# SCoPE: Sparse Cross-modal Prior Exchange for training-free Audio-Visual Event Perception

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

## Abstract

Audio-Visual Event Perception (AVEP) identifies which events occur, when they occur, and whether they are audible, visible, or both. This distinction matters for sound sources outside the camera's field of view or visible soundless objects. Frozen text-aligned encoders enable training-free recognition with flexible event vocabularies, but events often co-occur and are selected by thresholding scores. When a related incorrect event label scores as high as a correct one, no single threshold can separate them. We analyze this false co-activation (FCA) through two types, *shadows* and *rivals*. To address this, we introduce SCoPE, a framework for Sparse Cross-modal Prior Exchange. SCoPE first makes event labels compete to explain the evidence within each modality. Across modalities, prior exchange adjusts event selection costs. Then, target re-selection fits each modality's own embedding under these costs. SCoPE improves event prediction across diverse AVEP benchmarks using the same hyperparameter settings. SCoPE also supports causal streaming for applications that need continuous predictions from incoming audio and video.

## 1 Introduction

Audio-Visual Event Perception (AVEP) (Tian et al., 2020; Gao et al., 2023; Zhou et al., 2024) not only identifies which events occur and when, but also determines whether each is audible, visible, or both. Knowing which modality provides evidence for each event gives a more complete understanding of the scene. This distinction is useful in applications such as robotics (Chen et al., 2021), autonomous driving (Chakravarthula et al., 2023), and smart glasses, where sound sources may lie outside the camera's field of view (FoV) (Huang et al., 2023) or visible objects may be soundless (Li et al., 2026). These applications require streaming event recognition without access to future inputs (Yu et al., 2025b). They also need to recognize new event classes as application requirements change (Zhou et al., 2025; Shaar et al., 2025; Yu et al., 2025a).

Supervised AVEP models (Tian et al., 2020; Cheng et al., 2022; Gao et al., 2023; Zhou et al., 2024; Lai et al., 2025) learn to recognize a predefined set of event classes, called an event vocabulary. Supporting a new class usually requires collecting labeled videos and retraining the model. Training-free AVEP (Zhou et al., 2025; Shaar et al., 2025) avoids this process by using frozen text-aligned encoders, such as CLIP (Radford et al., 2021) and CLAP (Wu et al., 2023). For each time segment, candidate events are scored by the similarity between text embeddings of event label names and the audio or visual features. Candidate events are ranked by these scores, and a threshold determines which events are selected when multiple events coexist. In this setting, introducing cross-modal interactions without task-specific training remains challenging. Existing training-free AVEP methods (Zhou et al., 2025; Shaar et al., 2025) score each modality independently and improve predictions by refining the decision rules applied to these scores. **Our key observation is that a single threshold across event labels cannot remove every false positive while retaining all ground-truth labels.** When an incorrect event label receives a score equal to or higher than that of a ground-truth event label, even an oracle choice of a single threshold cannot reject it while retaining all ground-truth labels. We call this score conflict a *false co-activation* (FCA; Definition 1).



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.

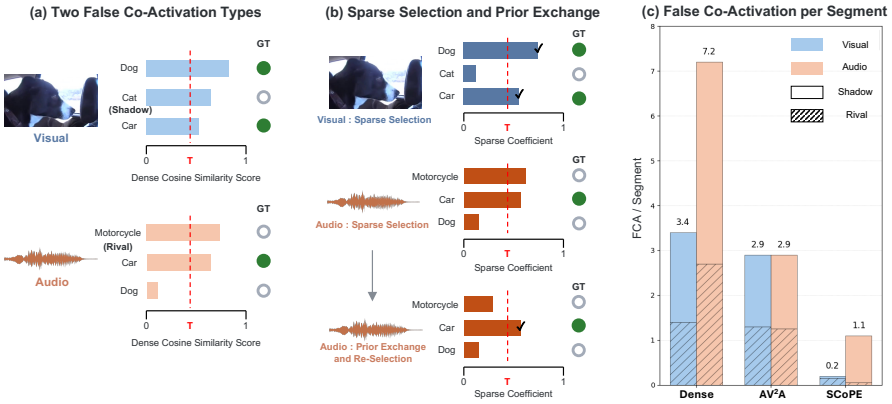



Figure 1: False co-activation and our approach to reducing it. (a) *Cat* is a shadow between two ground-truth visual events. *Motorcycle* is a rival of *Car* in audio. Both FCA types can occur in either modality. (b) Our approach. Sparse selection makes event labels compete within each modality and suppresses the visual shadow *Cat*. Prior exchange adjusts selection costs so that target re-selection can distinguish *Car* from its audio rival *Motorcycle*. (c) Mean FCA counts per multi-label segment on LLP. Each bar stacks shadow and rival counts, with the total height showing their combined count.

We divide FCAs into two types, and both types can occur in either the audio or visual modality. The first type, a ***shadow***, occurs when an incorrect label is sandwiched between a higher-scoring and a lower-scoring ground-truth label. In Figure 1(a), the video contains *Dog* and *Car*, but not *Cat*. The visual evidence supports *Dog*, but features shared by dogs and cats can raise the score of the incorrect label *Cat* above that of the ground-truth label *Car*. The second type, a ***rival***, occurs when a single modality cannot distinguish an incorrect label from a ground-truth label. From audio alone, a car engine can sound like a motorcycle, so *Motorcycle* remains plausible beside *Car*. *Car* and *Motorcycle*, however, are usually easy to distinguish visually. This motivates cross-modal guidance that helps distinguish competing events while preserving the target modality's role in deciding which events are present. Both types are formally defined in Definition 2. Figure 1(c) shows that shadows and rivals occur frequently under existing decision rules. AV$^2$A (Shaar et al., 2025) refines predictions through video-level candidate filtering and temporal interval re-evaluation, yet substantial FCAs remain.

We propose **SCoPE**, a training-free framework for **S**parse **C**ross-m**o**dal **P**rior **E**xchange. As illustrated in Figure 1(b), SCoPE first replaces independent similarity scoring with sparse selection. Event labels compete within each modality, reducing FCAs before cross-modal guidance. Cross-modal interaction should not simply transfer event scores from one modality to the other. For example, a visible dog may be silent, and an audible source may lie outside the camera's field of view. SCoPE exchanges cross-modal priors to adjust event-selection costs based on support from the other modality. Each target modality re-selects events using its own features, reducing the remaining FCAs. SCoPE's design is inspired by concept intervention in concept bottleneck models, where humans can correct the model's concept predictions to adjust its final predictions. In SCoPE, another modality provides guidance by adjusting event-selection costs without human intervention (Section 5.3). SCoPE also offers practical flexibility, supporting efficient causal streaming using only observed segments (Appendix B.1) and allowing integration with existing training-free decision rules (Appendix B.2).

Our contributions are as follows:

- We characterize false co-activation (FCA) through shadows and rivals, exposing a limitation of single-threshold refinement.
- We introduce SCoPE, which combines sparse prior exchange with target re-selection for training-free cross-modal guidance to reduce FCA.
- SCoPE supports causal streaming by predicting events from incoming audio and video without using future segments.
- SCoPE improves event prediction across AVEP benchmarks and frozen backbones using the same hyperparameter settings.

## 2 FALSE CO-ACTIVATION

Let a video be divided into $T$ aligned audio and visual segments, and let $\mathcal{Y}$ be the queried event vocabulary. For modality $m \in \{a, v\}$, $Y_m(t) \subseteq \mathcal{Y}$ is the ground-truth event set in segment $t$. A frozen modality encoder produces $z_m(t)$, and the corresponding frozen text encoder maps the candidate names to

$$C_m = [\mathbf{c}_m(1), \ldots, \mathbf{c}_m(|\mathcal{Y}|)].$$

Here, $\mathbf{c}_m(c)$ is the column of $C_m$ for event class $c$. Dense training-free inference scores each candidate independently as

$$s_m^{\cos}(t, c) = \cos(z_m(t), \mathbf{c}_m(c)).$$

Let $s_m^{\mathrm{dec}}(t, c)$ denote the decision score for candidate $c$ after any filtering or score refinement and before the method's final decision rule.

**Definition 1 (False Co-Activation).** Let $Y_m^{\mathrm{ret}}(t) \subseteq Y_m(t)$ denote the ground-truth labels that remain after any candidate filtering or support selection and before the final decision rule. Suppose $Y_m^{\mathrm{ret}}(t) \neq \varnothing$. A retained candidate $j$ is a false co-activation when

$$j \notin Y_m(t), \qquad s_m^{\mathrm{dec}}(t, j) \geq \min_{i \in Y_m^{\mathrm{ret}}(t)} s_m^{\mathrm{dec}}(t, i). \tag{1}$$

An FCA scores at least as high as the lowest-scoring label in $Y_m^{\mathrm{ret}}(t)$, so a scalar cutoff on the same values cannot remove it while preserving the entire set.

