration on a 0–5 scale. The ratings give a consistent perceptual picture: the requested sound is clearer, the competitor is less audible, and overall quality is higher for FocusV2A. Relative to SelVA, target presence rises from 3.84 to 4.36 and leakage falls from 1.86 to 1.14. The duration ratings add a temporal perspective, with longer perceived target presence and shorter perceived leakage. Appendix D describes the study protocol.

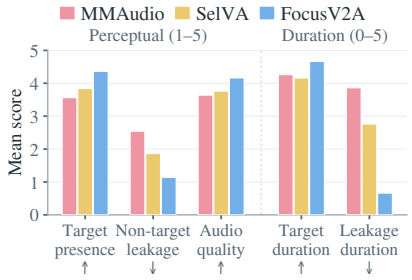


Figure 4: Human ratings for presence, leakage, and quality (1–5), and perceived duration (0–5).

5.3 Ablation Studies

Table 3 compares visual selection, generator adaptation, and competitive assignment. The w/o Focus Selector variant adapts the generator without visual selection, w/o Generator Adaptation combines the Stage-1 selector with the pretrained generator, and w/o Competition replaces target–complement softmax with independent sigmoid scores while retaining both training stages.

Table 3 distinguishes the effects of visual selection and generator adaptation. With the generator frozen, Stage 1 already reaches a RefSep of 0.3414, close to the full model's 0.3441. Generator adaptation then lowers FAD from 28.30 to 24.75 while keeping source correspondence at a similar level. Compared with generator-only adaptation, the full model achieves higher RefSep and lower leakage and FAD; the generator-only variant retains higher TempSuccess.

Competitive assignment increases RefSep and TempSuccess while reducing leakage relative to independent relevance scores, with nearly unchanged FAD.

5.4 Visual Selection Analysis

Figure 5(a) shows input frames and attention from the spatial aggregation layer under the two source requests. Switching the prompt shifts attention toward the corresponding source region. Figure 5(b) reports the fraction of spatial-key attention assigned to the target half of the encoder input. The examples and full-benchmark attention summary use full-frame letterboxing, which retains both source views. Across 512 requests, matched selection assigns 72.21% to the target half, compared with 27.79% for opposite selection.

Table 3: Component ablations on VGG-SourcePair. Best and second-best scores are highlighted; variant definitions are given in the text.

	Source Identity		Temporal Selection			Quality	Sync
Method	TextSep↑	RefSep↑	Leak@ Dur↓	Target@ Dur↑	Temp Success↑	FAD↓	DeSync↓
w/o Focus Selector	0.2634	0.3324	0.2025	0.9184	**0.7764**	30.10	0.7201
w/o Generator Adaptation	0.2734	<u>0.3414</u>	<u>0.1867</u>	0.9284	0.7632	28.30	**0.5879**
w/o Competition	<u>0.2744</u>	0.3400	0.1952	<u>0.9365</u>	0.7573	**24.72**	0.6771
FocusV2A	**0.2757**	**0.3441**	**0.1839**	**0.9418**	<u>0.7635</u>	<u>24.75</u>	<u>0.6422</u>

