CUE-TO-CONTEXT SUBSUMPTION LEARNING FOR AUDIO-VISUAL GENERALIZED ZERO-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-Visual Generalized Zero-Shot Learning (AV-GZSL) aims to recognize seen and unseen events by aligning audio-visual samples with textual categories. Although multimodal inputs provide rich information, the representations of full samples are often affected by changing backgrounds and surrounding contexts, making shared event cues difficult to capture across scenes. Existing alignment objectives specify which textual category a full sample should match, but do not explicitly indicate which audio and visual cues should support that match. We observe that recognizable localized cues can recur across different contexts, while each full sample presents these cues under a particular context. Based on this observation, we propose Cue-to-Context Subsumption Learning (CCSL). CCSL first extracts localized visual regions and auditory segments as box inputs. It then maps the box and full inputs into Gaussian distributions using a shared distribution mapper. A directional subsumption objective relates their means and variances, guiding the full representation with localized event cues while retaining information from the complete context. Experiments on VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL demonstrate consistent improvements over strong baselines. CCSL improves HM by 5.75 to 7.84 percentage points over the strongest baselines across the three benchmarks and achieves 30.63% unseen accuracy on VGGSound-GZSL. Our anonymous code link is here.

1 INTRODUCTION

In Audio-Visual Generalized Zero-Shot Learning (AV-GZSL), models are required to simultaneously process visual frames and auditory signals, utilizing both cues to recognize seen and unseen audio-visual events. The core challenge of AV-GZSL lies in aligning different multimodal data with textual categories. Although multimodal data provides rich features, aligning the entire sample with text categories may make the model rely on changing backgrounds rather than the event itself. For example, as illustrated in Figure 1, although these instances belong to the same "`basketball bounce`" category and share identical human actions, their scenes have different backgrounds. As a result, representations of most instances within the same category may differ because of background changes, making their shared event cues harder to capture. Yet the alignment objectives used by representative AV-GZSL methods (Mercea et al., 2022b; Kurzendörfer et al., 2024; Yang et al., 2025; Li et al., 2025b; Wang et al., 2026; Ma et al., 2026a) specify which category the full sample should match. They do not specify which audio and visual cues should support that match. The model may be overly influenced by changing background cues and struggle to learn shared event cues across scenes. Some methods separate motion from scene context (Li et al., 2025a), suppress redundant information with attention (Yu et al., 2025), or use the middle frame to provide object cues (Dong et al., 2024), but they do not explicitly constrain how these cues relate to the full sample representation.

Our observation is that the same event can retain recognizable local audio and visual cues across scenes with different backgrounds. These cues can therefore correspond to multiple representations of the full sample, depending on the surrounding context. We view the local representation as describing a range of plausible realizations of the event, with each full sample representing one realization in a particular context. This motivates us to introduce a subsumption relation, inspired

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: Instances of the same "`basketball bounce`" event under different background contexts.

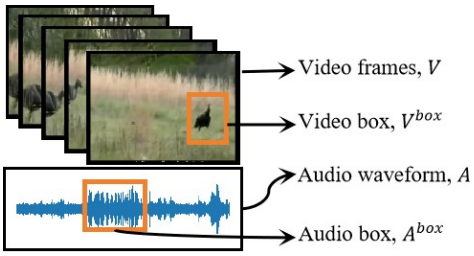

Figure 2: V and A denote the full visual and auditory inputs, while V^{box} and A^{box} denote visual and auditory boxes.

by compositional entailment learning (Pal et al., 2025). The local representation subsumes the full representation. We aim to preserve the event cues in the full representation while allowing it to vary with the surrounding context.

Motivated by this insight, we propose Cue-to-Context Subsumption Learning (CCSL), a probabilistic framework that uses localized audio and visual cues to guide the representation of the full audio-visual sample. Specifically, we first extract visual and auditory regions that are likely to contain recognizable event cues, as shown in Figure 2. We then map these localized cues and the full sample into Gaussian distributions parameterized by their means and variances. On this basis, we introduce a subsumption objective between the cue distribution and the contextualized full distribution. This objective aligns the mean of the full distribution with that of the cue distribution and constrains the variance of the full distribution relative to that of the cue distribution. In this way, the full representation is encouraged to preserve the event cues identified from localized regions while retaining contextual information useful for recognizing the complete audio-visual event.