In a joint sparse fit (Lee et al., 2006; Mairal et al., 2010; Zhang et al., 2024), correlated labels no longer receive shared evidence separately. Once one label explains their common direction, another is selected only if its remaining contribution justifies an additional selection cost. We next analyze the smallest sparse problem that contains this competition, with one annotated event and one competing candidate.

**Lemma 1 (Two-Label Sparse Competition).** For either modality, consider a centered unit-norm embedding $z$ (Section 3.2) and two unit-norm text atoms, $c_i$ for an annotated event and $c_j$ for an unannotated candidate. Their two-label non-negative sparse fit is

$$w^* = \arg\min_{w \geq 0} \|z - [c_i, c_j]w\|_2^2 + \lambda_0 \|w\|_1. \tag{2}$$

Let $\rho = c_i^\top c_j$ with $0 \leq \rho < 1$, and order the atoms so that $c_i^\top z \geq c_j^\top z$. We restrict the analysis to non-negative correlation because the shadow/rival diagnostic concerns labels that share a positively aligned text direction. Negatively correlated atom pairs fall outside this diagnostic. Define $\gamma = \lambda_0/2$ and assume $c_i^\top z > \gamma$. The solution selects both atoms exactly when

$$\Delta_{j|i}(z) := (c_j - \rho c_i)^\top z - \gamma(1 - \rho) > 0. \tag{3}$$

If $\Delta_{j|i}(z) \leq 0$, the fit selects only $c_i$ and removes candidate $c_j$. Appendix A.1 gives the proof.

Lemma 1 identifies a shadow when a stronger annotated atom explains enough of the candidate's evidence to exclude it from the two-label sparse fit.

**Definition 2 (Shadow and Rival Co-Activations).** Take an FCA $j$. We compare it with each related annotated event $i$ that has at least as much dense evidence and can enter the sparse fit. These reference events form

$$\mathcal{I}_j(z) = \left\{ i \in Y_m(t) : c_i^\top z \geq c_j^\top z,\ c_i^\top z > \gamma,\ 0 < c_i^\top c_j < 1 \right\}.$$

The lemma permits $\rho = 0$, but the diagnostic below requires $\rho > 0$ because a shadow specifically denotes evidence shared by related atoms. For each $i \in \mathcal{I}_j(z)$, consider the two-label sparse problem in Equation (2). We define

$$j \text{ is a shadow} \quad \Longleftrightarrow \quad \exists i \in \mathcal{I}_j(z) : \Delta_{j|i}(z) \leq 0,$$

and call it a rival otherwise. In Figure 1(a), *Dog* remains selected while *Cat* disappears, so *Cat* is a shadow. A rival is any FCA for which no eligible stronger annotated event removes the candidate in this two-label analysis. This includes $\mathcal{I}_j(z) = \varnothing$, where no such reference event exists. Figure 1(c) reports both types on multi-label segments for a matched comparison.

Lemma 1 is exact for two labels and does not fully characterize selection with the complete dictionary. Shadow and rival are therefore diagnostic tags, not hidden ground-truth causes. Appendix A.1 tests how the two-label condition relates to the full 25-label solve.

# 3 METHOD

Figure 2 summarizes the SCoPE pipeline.

## 3.1 PROBLEM FORMULATION

Given $T$ aligned audio and visual segments and a queried event vocabulary $\mathcal{Y}$, AVEP predicts three binary decisions $\hat{y}_a(t,c)$, $\hat{y}_v(t,c)$, and $\hat{y}_{av}(t,c)$ for every segment $t$ and candidate $c$. The first two indicate whether the event is audible or visible, respectively, while $\hat{y}_{av}$ denotes an event that is jointly audible and visible. Predictions are made from segment embeddings $z_m(t)$ and text dictionaries $C_m$ defined in Section 2. Each text dictionary is encoded once and reused across videos.

## 3.2 MEAN CENTERING

Multimodal contrastive models can place image or audio embeddings and text embeddings in different regions of their shared space, creating a modality gap (Liang et al., 2022; Wang et al., 2023). Following prior work (Bhalla et al., 2024), we center and renormalize the embeddings before sparse matching. After L2-normalizing $z_m(t)$ and $\mathbf{c}_m(c)$, we subtract a modality-specific feature mean $\mu_m$ and queried-dictionary mean $\bar{\mu}_m$, then renormalize:

$$\widetilde{z}_m(t) = \mathrm{norm}(\mathrm{norm}(z_m(t)) - \mu_m)\,, \qquad \widetilde{\mathbf{c}}_m(c) = \mathrm{norm}(\mathrm{norm}(\mathbf{c}_m(c)) - \bar{\mu}_m)\,.$$

Here, $\mathrm{norm}$ denotes L2 normalization, and $\widetilde{C}_m$ stacks $\widetilde{\mathbf{c}}_m(c)$ over $c \in \mathcal{Y}$. Section 4.1 specifies the reference datasets.

## 3.3 STAGE 1: COMPETITIVE SPARSE SELECTION

Stage 1 makes all candidate event labels compete to explain the same centered segment embedding. Building on sparse concept representations for CLIP and CLAP (Bhalla et al., 2024; Zhang et al., 2025), it solves non-negative Lasso independently in each modality:

$$w_m(t) = \arg\min_{w \geq 0} \left\| \tilde{z}_m(t) - \tilde{C}_m w \right\|_2^2 + \lambda_0 \|w\|_1. \tag{4}$$

We replace independent cosine-similarity scores with sparse coefficients that reflect concept prominence (Bhalla et al., 2024; Zhang et al., 2025) and serve as event-selection scores. Equation (4) makes event labels compete for shared evidence. An event label is selected only if it reduces reconstruction error enough to offset its selection cost. A candidate can have high cosine similarity yet receive a small or zero coefficient when other event labels explain its shared evidence.

The support set contains event labels with coefficients above $\varepsilon_{\mathrm{supp}}$:

$$S_m(t) = \{c : w_m(t,c) > \varepsilon_{\mathrm{supp}}\}. \tag{5}$$

Equation (4) extends the two-atom fit in Equation (2) to all queried event labels.

## 3.4 STAGE 2: PRIOR EXCHANGE AND RE-SELECTION

Stage 2 uses support from a source modality to help a target modality distinguish rival event labels. However, mispredictions in the source modality can propagate to the target modality. Moreover, simply adding scores can be misleading when a sound source lies outside the camera's field of view or a visible object is silent. To address these issues, SCoPE converts sparse source support into a pooled soft prior over target selection costs. Then, the target modality re-selects events using its own embedding.

Figure 1(b) shows how this design handles *Dog* as a visible but soundless object. Visual support can lower the audio selection costs of *Dog* and *Car*, but it does not directly add visual scores to audio scores. Audio Stage 2 re-selects events using its own embedding, so *Car* can gain support without forcing *Dog* into the audio output.

**Reconstruction-weighted source prior.** To address error propagation, we weight source segments by their reliability. We describe visual-to-audio guidance below, with the modalities swapped for the

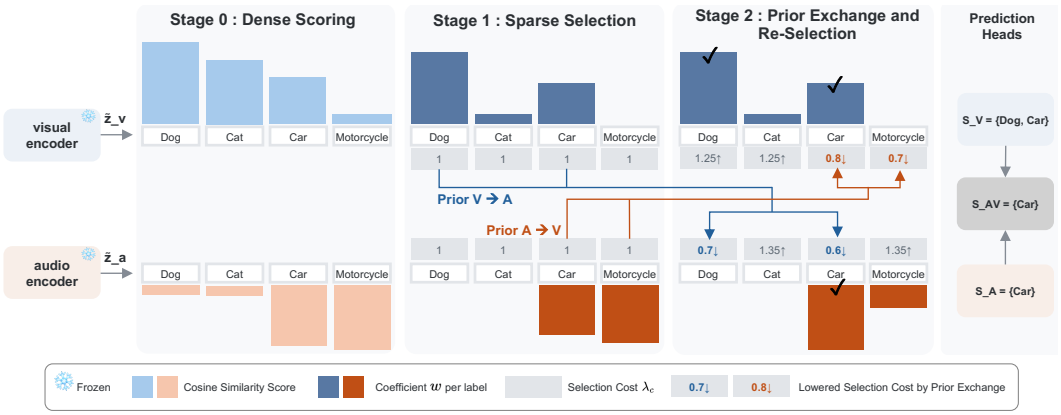



Figure 2: SCoPE on the example from Figure 1. Bars illustrate cosine-similarity scores in Stage 0 and sparse coefficients in Stages 1 and 2. Numbers show selection costs relative to $\lambda_0$. Stage 2 adjusts costs through prior exchange and performs re-selection using target evidence. The AV head fuses coefficients supported by both modalities.

reverse direction. Segment reliability $r_v(t)$ measures cosine similarity between the embedding and its Stage 1 reconstruction. The class-wise prior $P_v(c)$ averages normalized coefficients weighted by $r_v(t)$:

$$r_v(t) = \left[\cos(\tilde{z}_v(t), \tilde{C}_v w_v(t))\right]_+, \qquad P_v(c) = \frac{1}{T}\sum_{t=1}^{T} r_v(t) \frac{w_v(t,c)}{\|w_v(t)\|_\infty + \varepsilon_{\text{norm}}}. \tag{6}$$

Here, $[x]_+ = \max(0, x)$ and $\varepsilon_{\text{norm}}$ is a numerical stabilizer. We normalize coefficients by their segment-wise maximum. Averaging over all segments weakens the prior when reconstruction is consistently poor. We set $r_v(t) = 0$ when the support set is empty. If every $r_v(t)$ is zero, Stage 2 uses the Stage 1 cost.