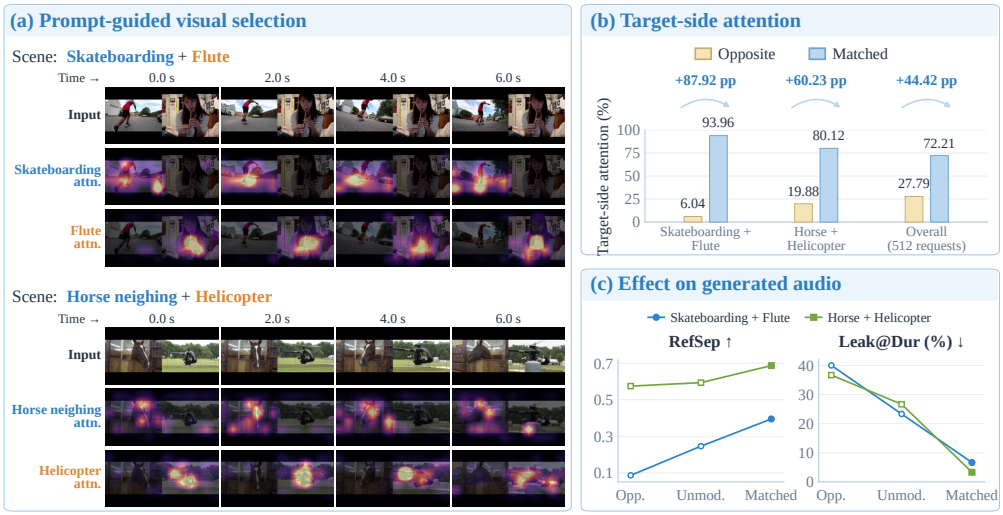


Figure 5: Visual selection and its effect on generated audio. (a) Input frames and spatial attention for two separate requests at four times. (b) Target-panel attention for the two cases and all 512 requests. (c) Case-level RefSep and Leak@Dur under opposite, unmodulated, and matched selection.

We next compare audio generated under three visual selection conditions while fixing the checkpoint, video preprocessing, native generator text, and initial noise. Unmodulated attention retains the pretrained aggregation scores with $b_i^t = 0$; matched and opposite selection use prompts for the requested and competing sources, respectively. In both displayed cases, matched selection has higher RefSep and lower Leak@Dur than unmodulated attention (Figure 5c). Target-matched visual evidence thus improves these outputs relative to the model's own unmodulated aggregation. Table 9 in Appendix E reports the full-benchmark comparison, with mean RefSep increasing from 0.2774 under opposite selection to 0.3441 under matched selection. The generated source content thus changes with the selected visual evidence while the generator's text condition remains fixed.

6 Conclusion

We introduced VGG-SourcePair and FocusV2A for selective video-to-audio generation. Paired requests and component references measure target retention and non-target leakage together, distinguishing source selection from target matching. FocusV2A uses competitive local assignments to guide frozen visual aggregation, then adapts the generator to the resulting condition. On VGG-SourcePair, it improves source selection and receives higher target-presence and quality ratings with lower perceived leakage. VGG-MonoAudio results show competitive audio quality and synchronization. Fixed-model comparisons link changes in selected visual evidence to generated source content while the generator's text condition remains fixed. Together, these results support prompt-guided control over which visible source contributes to a generated soundtrack.

AI USE STATEMENT

We used ChatGPT (OpenAI) to assist with literature retrieval and discovery, research ideation, selected aspects of research execution, and the writing and editing of portions of the research code. We also used ChatGPT to improve the clarity, grammar, and readability of the manuscript. All literature identified with AI assistance was independently checked against the original sources. AI-assisted research ideas were critically evaluated by the authors, and all AI-assisted code was reviewed and tested for correctness before being used in the reported experiments. All AI-assisted writing was reviewed and revised by the authors. The authors take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

REFERENCES

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18770–18781. IEEE, 2025.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Qiuhong Ke, Jianfei Cai, and Jun Zhu. Omni2sound: Towards unified video-text-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1661–1671, 2026.

Pengjun Fang, Yingqing He, Yazhou Xing, Qifeng Chen, Ser-Nam Lim, and Harry Yang. Acfoley: Reference-audio-guided video-to-audio synthesis with acoustic transfer. *arXiv preprint arXiv:2603.15597*, 2026.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 3590–3598, 2023.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025.

K Kilgour, M Zuluaga, D Roblek, and M Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.

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.

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

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.

Tingle Li, Siddharth Gururani, Kevin J Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. *arXiv preprint arXiv:2605.30339*, 2026a.

You Li, Dewei Zhou, Fan Ma, Fu Li, Dongliang He, and Yi Yang. Foleydirector: Fine-grained temporal steering for video-to-audio generation via structured scripts. *arXiv preprint arXiv:2603.19857*, 2026b.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal llms for audio generation and editing. *Advances in Neural Information Processing Systems*, 38:78526–78552, 2026a.

Huadai Liu, Kaicheng Luo, Wen Wang, Qian Chen, Peiwen Sun, Rongjie Huang, Xiangang Li, Jieping Ye, and Wei Xue. PrismAudio: Decomposed chain-of-thought and multi-dimensional rewards for video-to-audio generation. In *International Conference on Learning Representations*, volume 2026, pp. 25176–25204, 2026b.