**Soft-prior cost adaptation.** As in adaptive Lasso (Zou, 2006), we use class-specific selection costs. In SCoPE, these selection costs are derived from support in the other modality. To preserve a fixed total cost budget, we normalize the unnormalized cost multipliers $u_{a,c}$ to obtain the selection costs $\lambda_{a,c}$:

$$u_{a,c} = \exp(-\eta_{v\to a} P_v(c)), \qquad \lambda_{a,c} = \lambda_0 \frac{u_{a,c}}{\frac{1}{|\mathcal{Y}|}\sum_j u_{a,j}}. \tag{7}$$

Here, $\eta_{v\to a}$ controls guidance strength, with $\eta_{a\to v}$ used in the reverse direction. The normalization keeps every cost positive and enforces $\sum_c \lambda_{a,c} = |\mathcal{Y}|\lambda_0$. Stronger source support lowers a class's relative cost, while the other costs rise to preserve the mean.

**Target re-selection.** Using the same selection costs for all segments of the video, the audio modality re-selects events from each segment's own embedding:

$$w_a^{\text{new}}(t) = \arg\min_{w\geq 0} \left\|\tilde{z}_a(t) - \tilde{C}_a w\right\|_2^2 + \sum_c \lambda_{a,c}|w_c|. \tag{8}$$

The audio-to-visual direction applies the same target re-selection to produce $w_v^{\text{new}}(t)$. A lower cost can help the target reconsider a weak label, but it cannot insert the source label or guarantee a correct selection. The reconstruction term determines whether the target retains the candidate.

### 3.5 Sparse Readout

Sparse coefficient magnitudes vary across segments, complicating fixed-threshold calibration. We instead split the sorted coefficients at their largest gap (Gower & Ross, 1969). As in Equation (5), we define the Stage 2 support set $S_m^{\text{new}}(t) = \{c : w_m^{\text{new}}(t,c) > \varepsilon_{\text{supp}}\}$. With $K = |S_m^{\text{new}}(t)|$, we sort the coefficients $w_m^{\text{new}}(t,c)$ for $c \in S_m^{\text{new}}(t)$ in descending order as $w_{m,(1)}^{\text{new}}(t) \geq \cdots \geq w_{m,(K)}^{\text{new}}(t) > 0$

and append the zero anchor $w^{\text{new}}_{m,(K+1)}(t) = 0$. We select the largest gap:

$$k^*_m(t) = \min\left(\arg\max_{1\le k\le K}\left[w^{\text{new}}_{m,(k)}(t) - w^{\text{new}}_{m,(k+1)}(t)\right]\right), \qquad \hat{Y}_m(t) = \{c_{(1)}, \ldots, c_{(k^*_m(t))}\}. \quad (9)$$

We set $\hat{y}_m(t,c) = \mathbb{I}[c \in \hat{Y}_m(t)]$. The outer minimum retains fewer event labels when gaps tie. An empty support set produces no prediction, and a singleton support set retains its label. The zero anchor lets the rule keep the full support set when the final drop is strictly largest. The readout has no calibration threshold or support-size parameter.

For audio-visual prediction, we intersect the audio and visual support sets and fuse only the coefficients of event labels in this intersection:

$$S_{av}(t) = S^{\text{new}}_a(t) \cap S^{\text{new}}_v(t), \quad w_{av}(t,c) = \mathbb{I}[c \in S_{av}(t)]\left[\alpha w^{\text{new}}_a(t,c) + (1-\alpha) w^{\text{new}}_v(t,c)\right]. \quad (10)$$

Applying the same sparse readout to $w_{av}$ gives $\hat{y}_{av}$. For comparability, we set the fusion weight $\alpha$ to AV$^2$A's CLIP+CLAP value (Shaar et al., 2025) and do not retune it by dataset. The audio and visual outputs remain unchanged by this AV fusion.

# 4 Experiments

## 4.1 Setup

**Datasets and metrics.** We evaluate on LLP (Tian et al., 2020) (25 classes), OV-AVEBench (Zhou et al., 2025) (67 classes), and VGGSound-AVEL100k (Zhou et al., 2023) (141 classes). Following HAN (Tian et al., 2020), we report segment and event F1 on LLP for Audio, Visual, Audio-Visual, Type, and Event. Following the OV-AVEBench protocol (Zhou et al., 2025), we report chunk accuracy, segment F1, event F1, and their average. VGGSound-AVEL100k uses segment micro F1 and event F1. Appendix C.2 details the evaluation subsets and metrics.

**Frozen encoders and implementation.** We use CLIP ViT-B/32 (Radford et al., 2021) with LAION-CLAP HTSAT-tiny (Wu et al., 2023) across all three benchmarks, LanguageBind (Zhu et al., 2024) on LLP, and ImageBind (Girdhar et al., 2023) on the transfer benchmarks. SCoPE estimates visual and audio reference means from MS-COCO (Lin et al., 2014) and ESC-50 (Piczak, 2015), respectively, without using LLP features. Appendix B.5 evaluates the reference choice. We reproduce the training-free baselines in Tables 1 and 2 and execute each evaluated method–backbone–dataset configuration three times. Appendix C.1 specifies prompts, reference construction, baseline implementations, and decoding.

**Hyperparameters.** We select $\lambda_0 = 0.30$, $\eta_{v\to a} = 16$, and $\eta_{a\to v} = 4$ on 563 LLP validation videos with complete feature caches. The AV fusion weight $\alpha = 0.45$ follows AV$^2$A. This configuration is fixed across test datasets and backbones. Appendix B.4 gives the search grid and sensitivity.

## 4.2 Main Results

**LLP.** Table 1 compares training-free methods within each backbone block. With CLIP+CLAP, SCoPE raises Type@seg F1 from 43.11 to 52.33 and Event@seg F1 from 34.16 to 42.54, gains of 9.22 and 8.38 points over the AV$^2$A row. With LanguageBind, it reaches 56.60 and 46.83, gains of 4.79 and 4.08 points. Audio-Visual segment F1 improves by 0.57 and 0.65 points for the two backbones, respectively.

**Benchmark and backbone transfer.** Table 2 evaluates the same frozen SCoPE configuration on larger vocabularies and alternative backbones. On OV-AVEBench, SCoPE reaches 51.08 Avg. with CLIP+CLAP and 54.48 with ImageBind, including 51.55 and 56.43 on unseen classes. On strict VGGSound-AVEL100k, SCoPE reaches 50.09 segment F1 and 35.26 event F1 with CLIP+CLAP, compared with 28.47 and 24.21 for the AV$^2$A baseline. With ImageBind, OV-AVEL raises the dense baseline from 15.46 to 45.02 segment F1, and SCoPE further reaches 56.28. Event F1 rises from 5.43 to 22.97 and 41.71, respectively. Appendix B.6 lists trained reference results so that these training-free numbers can be read against supervised models.