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

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

Juncheng Wang, Chao Xu, Cheng Yu, Lei Shang, Zhe Hu, Shujun Wang, and Liefeng Bo. Synchronized video-to-audio generation via mel quantization-continuum decomposition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3111–3120. IEEE, 2025.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

A Benchmark Construction

Component screening. The candidate pool contains 960 VGGSound clips, with 20 candidates from each of 48 event categories. Candidates are deduplicated by parent video. Audio checks use the first eight seconds converted to mono at 16 kHz: RMS must exceed -35 dBFS, the fraction of samples with absolute amplitude below 10^{-3} must be less than 0.25, and the fraction with absolute amplitude at least 0.99 must be less than 0.01. Video checks require a minimum spatial dimension of 224 pixels and a duration of at least 7.85 seconds. Mean frame–label CLIP similarity (Radford et al., 2021) ranks candidates within each category. The final 512 clips pass these automatic checks and come from distinct parent videos in the official test split. Their original soundtracks are retained as component references.

Composition and references. Both source videos are trimmed to their first eight seconds and reset to a common starting time. Each is resized with its aspect ratio preserved, centrally padded to 384×384, and concatenated horizontally to form a 768×384 video at 25 fps. The mixed video contains a silent audio track. The two component references retain the source waveforms over the same interval, converted to mono 16 kHz PCM; they are not outputs of a source-separation model. The prompts are label-derived event descriptions without spatial side information. Each mixed video is used for both requests without re-composition.

Pairing strata. VGG-SourcePair contains 256 mixed-video clusters and 512 directed tasks. Easy cross-superclass pairs contribute 96 videos/192 tasks, hard cross-superclass pairs 64 videos/128 tasks, and hard same-superclass pairs 96 videos/192 tasks. The latter contain 16 pairs from each of the six superclasses, always with different event labels. The strata are defined by category-pairing templates. Each of the 512 source clips participates in one pair and therefore appears once as a target and once as a non-target. Bootstrap resampling retains both requests in each mixed-video cluster.

Evaluation scope. The experiments concern controlled two-source compositions and one generation backbone. Component references retain background audio, and source-activity scores are produced by a fixed pretrained detector.

Table 4: The 48 event categories and 512 source clips in VGG-SourcePair. Each clip appears once as the target and once as the competing source; counts are source clips, not mixed videos.

Superclass	Event categories (source clips per category)
Animal vocalization (84)	Bird chirping/tweeting (11), cat meowing (10), cat purring (11), dog barking (10), dog growling (10), duck quacking (11), horse neighing (11), sheep bleating (10).
Human non-speech action (80)	Baby crying, people cheering, clapping, coughing, eating, screaming, sneezing, and whistling (10 each).
Impact / physical action (80)	Basketball bounce, door slamming, fire crackling, lighting firecrackers, opening/closing drawers, skateboarding, splashing water, and toilet flushing (10 each).
Instrument / music (84)	Acoustic guitar (11), bass guitar (11), cello (10), drum kit (10), electric guitar (11), flute (10), piano (10), violin/fiddle (11).
Tool / machine / appliance (92)	Chainsawing trees (12), hair dryer drying (12), lawn mowing (12), printer printing (11), computer keyboard typing (11), typewriter typing (11), sewing machines (11), vacuum cleaner (12).
Vehicle / engine / siren (92)	Car engine starting (11), car passing by (11), motorcycle (12), engine accelerating/revving (12), helicopter (12), police siren (11), race car (12), train horn (11).

B Evaluation and Baseline Settings

The broader comparison also includes ThinkSound-Light (Liu et al., 2026a) and Mel-QCD (Wang et al., 2025). ThinkSound-Light denotes the `thinksound_light.ckpt` release in the official model repository.

Threshold fitting. Calibration uses 384 clips from the official VGGSound training split, with eight clips per benchmark category. Four clips per category establish its PANN prototype and four separate

clips provide threshold-calibration windows. These 192 threshold clips yield 2,880 windows: 60 positive and 2,820 other-category negatives per category. Each calibration window inherits the event label of its source clip. Thresholds maximize development true-positive rate minus false-positive rate and remain fixed for every method. The development and test sources have no clip or parent-video overlap; generated audio and test references are not used to fit thresholds.

Window-level evaluation. We use the 2,048-dimensional embedding of PANN Cnn14 (Kong et al., 2020) at 16 kHz (512-sample analysis window, 160-sample hop, 64 mel bins over 50–8,000 Hz). Each embedding is ℓ_2-normalized; a category prototype averages its four normalized clip embeddings and is normalized again. A window is positive when its cosine similarity is greater than or equal to the category threshold. Audio is channel-averaged, resampled to 16 kHz, then zero-padded or truncated to 128,000 samples. Fifteen one-second windows are evaluated at 0.5-second hops. Every request contributes all 15 windows, including quiet and padded intervals. Scores are averaged over windows and then equally over requests. Scoring requires a complete audio and category mapping for all requests.

Audio quality and output energy. TextSep and RefSep use normalized CLAP embeddings. FAD (Kilgour et al., 2019) is computed with the AVBench implementation from unnormalized 2,048-dimensional PANN Cnn14 embeddings. Audio is channel-averaged, mean-centered, resampled to 16 kHz, and limited to eight seconds; shorter waveforms retain their decoded length. On VGG-SourcePair, FAD compares the generated outputs with the target component recordings from the same 512 requests. VGG-MonoAudio uses its official evaluator and split-specific reference sets. A separate clip-level RMS diagnostic uses a fixed cutoff of -32.002028 dBFS. It flags 16.99% of FocusV2A outputs and 17.97% of SelVA outputs as low-energy. The flag is recorded alongside the source-level scores; all outputs and all 15 evaluation windows remain in the temporal metrics.

For the RMS diagnostic, decoded audio is channel-averaged, resampled to 16 kHz, and truncated or zero-padded at the end to 128,000 samples, without mean subtraction or amplitude normalization. For these samples $a_1, \ldots, a_L$, where $L = 128{,}000$ includes padding and unit amplitude denotes digital full scale, the statistic is $E(a) = 20 \log_{10}\left(\max\left(\sqrt{L^{-1}\sum_{n=1}^{L} a_n^2}, 10^{-12}\right)\right)$. The cutoff was fixed using train-origin development controls: recordings RMS-normalized to -20 dBFS as positives, and their 24-dB-attenuated versions plus silence as negatives, maximizing the true-positive rate minus the false-positive rate. This cutoff is used only for the waveform-energy diagnostic.

DeSync evaluation. On VGG-SourcePair, DeSync compares each generated waveform with the mixed video, not the target component video. The evaluator reads the first eight seconds at 25 fps and applies full-frame letterboxing: the 768×384 composite is resized proportionally to 224×112 and padded by 56 pixels above and below to 224×224. Synchformer (Iashin et al., 2024) extracts 24 overlapping 16-frame segments with an eight-frame stride. The scorer compares the first 14 and last 14 audiovisual segments separately, predicts an offset on a 21-bin grid from -2 to 2 seconds, and averages the absolute predicted offsets from the two temporal windows over all requests. Lower DeSync indicates better audiovisual synchronization.

Controlled-mixture validation. Each full eight-second source is RMS-normalized to -20 dBFS before central-segment gating. Non-target level is relative to this full-length reference, and coverage is the central retained waveform fraction. No post-gate RMS normalization is applied. A common peak guard $\min(1, 0.98/\text{peak})$ rescales the mixture when needed. At zero coverage, the mixture contains only the target component recording.

Figure 6(a) varies the non-target level from -12 to $+12$ dB in 0.5-dB steps at 25%, 50%, and 100% coverage. Figures 6(b–c) vary coverage from 0% to 100% in one-percentage-point steps at -6, 0, and $+6$ dB; panel (c) uses the 0-dB slice. The union contains 441 distinct settings, each evaluated on all 512 requests. The curves connect measured values without fitting or smoothing. Each bootstrap replicate resamples mixed videos, retaining both directions and averaging within cluster. The 10,000 shared replicates use seed 260620; percentile intervals are pointwise rather than simultaneous bands. Collection-level distances are not assigned per-task intervals.

Figure 6 tests sensitivity to controlled changes in a competing recording. Raising its level reduces RefSep, while extending its coverage increases Leak@Dur and reduces TempSuccess. On the 0-dB