Table 1: Training-free audio-visual video parsing on LLP (F1, %).

| Method | Audio | | Visual | | Audio-Visual | | Type | | Event | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Seg | Evt | Seg | Evt | Seg | Evt | Seg | Evt | Seg | Evt |
| CLIP+CLAP | 17.63 | 15.14 | 18.72 | 17.76 | 29.85 | 25.48 | 22.07 | 19.46 | 18.26 | 16.15 |
| with AV$^2$A | 32.48 | 28.29 | 43.27 | 41.18 | 53.58 | 46.68 | 43.11 | 38.72 | 34.16 | 30.64 |
| with **SCoPE** | **42.85** | **38.41** | **59.97** | **56.82** | **54.15** | **49.06** | **52.33** | **48.10** | **42.54** | **37.88** |
| LanguageBind | 20.28 | 17.43 | 21.25 | 19.92 | 24.18 | 20.87 | 21.94 | 19.41 | 20.73 | 18.35 |
| with AV$^2$A | 40.66 | 35.61 | 57.14 | 54.08 | 57.62 | 50.76 | 51.81 | 46.83 | 42.75 | 38.32 |
| with **SCoPE** | **45.87** | **41.52** | **65.66** | **62.99** | **58.27** | **52.70** | **56.60** | **52.40** | **46.83** | **42.22** |

Table 2: Training-free event localization across datasets and frozen backbones (%).

| Method | OV-AVEBench | | | | | VGGSound-AVEL100k | |
|---|---|---|---|---|---|---|---|
| | Acc. | Seg-F1 | Event-F1 | Avg. | Unseen Avg. | Seg F1 | Event F1 |
| CLIP+CLAP | 44.46 | 37.57 | 30.89 | 37.64 | 38.17 | 19.09 | 6.09 |
| with OV-AVEL | 51.86 | 42.92 | 32.85 | 42.54 | 42.05 | 26.98 | 14.83 |
| with AV$^2$A | 43.18 | 31.90 | 24.34 | 33.14 | 32.69 | 28.47 | 24.21 |
| with **SCoPE** | **57.38** | **50.10** | **45.75** | **51.08** | **51.55** | **50.09** | **35.26** |
| ImageBind | 36.98 | 30.61 | 24.53 | 30.71 | 30.53 | 15.46 | 5.43 |
| with OV-AVEL | 59.41 | 48.07 | 35.66 | 47.71 | 48.14 | 45.02 | 22.97 |
| with **SCoPE** | **60.97** | **53.51** | **48.97** | **54.48** | **56.43** | **56.28** | **41.71** |

### 4.3 Mechanism Analysis

Table 3 summarizes the controlled exchange variants. Table 4 follows fixed pairwise-tagged shadows from multi-label segments and rivals from all event-containing segments through the pipeline. A candidate counts as a GT-retaining rejection when the evaluated output excludes it while at least one annotated event in the same segment remains selected. We evaluate the raw Stage 1 readout and the final Stage 2 output.

| Variant | Soft prior | Reconstruction weighting | Target re-selection | Type@seg | Event@seg |
|---|---|---|---|---|---|
| Stage 1 (no exchange) | × | × | × | 42.36 | 34.58 |
| Hard prior | × | ✓ | ✓ | 51.53 | 42.13 |
| Direct fusion | ✓ | ✓ | × | 43.90 | 39.54 |
| Soft prior, unweighted | ✓ | × | ✓ | 51.91 | 42.47 |
| SCoPE | ✓ | ✓ | ✓ | **52.33** | **42.54** |

**Table 3:** Cross-modal prior ablation on LLP (%).

| GT-retaining rejection (%) | | | |
|---|---|---|---|
| Modality | Type | Stage 1 | Stage 2 |
| Audio | Shadow | **96.07** | 92.95 |
| | Rival | 13.17 | **40.68** |
| Visual | Shadow | **100.00** | 95.06 |
| | Rival | 5.71 | **45.60** |

**Table 4:** GT-retaining rejection of fixed pairwise-tagged FCA candidates.

**What should be exchanged, and how?** We evaluate the design choices in Section 3.4 intended to limit inappropriate transfer of evidence from the source modality to the target modality. All variants start from the same Stage 1 coefficients. The Stage 1 baseline uses no cross-modal exchange. A hard prior weights selected source event labels equally, whereas a soft prior gives stronger labels more influence. The hard and soft prior variants use target re-selection. The unweighted soft prior gives source segments equal influence, while SCoPE weights them by reconstruction reliability. Direct fusion instead adds the reconstruction-weighted soft prior to the target coefficients without target re-selection.

Target re-selection has the largest effect. Replacing it with direct fusion lowers Type@seg and Event@seg F1 by 8.43 and 3.00 points, respectively. The unweighted soft prior outperforms the hard prior by 0.38 and 0.34 points. Adding reconstruction weighting yields the full SCoPE model and further improves these metrics by 0.42 and 0.07 points. Section 4.4 provides clip-level examples of target re-selection.

Direct fusion and SCoPE use the same reconstruction-weighted soft prior for coefficient addition and cost-guided target re-selection, respectively. Direct fusion and the hard prior retain SCoPE's fixed 4:1 directional ratio, with only a global scale tuned on LLP validation. For the unweighted soft prior, we rescale $\eta$ using label-free validation features to match SCoPE's mean absolute selection-cost change, separating reconstruction weighting from prior strength. All variants use the same sparse readout and temporal decoding as the main experiments, isolating the effect of the exchange design.

**Which candidates does each stage remove?** Table 4 follows the frozen tags. Sparse selection alone already rejects almost every shadow candidate in both modalities (96.07% and 100.00%), consistent with the two-label condition, under which a shadow is removed once its stronger annotated atom is selected. It rejects few rival candidates (13.17% and 5.71%), because a rival carries evidence that the same modality supports on its own. After the full pipeline, rival rejection rises to 40.68% and 45.60% while shadow rejection stays above 92%.

### 4.4 Qualitative Analysis

Figure 3 shows how cross-modal guidance changes event selection in two LLP clips. Blue cells mark positive Stage 1 support, and green cells show final SCoPE predictions.

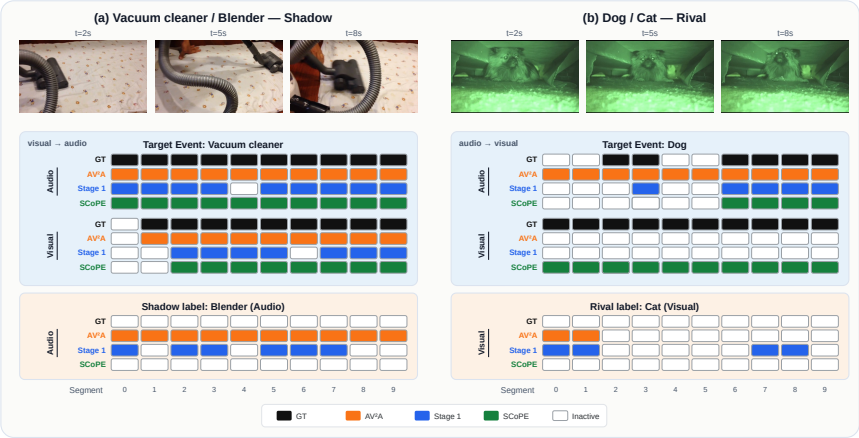



Figure 3: Detailed LLP cases. Blue cells show positive Stage 1 support, and green cells show final SCoPE predictions. Upper blocks follow the annotated target event in both modalities. Lower blocks follow a related incorrect event label.

**Suppressing a competing event.** In Figure 3(a), audio Stage 1 activates both the correct label Vacuum cleaner and the incorrect label Blender. Visual prior exchange lowers Vacuum cleaner's audio selection cost, helping retain it in all ten audio segments and suppress Blender. In the reverse direction, incorrect audio support for Blender lowers its visual selection cost. Target re-selection limits Blender's visual support. As a result, Blender is excluded from the final predictions in both modalities.

**Recovering a missed event.** In Figure 3(b), visual Stage 1 misses Dog. Audio guidance helps recover Dog, which appears in the final visual predictions across all ten segments. In the reverse direction, the visual prior also makes the incorrect label Cat cheaper to select than Dog. Despite its higher selection cost, Dog remains selected based on audio evidence. Appendix B.3 gives the priors, costs, and coefficient changes for both clips, together with additional examples.

## 5 Related Work

### 5.1 Fixed-Vocabulary AVEP

Audio-visual event localization (Tian et al., 2018) identifies events that are both audible and visible in temporal video segments. Audio-visual video parsing (Tian et al., 2020) broadens this task to

audio-only, visual-only, and jointly audio-visual events under video-level supervision. Later methods improve temporal event localization and modality assignment with task-specific models (Wu & Yang, 2021; Fan et al., 2023; Lai et al., 2023; Yu et al., 2025b; Gao et al., 2023; Lai et al., 2025). Because these models are trained for a fixed label set, extending them to new event names or domains requires new annotations or retraining.

### 5.2 Open-Vocabulary AVEP

Open-vocabulary AVEP queries event classes through text embeddings. Learning-based methods use class-agnostic proposals and multimodal classification for untrimmed videos (Yu et al., 2025a) or learn hierarchical cross-modal interactions (Yang et al., 2026), with task-specific training on seen categories. However, generalizing cross-modal fusion learned in closed-set settings to unseen categories remains challenging (Zhou et al., 2025). Training-free approaches instead refine frozen text-matching scores through fixed thresholds (Zhou et al., 2025) or video-level candidate filtering and adaptive class thresholds (Shaar et al., 2025). Decision rules can also be adapted using a learned policy over frozen encoder outputs (Jo et al., 2026). Section 2 analyzes score conflicts that remain under these decision rules. SCoPE instead exchanges selection-cost priors without training, retaining each modality's own evidence for re-selection.