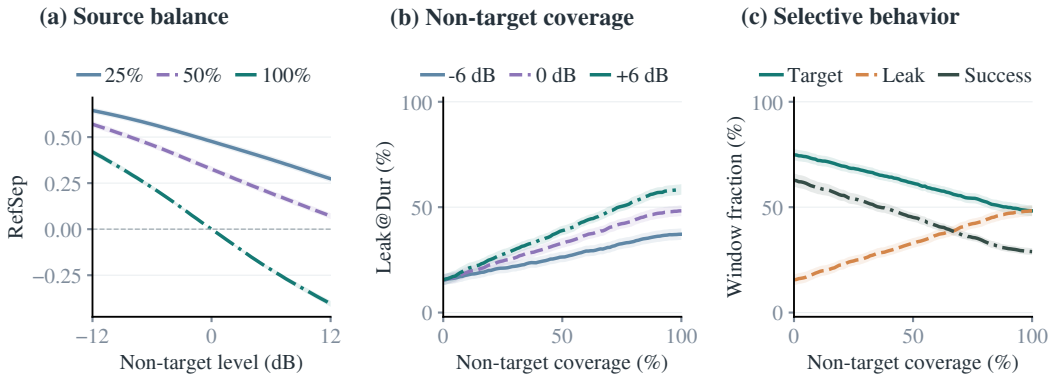


Figure 6: Sensitivity to controlled source changes. (a) RefSep under different competing-source levels and coverage settings. (b) Leak@Dur as competing-source coverage increases. (c) Target@Dur, Leak@Dur, and TempSuccess at 0 dB. Bands are pointwise 95% cluster-bootstrap intervals.

slice, Leak@Dur rises from 15.55% with no competing waveform added to 48.27% with the full waveform. Coverage describes the retained waveform fraction; Leak@Dur describes the fraction of windows classified as the competing category. The zero-coverage value records the detector's response to the target component alone. Background sounds and category confusions can contribute to this response, so the scores are reported as unadjusted detection rates.

Generation settings and preprocessing. For the VGG-SourcePair main comparison and ablations (Tables 1 and 3), all video-conditioned methods use full-frame letterboxing, preserving aspect ratio and both source views at each model's input resolution. Text-only controls receive no video. The human-evaluation waveforms in Figure 4 follow the same generation protocol as Table 1. In the qualitative examples of Figure 3, MMAudio, SelVA, and FocusV2A all use full-frame letterboxing.

All attention and audio comparisons in Figure 5, and the aggregate selection-path comparison in Table 9, use full-frame letterboxing. For the 224×224 spatial encoder, the 768×384 composite is resized proportionally to 224×112 and padded by 56 pixels above and below. This differs from the half-width composition used during training. Table 5 lists the qualitative and human-generation settings; all three methods use 25 Euler steps. Window-level evaluation pads short decoded audio to eight seconds, whereas spectrograms leave the ungenerated tail blank. VGG-MonoAudio baselines are reported by SelVA (Lee et al., 2026); FocusV2A uses the official evaluator.

Table 5: Generation settings for qualitative and human evaluation. All three methods use 25 Euler steps to generate eight-second outputs.

Method	Model release / checkpoint	Generation
MMAudio	Small 16k pretrained generator	25 Euler steps; CFG 3.5
SelVA	Small 16k selective generator and video encoder	25 Euler steps; CFG 3.5
FocusV2A	Stage-1 final selector; Stage-2 final EMA	25 Euler steps; CFG 3.5

C Architecture and Training

Data and optimization. The training inventory contains 176,289 VGGSound source clips, excluding benchmark sources at clip and parent-video levels. Stage 1 schedules 100,000 pairs using 108,255 distinct clips: 75,000 mixed and 25,000 unmixed pairs. Stage 2 schedules 150,000 pairs using 131,278 distinct clips: 112,500 mixed and 37,500 unmixed pairs. Both directions provide supervision for every pair. Table 6 summarizes optimization.

Table 6: Optimization settings. Stage 2 uses the final step-12,500 EMA for inference.

Setting	Stage 1	Stage 2
Updates	25,000	12,500
Effective batch (directed requests)	8	24
Learning rate	10^{-4}	10^{-5}
Warmup steps	1,000	1,000
Trainable parameters	15,500,288	22,058,944
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$	
Numerical settings	$\epsilon = 10^{-6}$, weight decay 10^{-6}, gradient clip 1.0	
Precision	bfloat16	

Training preprocessing. Both stages read the first eight seconds of each source at 25 fps. Frames are resized proportionally to a 224-pixel short side, center-cropped to 224×224, and normalized to $[-1, 1]$. For mixed pairs, each square crop is resized to 112×224, then concatenated with its partner to form a 224×224 frame, with randomized left/right order. Unmixed inputs and component teachers retain their individual 224×224 views. Stage-2 audio supervision uses the requested source's waveform, converted to mono at 16 kHz, rather than a mixture of both sources.