### 5.3 Concept Bottleneck Models and Intervention

Concept bottleneck models (Koh et al., 2020; Oikarinen et al., 2023; Yuksekgonul et al., 2023) and sparse concept representations such as SpLiCE (Bhalla et al., 2024) allow users to inspect and edit model behavior through named concepts. SpLiCE decomposes frozen CLIP embeddings into sparse text concepts, and related work constructs interpretable concept representations for audio (Zhang et al., 2025). However, using these representations to correct model behavior still requires human guidance. Within CBMs, later work reduces this burden by selecting which concepts to query (Chauhan et al., 2023; Shin et al., 2023; Espinosa Zarlenga et al., 2023), propagating one correction to correlated concepts (Singhi et al., 2024; Vandenhirtz et al., 2024), or reusing past corrections on new inputs (Steinmann et al., 2024). These approaches reduce the amount of feedback needed, but their corrections still originate from people.

SCoPE instead uses evidence from the other modality to guide sparse event selection without human feedback at inference time. A shared event vocabulary allows the support selected in one modality to guide the selection of the same events in the other.

## 6 Discussion and Conclusion

**Causal streaming.** Causal streaming inference is important for real-world AVEP applications that must respond as events occur. SCoPE supports efficient causal streaming by forming priors from observed segments alone. With one-segment confirmation, Type@seg and Event@seg F1 remain within 2.71 and 1.67 points of offline inference across both backbones (Appendix B.1).

**Future directions.** First, reducing the confirmation delay while preserving accuracy could improve causal streaming. Second, adapting selection costs to each input segment could make cross-modal guidance more flexible. Third, frozen object detectors (Liu et al., 2024) and audio source separation models (Shi et al., 2026; Huang et al., 2025) could provide stronger modality-specific evidence without task-specific training.

**Limitations.** First, Lemma 1 applies only to two atoms and does not fully characterize selection over the full dictionary. Second, SCoPE depends on the recognition ability and representation quality of its frozen backbones.

**Conclusion.** Training-free AVEP often selects events by thresholding independently computed similarity scores. When an incorrect event label scores above a ground-truth label, a single threshold cannot reject it while retaining all ground-truth labels. We characterize this false co-activation through shadows and rivals and introduce SCoPE to address it. SCoPE combines sparse competition within each modality with cross-modal prior exchange. Each modality guides the other, but event selection relies on the target modality's own evidence. Experiments across AVEP benchmarks and frozen backbones demonstrate improved event prediction using the same hyperparameter settings.

## AI USE STATEMENT

We used a large language model as an assistant. For implementation, parts of the inference code and experiment scripts were written with AI assistance, and some experiments were run by the assistant under the authors' direction. The authors reviewed this code and verified its outputs against reference implementations and recorded parity checks. For the research, the assistant provided feedback on experimental design and helped interpret results. For the paper, it assessed whether the experimental evidence was sufficient to support each claim, helped write and edit the text, and suggested candidate references that the authors verified. We did not use generative AI to formulate the method, Lemma 1, or its proof, to generate synthetic data, or to modify experimental results, and the remaining required-disclosure tasks are not applicable. The authors verified every AI-assisted result and sentence against the recorded experimental artifacts and take responsibility for the final content of this work.

## REPRODUCIBILITY STATEMENT

Appendix C.1 specifies encoders, prompts, reference-mean construction, solver constants, and temporal decoding. Appendix C.2 gives evaluation subsets and metrics. Anonymized code for preprocessing, inference, and evaluation is included in the supplementary material.

## REFERENCES


Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

Praneeth Chakravarthula, Jim Aldon D'Souza, Ethan Tseng, Joe Bartusek, and Felix Heide. Seeing with sound: Long-range acoustic beamforming for multimodal scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 982–991, 2023.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5948–5955, 2023.

Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15511–15520. IEEE, 2021.

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.

Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pp. 431–448. Springer, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36:37849–37875, 2023.

Yingying Fan, Yu Wu, Bo Du, and Yutian Lin. Revisit weakly-supervised audio-visual video parsing from the language perspective. *Advances in Neural Information Processing Systems*, 36:40610–40622, 2023.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18827–18836. IEEE, 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

John C Gower and Gavin JS Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 18(1):54–64, 1969.

Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22910–22921. IEEE, 2023.

Chao Huang, Yuesheng Ma, Junxuan Huang, Susan Liang, Yunlong Tang, Jing Bi, Wenqiang Liu, Nima Mesgarani, and Chenliang Xu. ZeroSep: Separate anything in audio with zero training. *Advances in Neural Information Processing Systems*, 38:111844–111873, 2025.

Beomgwon Jo, Sunchan Park, and Kyeongbo Kong. Compositional agentic formulation search for open-vocabulary audio-visual event localization. In *2nd Workshop on Compositional Learning: Safety, Interpretability, and Agents*, 2026.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yung-Hsuan Lai, Yen-Chun Chen, and Yu-Chiang Frank Wang. Modality-independent teachers meet weakly-supervised audio-visual event parser. *Advances in Neural Information Processing systems*, 36:73633–73651, 2023.

Yung-Hsuan Lai, Janek Ebbers, Yu-Chiang Frank Wang, François Germain, Michael Jeffrey Jones, and Moitreya Chatterjee. UWAV: Uncertainty-weighted weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13561–13570, 2025.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19, 2006.

Jia Li, Wenjie Zhao, Ziru Huang, Yunhui Guo, and Yapeng Tian. Do audio-visual segmentation models truly segment sounding objects? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6172–6180, 2026.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of machine learning research*, 11(1), 2010.

Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 35:34722–34733, 2022.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

Karol J Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

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.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014.

Eitan Shaar, Ariel Shaulov, Gal Chechik, and Lior Wolf. Adapting to the unknown: Training-free audio-visual event perception with dynamic thresholds. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3142–3151. IEEE, 2025.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. SAM Audio: Segment anything in audio. In *International Conference on Machine Learning*, 2026.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International conference on machine learning*, pp. 31504–31520. PMLR, 2023.

Nishad Singhi, Jae Myung Kim, Karsten Roth, and Zeynep Akata. Improving intervention efficacy via concept realignment in concept bottleneck models. In *European Conference on Computer Vision*, pp. 422–438. Springer, 2024.

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 46556–46571. PMLR, 2024.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pp. 436–454. Springer, 2020.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia E Vogt. Stochastic concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:51787–51810, 2024.

Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023.

Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1326–1335, 2021.

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.

Zhe Yang, Ruyi Zhang, Hongtao Chen, Wenrui Li, Hengyu Man, Wangmeng Zuo, and Xiaopeng Fan. Hierarchical semantic-constrained heterogeneous graph for audio-visual event localization. *arXiv preprint arXiv:2606.07033*, 2026.

Jiale Yu, Baopeng Zhang, Zhu Teng, and Jianping Fan. OV-DAVEL: Towards open-vocabulary dense audio-visual event localization in untrimmed videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 553–562, 2025a.

Xiao Yu, Yan Fang, Yao Zhao, and Yunchao Wei. PreFM: Online audio-visual event parsing via predictive future modeling. *Advances in Neural Information Processing Systems*, 38:140138–140165, 2025b.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023.

Alice Zhang, Edison Thomaz, and Lie Lu. Transformation of audio embeddings into interpretable, concept-based representations. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Yi Zhang, Ke Yu, Angelica I Aviles-Rivero, Jiyuan Jia, Yushun Tang, and Zhihai He. Training-free feature reconstruction with sparse optimization for vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4387–4396, 2024.

Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2023.

Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. Label-anticipated event disentanglement for audio-visual video parsing. In *European Conference on Computer Vision*, pp. 35–51. Springer, 2024.

Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371. IEEE, 2025.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pp. 9588–9608, 2024.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

# SUPPLEMENTARY MATERIAL

## A SPARSE CANDIDATE AND EVENT ANALYSIS

### A.1 PAIRWISE RESIDUAL CRITERION

**Lemma 1 (restated).** Consider the two-label non-negative sparse fit in Equation (2), with unit-norm atoms satisfying $0 \le \rho = c_i^\top c_j < 1$, $c_i^\top z \ge c_j^\top z$, and $c_i^\top z > \gamma = \lambda_0/2$. The solution selects both atoms if and only if

$$(c_j - \rho c_i)^\top z > \gamma(1-\rho).$$

Otherwise, only $c_i$ is selected.

*Proof.* On the nonnegative orthant, Equation (2) has a differentiable convex objective and affine constraints with strictly feasible points. The Karush–Kuhn–Tucker (KKT) conditions are therefore necessary and sufficient for optimality (Boyd & Vandenberghe, 2004). For residual $r = z - Cw$ and $\gamma = \lambda_0/2$, they reduce to

$$w_k \ge 0, \qquad c_k^\top r \le \gamma, \qquad w_k(c_k^\top r - \gamma) = 0.$$

Thus an active atom has residual correlation $\gamma$, while an inactive atom has residual correlation at most $\gamma$. Suppose $c_j$ is inactive. The one-atom optimum for $c_i$ under Equation (2) is

$$w_i^\star = c_i^\top z - \gamma, \qquad \gamma = \lambda_0/2.$$

The KKT condition keeps $c_j$ inactive exactly when

$$c_j^\top (z - w_i^\star c_i) \le \gamma.$$

Substituting $w_i^\star$ and using $\rho = c_i^\top c_j$ gives

$$(c_j - \rho c_i)^\top z \le \gamma(1-\rho).$$

When this inequality is violated, the coefficient of $c_j$ is positive. The score ordering keeps the coefficient of $c_i$ at least as large, so the pairwise solution selects both atoms. Otherwise, it selects only $c_i$. Since $\rho < 1$, the two-atom objective is strictly convex on their span and the solution is unique. □

For the idealized mixture $z = ac_i + bc_j$ with $a, b \ge 0$, dense matching assigns $c_j$ the score $b + \rho a$. After conditioning on $c_i$, the shared term $\rho a$ cancels, and Equation (3) reduces to

$$b > \frac{\gamma}{1+\rho}.$$

The second atom therefore survives only when its independent mixture weight exceeds the event selection cost implied by its similarity to the selected atom.

**Pairwise diagnostic tags.** For every candidate $j$ outside the annotated event set in a segment with at least one annotated event, we compute one tag from the centered target embedding and text dictionary. Let $s_k = c_k^\top z$ and $\rho_{ij} = c_i^\top c_j$. We collect all annotated reference events $i$ with $s_i \ge s_j$, $s_i > \gamma$, and $\rho_{ij} > 10^{-8}$. If no such reference exists, $j$ is a rival. Otherwise, we evaluate

$$\delta_{ij} = s_j - \rho_{ij} s_i - \gamma(1-\rho_{ij})$$

for every eligible reference. Candidate $j$ is a shadow when $\min_i \delta_{ij} \le 0$ and a rival otherwise. We record the minimizing reference and margin for audit, but evaluate all eligible references when assigning the tag.

The tag is fixed from the target embedding before methods are compared. Dense matching compares all candidates using centered cosine values, AV$^2$A compares its retained candidates using the released similarity scores for each prediction head, and SCoPE compares positive sparse coefficients. Because the methods retain different numbers of ground-truth labels, each score floor is the lowest-valued ground-truth label retained by that method.

This is a pairwise diagnostic rather than an explanation of the full-dictionary prediction. It tests whether any one annotated event can explain the competing candidate under the two-atom model, while the full solve may also depend on interactions with the remaining atoms.

**Full-dictionary check.** The lemma is exact for a two-atom dictionary: the pairwise solution retains both atoms when Equation (3) holds and only the stronger atom otherwise. We compare this exact implication with Stage 1 solutions over all 25 LLP atoms. On segments with exactly two annotated events, we order their atoms by dense score and apply the condition to the lower-scoring atom. This checks whether annotated events remain in the support and is not an accuracy claim for the candidate tags.

**Table 5:** Pairwise-condition agreement under two-label and full 25-label sparse solves on LLP. Agreement means that the condition correctly predicts whether both annotated events remain selected. The test uses segments with exactly two annotated events and does not by itself validate the tags assigned to unannotated candidates.

| Dictionary | Audio agreement | Visual agreement |
|---|---|---|
| 2 labels | 100.00% | 100.00% |
| 25 labels | 96.15% | 99.45% |

The two-label solve follows Lemma 1 exactly. With all 25 labels competing, its outcome agrees with the pairwise condition in 96.15% of audio cases and 99.45% of visual cases. In the removal direction, where $\Delta_{j|i}(z) \leq 0$, the lower-scoring event incorrectly remains selected in only 0.55% of audio cases and no visual cases. Most remaining disagreement occurs in the opposite direction, when additional atoms remove an event that the two-label fit would retain.

# B Additional Results and Diagnostics

## B.1 Efficiency and Causal Streaming

Real-world AVEP applications require efficient inference and predictions based only on observations available so far. We therefore evaluate video-level runtime, vocabulary scaling, and causal streaming.

Each stage of SCoPE solves a convex non-negative Lasso problem with fixed selection costs, so a local minimum is also a global minimum (Boyd & Vandenberghe, 2004). This design is also efficient in practice. On OV-AVEBench videos, SCoPE takes 383.66 ms per video, compared with 5,026.40 ms for AV$^2$A and 356.23 ms for OV-AVEL (Table 6).

For vocabulary scaling, solver-only latency with cached features stays near 3 ms for the three benchmark dictionaries (25 to 141 labels) and rises to 17.05 ms for approximately 10k labels (Figure 4). Processing beyond the 1 s segment interval would cause latency to accumulate, but a single NN-Lasso solve with the largest dictionary takes only 1.7% of that interval.

| OV-AVEBench runtime (ms/video) | | | |
|---|---|---|---|
| Method | Encoding | Inference | Total |
| OV-AVEL | 349.09 | **7.14** | **356.23** |
| AV$^2$A | 349.09 | 4,677.31 | 5,026.40 |
| SCoPE | 349.09 | <u>34.57</u> | <u>383.66</u> |

**Table 6:** Video-level runtime on 100 OV-AVEBench videos. Encoding is the shared reference; Inference includes additional method-specific encoding.

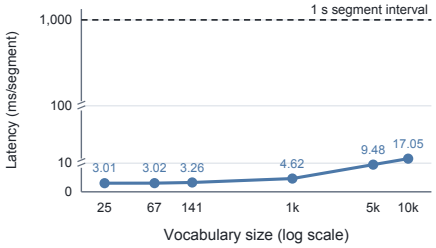



**Figure 4:** Vocabulary scaling for one visual solve on 100 LLP segments, using the benchmark dictionaries (25, 67, and 141 atoms) and larger ImageNet (1,019) and LAION (5,025/10,025) vocabularies.

**Causal streaming.** Equation (6) computes the source prior over the full video for offline inference. For causal streaming, we modify this computation to use only the segments observed up to time $t$.

$$P_m^{\leq t}(c) = \frac{1}{t}\sum_{\tau=1}^{t} r_m(\tau)\frac{w_m(\tau,c)}{\|w_m(\tau)\|_\infty + \varepsilon_{\text{norm}}}, \qquad m \in \{a, v\}.$$

The coefficients and reliability weights follow Equation (6). We then perform bidirectional prior exchange for the current segment. The 0 s setting emits the sparse readout immediately without

temporal post-processing. The 1 s setting performs no gap filling, emits time $t$ after observing $t+1$, and keeps it only with an adjacent positive segment. Across both backbones, it remains within 2.71 Type@seg points and 1.67 Event@seg points of offline inference (Table 7).

| | Offline | | Causal, 0 s | | Causal, 1 s | |
|---|---|---|---|---|---|---|
| Backbone | Type@seg | Event@seg | Type@seg | Event@seg | Type@seg | Event@seg |
| CLIP+CLAP | 52.33 | 42.54 | 43.00 | 29.36 | 49.62 | 40.87 |
| LanguageBind | 56.60 | 46.83 | 51.52 | 39.62 | 54.95 | 45.61 |

**Table 7:** Causal streaming on LLP (F1, %). The 0 s output has no temporal post-processing, whereas the 1 s output uses one-segment confirmation.

Appendix C.1 provides implementation details including runtime measurement protocols and temporal post-processing.

### B.2 Compatibility with Training-Free Decision Rules

Using the same frozen Stage 2 coefficients, we test fixed AV$^2$A-inspired dynamic thresholding (Shaar et al., 2025) and video-level candidate filtering on all three prediction heads. The former replaces SCoPE's largest-gap decoder, while the latter filters candidates separately for each prediction head before SCoPE's original decoder. Dynamic thresholding preserves Type@seg and improves Event@seg, while candidate filtering improves both.

Table 8: Compatibility with training-free decision rules on LLP (F1, %).

| Variant | Type@seg | Event@seg |
|---|---|---|
| SCoPE | 52.33 | 42.54 |
| Dynamic-threshold decoder | 52.31 | 43.40 |
| Candidate filter + SCoPE | **53.47** | **47.45** |

### B.3 Qualitative Examples

We detail the priors, selection costs, and coefficient changes behind the two clips in Section 4.4 (Figure 3).