Model dimensions and updated modules. Synchformer exposes features of shape $[B, S, 8, 14, 14, 768]$. The selector uses width 512, four target and four complement queries, two decoder layers, eight heads, and a 4,096-dimensional feed-forward block. FLAN-T5-base remains frozen. We use assignment temperature $\tau = 0.5$, selection-bias scale $\beta = 2.0$, and assignment floor $\varepsilon_m = 10^{-4}$. The key bias is added after input normalization during both training and inference. Spatial and temporal aggregators remain frozen, producing an eight-second synchronization sequence of shape $[B, 192, 768]$.

Stage 2 updates only the synchronization-input projection and positional embedding, latent AdaLN modulation in joint blocks 0–3, and AdaLN modulation in fused blocks 0–7. The selector, visual backbone, aggregators, and text encoders remain frozen. Inference uses 25 Euler steps and CFG 3.5. Both Full and the generator-only variant optimize conditional flow matching. Native text dropout is 0.1 for all Stage-2 training variants. The Stage-1-only variant retains the pretrained generator; independent relevance replaces the competitive assignment by sigmoid scores.

C.1 Stage 1 Objective

For pair j, let $F_j^A, F_j^B \in \mathbb{R}^{T \times d}$ be the selected synchronization sequences for requests A and B, and let H_j^A, H_j^B be the frozen unmixed-component teacher sequences. Define sequence similarity as the mean of token-wise cosine similarities,

$$s(U, V) = \frac{1}{T} \sum_{t=1}^{T} \cos(U_t, V_t), \qquad \cos(u, v) = \frac{u^\top v}{\max(\|u\|_2, \epsilon) \max(\|v\|_2, \epsilon)}, \quad \epsilon = 10^{-8}. \tag{8}$$

The cosine operations use float32. For a batch of B pairs, the alignment term sums the alignment penalties for both directions and averages over pairs:

$$\mathcal{L}_{\text{align}} = \frac{1}{B} \sum_{j=1}^{B} \left[2 - s(F_j^A, H_j^A) - s(F_j^B, H_j^B) \right]. \tag{9}$$

Let $q_j = 1$ for a mixed-scene training pair and 0 for an unmixed-source pair, and let $[x]_+ = \max(x, 0)$. The separation term uses a relative-similarity margin $\mu = 0.2$:

$$\begin{aligned} \mathcal{L}_{\text{sep}} = \frac{1}{2B} \sum_{j=1}^{B} q_j \big(& [\mu - s(F_j^A, H_j^A) + s(F_j^A, H_j^B)]_+ \\ & + [\mu - s(F_j^B, H_j^B) + s(F_j^B, H_j^A)]_+ \big). \end{aligned} \tag{10}$$

For the paired regularizer, flatten all spatiotemporal target-assignment entries into $m_j^A, m_j^B \in [0,1]^P$ and center them as $\widetilde{m}_j^u = m_j^u - \frac{1}{2}\mathbf{1}$. The objective combines a complementarity penalty and an anti-correlation penalty:

$$\mathcal{R}_{\text{pair}} = \frac{1}{2B}\sum_{j=1}^{B} q_j \left[\frac{\|m_j^A + m_j^B - \mathbf{1}\|_1}{P} + 1 + \cos(\widetilde{m}_j^A, \widetilde{m}_j^B) \right]. \tag{11}$$

Equation 4 uses $\lambda_{\text{sep}} = 0.5$ and $\lambda_{\text{pair}} = 0.2$. Unmixed pairs contribute only $\mathcal{L}_{\text{align}}$. The paired regularizer encourages complementary target assignments across the two requests, while Equation 2 normalizes target and complement assignments within each request. Teacher sequences and encoders remain frozen.

D Human Evaluation

Stimuli and allocation. The study uses 12 mixed videos selected for source visibility across the three pairing strata. Both requests are evaluated for MMAudio, SelVA, and FocusV2A, yielding 72 outputs. Scenes were selected using visual information, without comparing generated audio or automatic scores. Participants hear the original generated waveforms alongside silent video, without gain adjustment, denoising, or duration matching.

Thirty participants each rated 36 outputs: all three methods for one request direction per video. Fifteen six-left/six-right schedules and their complements balance directions, so every output receives 15 ratings. The recorded assignment seed is 260920. The interface hides method names behind anonymous media IDs; trial order uses three hidden rounds, with each scene occurring once per round and shuffled model order within each scene. At least three trials separate repeats of a scene.

Questions and procedure. The interface identifies the requested source and the other visible source. Participants are instructed to use headphones in a quiet environment, keep a comfortable consistent volume, and rate each output independently. Replay, pauses, and withdrawal are allowed. The study includes three perceptual items and two perceived-duration items. The perceptual items rate target presence, competing-source audibility, and audio quality on a 1–5 scale (Table 7); ratings 2 and 4 are intermediate anchors. The duration items rate the portion of the clip during which each source is audible, from 0 (not audible at any time) to 5 (audible for almost the entire clip). All five items were collected in the same listening session for each output. These are subjective temporal-coverage ratings, not durations in seconds.

Table 7: Human-rating questions (English translation). Source names in brackets are replaced according to the video and request. Lower leakage scores indicate less competing-source audibility.

Dimension	Question and anchors
Target presence	How clearly can you hear the requested sound? 1: not at all; 3: moderately audible; 5: very clear.
Non-target leakage	How clearly can you hear the other visible sound? 1: not at all; 3: moderately audible; 5: very clear.
Audio quality	How would you rate the overall quality of the generated audio? 1: very poor; 3: fair; 5: excellent.
Target duration	During what proportion of this audio can you hear [target source]?
Leakage duration	During what proportion of this audio can you hear [non-target source]?
Duration anchors	0: not audible at any time; 1: audible for only a small portion; 2: audible for less than half; 3: audible for about half; 4: audible for most of the clip; 5: audible for almost the entire clip.

Aggregation. For each item, we average ratings within each output and then equally across the 24 outputs per method. Every output has 15 ratings, and the 24 outputs cover both requests from 12 mixed videos. Figure 4 and Table 8 report these means.

Table 8: Mean human ratings. Perceptual items use a 1–5 scale; perceived-duration items use a 0–5 scale.

Method	Target presence	Non-target leakage	Audio quality	Perceived target duration	Perceived leakage duration
MMAudio	3.56	2.54	3.64	4.26	3.86
SelVA	3.84	1.86	3.76	4.16	2.76
FocusV2A	4.36	1.14	4.16	4.66	0.66

E Additional Results and Visualization

Matched and opposite selection. Table 9 compares matched and opposite selection over all 512 requests using the main letterbox protocol. The matched condition reuses the FocusV2A outputs in Table 1. Within each request, the checkpoint, input video and preprocessing, native generator text, and initial noise are fixed. Figure 5(c) additionally includes unmodulated attention for the two displayed cases, averaging both request directions. All differences are calculated from unrounded scores and rounded only for display.

Table 9: Aggregate selection-path comparison. RefSep is in raw score units, with its 95% confidence interval in brackets; temporal differences are in percentage points.

Metric	Matched	Opposite	Matched − Opposite
RefSep	0.3441	0.2774	$+0.0667$ $[0.045, 0.089]$
Leak@Dur	18.39%	21.55%	-3.16 pp
Target@Dur	94.18%	89.26%	+4.92 pp
TempSuccess	76.35%	70.42%	+5.94 pp

Assignments and aggregation. Figure 7 separates the selector's target assignments from the attention weights used to aggregate visual features. Both maps are obtained from the same letterboxed input and temporal position. The assignments express prompt-dependent relevance, while spatial attention combines these assignments with the pretrained query–key scores (Equation 3). The resulting aggregation therefore reflects both the request and the visual content.

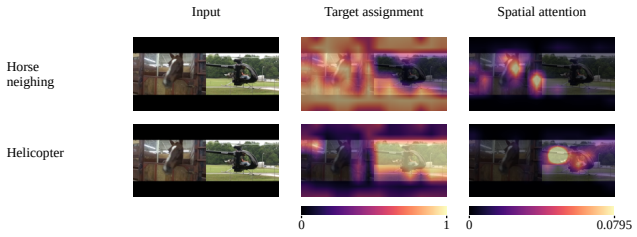


Figure 7: Selector target assignments and spatial aggregation attention for two requests on the same letterboxed input. Assignments use a $[0, 1]$ scale; attention is normalized over spatial keys.

Source-panel attention. Let $A_{\tau,h,i}$ denote post-softmax attention from the aggregation query to spatial key i at temporal position τ and head h. For a given request, let $\mathcal{P}_{\text{tar}} \subset \mathcal{P}$ contain the keys in the target half of the letterboxed input. We compute $\sum_{i\in\mathcal{P}_{\text{tar}}} A_{\tau,h,i} / \sum_{i\in\mathcal{P}} A_{\tau,h,i}$, then average over heads, temporal positions, and requests. $\mathcal{P}$ includes all spatial keys, including padding, and excludes the aggregation-token key. Across 512 requests, Matched and Opposite means are 72.21% and 27.79%, with a paired difference of 44.42 percentage points and a 95% mixed-video cluster interval of $[40.25, 48.41]$. Averaging a request-invariant map over both directions gives 50%, without requiring spatially uniform attention.