**Cross-modal guidance safeguards.** As described in Section 3.4, cross-modal guidance must account for both source mispredictions and events that occur only in one modality. SCoPE addresses these issues through reconstruction-weighted source priors, soft-prior cost adaptation, and target re-selection. The following examples trace how these three components affect event selection across time segments, from source support to final target predictions.

**Vacuum cleaner and Blender.** The first clip in Figure 3(a) is annotated with Vacuum cleaner but not Blender. Although audio Stage 1 also activates Blender, this misprediction does not become a final visual prediction. In the reverse direction, visual support for Vacuum cleaner helps the audio modality retain the correct event and suppress Blender.

In the visual-to-audio direction highlighted in Figure 3(a), the visual source supports Vacuum cleaner but not Blender, with $P_v(\text{Vacuum cleaner}) = 0.150$ and $P_v(\text{Blender}) = 0$. With $\eta_{v\to a} = 16$, Equation (7) lowers the audio cost of Vacuum cleaner from $\lambda_0 = 0.300$ to $0.028$ and raises that of Blender to $0.311$. Target re-selection retains Vacuum cleaner in all ten audio segments and reduces Blender's positive coefficients from six segments to two. The sparse readout removes the remaining Blender coefficients, $0.085$ and $0.110$.

In the audio-to-visual direction, the audio source supports both events, with $P_a(\text{Vacuum cleaner}) = 0.353$ and $P_a(\text{Blender}) = 0.135$. With $\eta_{a\to v} = 4$, their visual selection costs become $0.078$ and $0.187$, respectively. Under these costs, target re-selection expands Vacuum cleaner's visual support from seven to eight segments. Despite Blender's lower selection cost, re-solving from the visual evidence gives it only a small positive coefficient of $0.022$ in one segment, which the sparse readout then removes.

**Dog and Cat.** Unlike the previous example, the second clip in Figure 3(b) shows audio helping visual recognition. Visual Stage 1 misses Dog despite its presence in the clip. We examine how audio guidance recovers Dog and how visual guidance, in turn, affects the audio predictions. The audio source provides $P_a(\text{Dog}) = 0.236$ and $P_a(\text{Cat}) = 0.050$, giving visual selection costs of $0.124$ and $0.260$ with $\eta_{a\to v} = 4$. Target re-selection gives Dog positive coefficients in eight visual segments, and temporal gap filling yields a final Dog prediction across all ten segments.

In the reverse visual-to-audio direction, visual guidance sets the audio selection costs of Dog and Cat to $0.307$ and $0.140$, respectively. Despite its higher cost, Dog remains selected based on the audio evidence, showing that selection is not determined by the prior alone.

Figure 5 provides four additional comparisons across the audio and visual tracks.

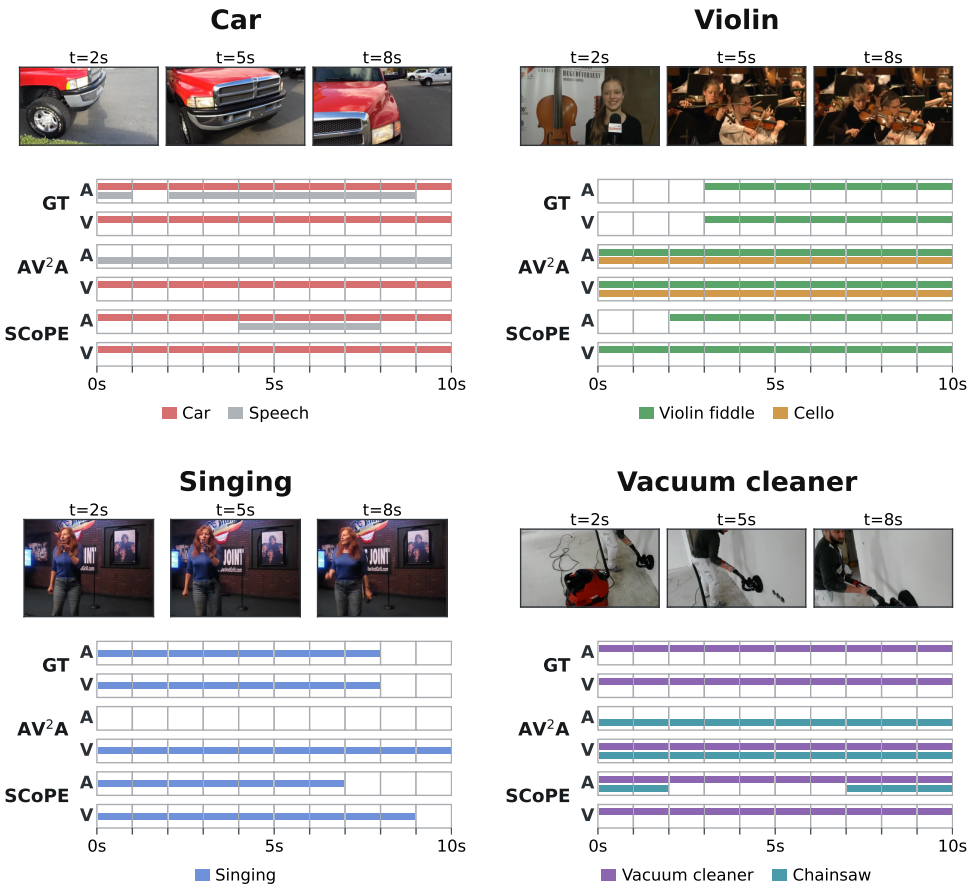



Figure 5: Additional qualitative LLP examples. Timelines compare ground-truth annotations, AV$^2$A predictions, and SCoPE predictions for audio and visual events. Each legend lists only the labels displayed in that panel.

### B.4 Hyperparameter Sensitivity

Figure 6 varies one setting at a time on the LLP validation subset while fixing the others to assess robustness around the reported configuration. The reported setting is $\lambda_0 = 0.3$, geometric prior scale $g = \sqrt{\eta_{v\to a}\eta_{a\to v}} = 8$, direction ratio $\eta_{v\to a} : \eta_{a\to v} = 4{:}1$, and AV fusion weight $\alpha = 0.45$.

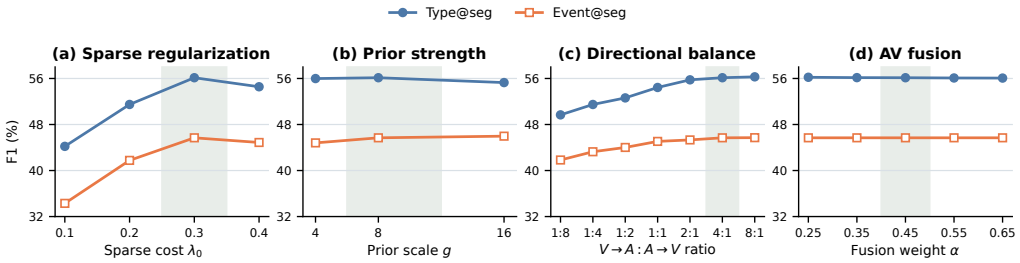



Figure 6: Hyperparameter sensitivity on LLP validation with CLIP+CLAP. Curves report Type@seg and Event@seg while varying (a) $\lambda_0$, (b) the geometric mean $g$ of the prior strengths, (c) their direction ratio, and (d) the AV fusion weight $\alpha$. Shading marks the reported setting.

### B.5 Mean Centering Ablation

We examine the effect of mean centering on Dense predictions and SCoPE's sensitivity to the choice of external reference data. Centering yields only small gains for Dense, which remains well below SCoPE. SCoPE also maintains similar performance across alternative reference datasets.

Table 9: Mean-centering controls and reference sensitivity for CLIP+CLAP on LLP (%).

| Method | Audio reference | Visual reference | A@seg | V@seg | AV@seg | Type@seg | Event@seg |
|---|---|---|---|---|---|---|---|
| Dense | None | None | 31.15 | 48.79 | 38.54 | 39.49 | 32.17 |
| Dense | ESC-50 | MS-COCO | 32.92 | 49.21 | 38.91 | 40.35 | 33.32 |
| SCoPE | ESC-50 | MS-COCO | **42.85** | 59.97 | 54.15 | **52.33** | **42.54** |
| SCoPE | UrbanSound8K | MS-COCO | 41.32 | 59.46 | 51.84 | 50.87 | 41.58 |
| SCoPE | FSD50K | MS-COCO | 41.15 | **60.78** | **54.21** | 52.05 | 41.11 |
| SCoPE | ESC-50 | ImageNet-1k | 42.71 | 59.01 | 53.02 | 51.58 | 42.41 |
| SCoPE | ESC-50 | CIFAR-100 | 41.57 | 56.38 | 51.65 | 49.87 | 41.00 |

Centering improves Dense Type@seg from 39.49 to 40.35 and Event@seg from 32.17 to 33.32, leaving gaps of 11.98 and 9.22 points to SCoPE, respectively. For SCoPE, using UrbanSound8K (Salamon et al., 2014), FSD50K (Fonseca et al., 2022), ImageNet (Deng et al., 2009), or CIFAR-100 (Krizhevsky, 2009) as alternative reference datasets yields Type@seg scores of 49.87–52.05 and Event@seg scores of 41.00–42.41.