Spectrogram settings. Spectrograms use 16-kHz mono audio, a 1,024-sample periodic Hann window, a 160-sample hop, and 80 unit-peak HTK mel filters over 0–8 kHz. References and outputs share the first eight seconds and $[-90, 0]$ dB scale, with unchanged waveform amplitudes. Reflect boundaries are used, and one-sided FFT power is divided by $N \sum_n w_n^2$ before conversion by $10 \log_{10}(P/1)$. Time guides and boxes are visual annotations, not automatic event boundaries.

Attention display. We extract post-softmax attention from the aggregation query of `encoder.extractor.spatial_attn_agg.self_attn`, average the 12 heads, and normalize over the 196 spatial keys. Query 0 corresponds to the aggregation token, and patch keys follow row-major order on the 14×14 grid. The 768×384 composite is resized to 224×112 and padded by 56 pixels above and below. Figure 5(a) uses display times of 0, 2, 4, and 6 seconds (frames 0, 50, 100, and 150), with zero-based (segment, token) indices $(0, 0)$, $(5, 5)$, $(11, 6)$, and $(17, 7)$. Figure 7 uses the same geometry at 4 seconds. Display rows 56–167 retain the composite and omit only the outer encoder padding.

Figures 5 and 7 use a shared attention color range of $[0, 0.07949615]$, fixed from the 99.5th percentile of an earlier diagnostic pool and retained unchanged for the letterboxed maps. Overlays use magma colors, bilinear upsampling, and opacity 0.62. Target assignments in Figure 7 use $[0, 1]$. No per-image rescaling, target-mask multiplication, or additional smoothing is applied.

Window-level source analysis. Figure 8 aligns generated spectrograms with target and competing-source detections for the hair-dryer/printer scene. Some windows contain both detections, including in FocusV2A outputs. These windows contribute to Target@Dur but not TempSuccess, illustrating the distinction between target retention and source selectivity.

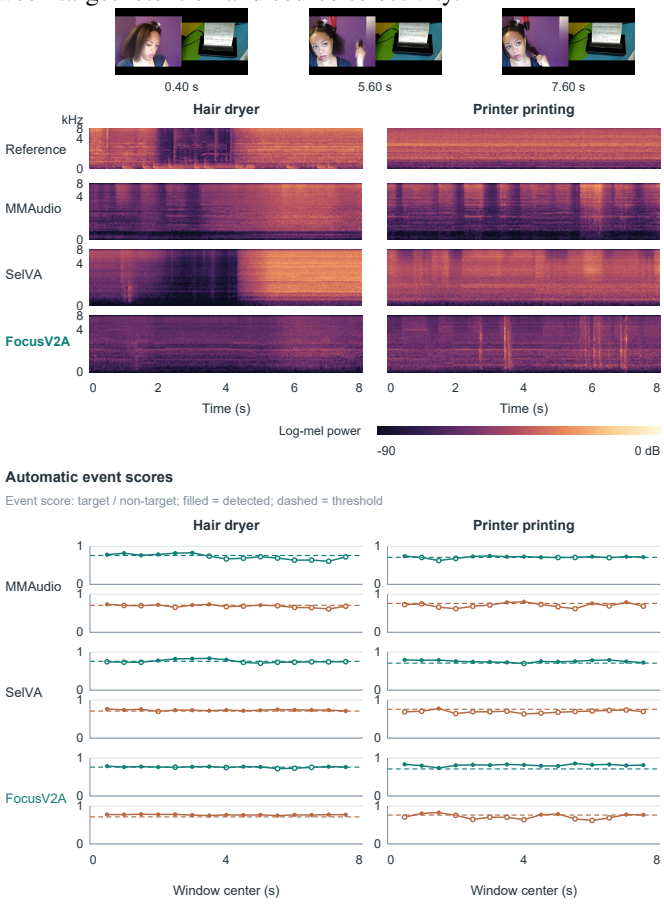


Figure 8: Hair-dryer/printer example with residual competing-source detections. Spectrograms and window-level scores share the same timeline. Dashed lines mark category thresholds, and filled markers indicate detections.