Both Dense controls use thresholds calibrated separately on LLP validation, with performance evaluated on LLP test. Both controls use the same frozen features, prompts, score normalization, and temporal decoder, differing only in whether segment and text embeddings are centered. Our choice of alternative reference datasets is motivated by the datasets used in prior work on sparse concept representations (Bhalla et al., 2024; Zhang et al., 2025). For SCoPE, we change only the external reference dataset used to compute each modality's mean. Every reference embedding is L2-normalized before averaging, and no LLP feature is used to estimate a displayed mean.

### B.6 Trained Reference Results

Table 10 compares SCoPE with task-specifically trained methods as reference points.

For LLP, results for MGN-MA (Mo & Tian, 2022), JoMoLD (Cheng et al., 2022), and CMPAE (Gao et al., 2023) are taken from the supplementary material of AV$^2$A (Shaar et al., 2025). The trained systems use task-specific networks and conventional AVVP features, while SCoPE performs inference with frozen CLIP+CLAP encoders.

For OV-AVEBench, both methods use ImageBind. OV-AVEL (FT) denotes the reported temporally trained model (Zhou et al., 2025), while the SCoPE row reports our reproduced training-free results.

**Table 10:** Supervised reference results (%), grouped by dataset. Backbones differ in (a), whereas both methods use ImageBind in (b). Training settings are listed for each method.

**(a) LLP segment-level results**

| Method | Training | Audio | Visual | Audio-Visual | Type | Event |
|---|---|---|---|---|---|---|
| MGN-MA | Weak | 60.3 | 55.3 | 50.1 | 55.3 | 56.9 |
| JoMoLD | Weak | 61.1 | 63.5 | 56.8 | 60.5 | 59.7 |
| CMPAE | Weak | **64.1** | **66.1** | **59.1** | **63.3** | **62.9** |
| **SCoPE** | Free | 42.9 | 60.0 | 54.2 | 52.3 | 42.5 |

**(b) OV-AVEBench results**

| Method | Backbone | Training | Acc. | Seg-F1 | Event-F1 | Avg. | Unseen |
|---|---|---|---|---|---|---|---|
| OV-AVEL (FT) | ImageBind | Temporal | **67.1** | **56.9** | **49.5** | **57.8** | 55.8 |
| **SCoPE** | ImageBind | Free | 61.0 | 53.5 | 49.0 | 54.5 | **56.4** |

# C Implementation Details

## C.1 Inference Details and Runtime Measurement

**Encoders and prompts.** The CLIP+CLAP setup uses CLIP ViT-B/32 (Radford et al., 2021) and LAION-CLAP HTSAT-tiny (Wu et al., 2023). Following AV$^2$A's released implementation, the audio prompt is `This is a sound of {label}` and the visual prompt is `A {label}`. LanguageBind uses `A {label}` in both modalities. ImageBind uses the same modality-specific templates as CLIP+CLAP.

On OV-AVEBench, CLIP+CLAP uses the 67 evaluation prompts. ImageBind adds one fixed *other* prompt, collapsed to background, for both its base and SCoPE rows.

**Runtime measurement.** Table 6 uses the same 100 evenly spaced OV-AVEBench test videos, one TITAN RTX, PyTorch 2.4.0/CUDA 12.1, and a shared CLIP+CLAP encoder instance. At video batch size one, basic A/V encoding runs in parallel on separate workers and CUDA streams, using ten RGB frames of size $224 \times 224$ and ten one-second audio segments sampled at 48 kHz. After 20 warm-up videos per method, we average 300 measurements from three passes with rotating method order. Each timed run re-encodes the media, while text embeddings, dictionary geometry, and CUDA Graph setup are prepared beforehand. OV-AVEL retains its released decision rule, and SCoPE uses its fixed 200-iteration solves and main-experiment decoder. Total spans data loading through final predictions, excluding warm-up, evaluation, and artifact writing. Encoding is a shared reference of 349.09 ms for loading, preprocessing, and basic A/V encoding, and Inference is the difference between Total and this reference, so it includes any additional encoding a method performs for filtering or refinement.

All runtime measurements use one TITAN RTX, and SCoPE's sparse solves use CUDA Graph replay. For vocabulary scaling, we measure one visual NN-Lasso solve with 200 FISTA iterations on a cached LLP segment at segment batch size one. The measurement includes full coefficient transfer to CPU but excludes encoding, dictionary setup, and graph capture. For the vocabulary-scaling curve, 20 warm-up calls precede three passes over 100 fixed LLP segments for each dictionary, and we report the median of the three pass means. The six dictionaries are the LLP, OV-AVEBench, and VGGSound-AVEL100k label sets (25, 67, and 141 atoms) and external ImageNet and LAION vocabularies with 1,019, 5,025, and 10,025 atoms.

For causal streaming, we measure four passes over 100 LLP videos with cached features, processing one incoming segment at a time. The timed computation includes both sparse stages, prefix-prior updates, readout, and confirmation bookkeeping, but excludes feature encoding, graph setup, and the one-second confirmation wait.

**Temporal post-processing.** For each offline SCoPE audio, visual, and Audio-Visual output, temporal post-processing first fills an exact one-segment zero gap bounded by two positive predictions in the unmodified sparse readout. It then removes remaining singleton positive runs, enforcing a

minimum duration of two segments. AV$^2$A retains its released interval processing, which removes singleton runs before merging intervals separated by at most one segment.

For Table 2, the base rows use a frozen dense decoder, and OV-AVEL applies its training-free A/V agreement rule to the same dense scores. On OV-AVEBench, both use the offline gap-filling and singleton-removal rule described above before longest-duration reduction.

For causal SCoPE, the 0 s setting emits the sparse readout immediately without temporal post-processing. The 1 s setting performs no gap filling. It emits time $t$ after observing $t + 1$ and retains a positive prediction at $t$ only if the same class is positive at $t - 1$ or $t + 1$ in the raw readout. These rules apply to all three output types.

**Solver and centering.** Equations (4) and (8) are solved by non-negative FISTA (Beck & Teboulle, 2009) for 200 fixed iterations from zero initialization in both stages. The same batched solver handles uniform and class-specific costs through elementwise non-negative shrinkage. Equation (6) uses $\varepsilon_{\mathrm{norm}} = 10^{-8}$, and coefficients above $\varepsilon_{\mathrm{supp}} = 10^{-6}$ define the numerical support. For every backbone, reference means use all 118,287 MS-COCO train2017 images and 1,600 ESC-50 clips, encoded by the corresponding visual and audio encoders. Embeddings are L2-normalized before averaging. Each mean is backbone- and modality-specific and uses no LLP features.

### C.2 Evaluation Datasets and Protocols

**LLP.** We evaluate Dense, AV$^2$A, and SCoPE with both CLIP+CLAP and LanguageBind on the same 1,109 raw-available LLP test videos, whose audio-visual labels are the intersection of the audio and visual annotations, using the official AVVP evaluator and reporting video-averaged segment and event F1 for Audio, Visual, Audio-Visual, Type, and Event. Predictions remain multi-label, without longest-duration reduction.

**OV-AVEBench.** We use all 5,820 test videos and 67 classes, including 46 seen and 21 unseen classes. Following the benchmark protocol (Zhou et al., 2025), each method's multi-label output is reduced to the class with the longest predicted duration. We report chunk accuracy, segment F1, event F1, and their average, along with the average on unseen classes.

**VGGSound-AVEL100k.** VGGSound-AVEL100k (Zhou et al., 2023) is constructed from VGGSound (Chen et al., 2020). We use the 19,378 of 20,216 test videos with complete inputs and feature caches. Every method uses this subset and the same 141-class dictionary. We report segment micro F1 and event F1 averaged over temporal IoU thresholds $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.

**Assets and licenses.** We use public datasets and frozen backbones only for research evaluation and feature extraction and do not redistribute raw media or checkpoints. CLIP is released under MIT and LAION-CLAP under CC0-1.0. LLP, OV-AVEBench, VGGSound, MS-COCO, ESC-50, UrbanSound8K, FSD50K, ImageNet, and CIFAR-100 remain subject to their source licenses. The still frame in Figure 1 comes from the public LLP benchmark and is used only to illustrate the benchmark example under its original access conditions.