In summary, our main contributions are as follows:

- We introduce a cue-to-context view for AV-GZSL. Under this view, localized audio and visual cues provide a reference for learning representations of full samples across different contexts, while the contextual information in each sample is retained.
- We propose Cue-to-Context Subsumption Learning (CCSL), which represents localized cues and full samples as Gaussian distributions. CCSL introduces a subsumption objective that aligns their means and constrains the variance of the full distribution relative to that of the cue distribution.
- CCSL establishes new state-of-the-art performance across VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL, yielding Harmonic Mean gains ranging from 5.75% to 7.84% over previous baselines.

2 Related Works

2.1 Audio-Visual Learning

Audio-visual learning aims to model the correspondence and complementary information between acoustic signals and visual scenes. Existing research covers several related tasks. Audio-visual event localization (Liu et al., 2025a; Zhou et al., 2024; 2025; Cheng et al., 2022; 2024) identifies the temporal boundaries and categories of sounding events in unconstrained videos. Audio-Visual retrieval (Araujo et al., 2025; Ryu et al., 2025; Zhu et al., 2024; Tang et al., 2026; Luo et al., 2026a; Mustafa et al., 2022) learns a shared embedding space to retrieve relevant visual frames or audio clips from multimodal queries. Audio-visual sound source localization (Um et al., 2025; Park & Kim, 2026) associates acoustic signals with their corresponding sounding objects or visual regions. Audio-visual segmentation (Guo et al., 2024; Liu et al., 2023; Luo et al., 2026b; Shen et al., 2026) further establishes pixel-level correspondence by segmenting sounding objects in complex visual scenes. In this work, we study audio-visual generalized zero-shot learning, which requires models to recognize both seen and unseen event categories without using training samples from unseen categories.

2.2 AUDIO-VISUAL GENERALIZED ZERO-SHOT LEARNING

Early AV-GZSL methods learn joint audio-visual-text representations through different alignment strategies. CJME (Parida et al., 2020) and AVGZSLNet (Mazumder et al., 2021) use coordinated embeddings and cross-modal reconstruction, while AVCA (Mercea et al., 2022b) and TCaF (Mercea et al., 2022a) employ attention mechanisms for multimodal fusion and temporal modeling. Hypermultiple (Hong et al., 2023) introduces hyperbolic geometry to model hierarchical structures. With pretrained multimodal models, ClipClap-GZSL (Kurzendörfer et al., 2024) uses CLIP (Radford et al., 2021) and CLAP (Wu et al., 2023) features. EZ-AVGZL (Mo & Morgado, 2024), EZ-AVOOD (Liu et al., 2025b), LSC-AVZSL (Li et al., 2025b), and SACMA (Yang et al., 2025) further improve category embeddings, the separation of seen and unseen samples, hyperbolic fusion, and cross-modal relations, respectively. More recent methods, including Spiking Temporal-Enhanced Network (Wang et al., 2026), Semantic-Guided Pseudo-Feature Attention Network (Ma et al., 2026a), AHSE (Zhang et al., 2026), and FAMAV (Ma et al., 2026b), continue to refine temporal modeling and multimodal alignment. Other approaches focus more directly on event-related information. MDST++ (Li et al., 2025a) decouples contextual and motion information, DAAN (Yu et al., 2025) suppresses redundant cues, and OAIA (Dong et al., 2024) augments multimodal representations with object-aware images. Although these methods improve feature extraction and fusion, their objectives do not explicitly model the subsumption relation between localized event cues and corresponding full representations. To address this gap, CCSL uses directional subsumption to guide full representations with localized event cues.

3 METHOD

3.1 PROBLEM FORMULATION

In the AV-GZSL task, the training set belonging to the seen categories $\mathcal{S}$, is defined as $\mathcal{D}^{train} = \{(v_i, a_i, t_i, y_i)\}_{i=1}^{N}$, where v_i and a_i represent the visual and audio features, t_i is the textual category descriptions, and y_i is the corresponding class label. N is the number of samples in training set. The test set comprises samples from both seen categories $\mathcal{S}$ and unseen categories $\mathcal{U}$, where $\mathcal{S} \cap \mathcal{U} = \emptyset$. It is defined as $\mathcal{D}^{test} = \{(v_j, a_j, t_j, y_j)\}_{j=1}^{M}$, where M is the number of samples in test set. The AV-GZSL goal is to learn a multi-modal mapping function $f : (v_i, a_i) \rightarrow t_i$, which can then be applied to test samples $f(v_j, a_j) = t_j$. In this work, the localized audio-visual cues are denoted as the box input $\mathbf{x}_i^{\text{box}}$, while the complete audio-visual sample is denoted as the full input $\mathbf{x}_i^{\text{full}}$.

3.2 AUDIO-VISUAL BOX EXTRACTION

3.2.1 VISUAL BOX EXTRACTION

As shown in Figure 3, for each training sample, we first select a keyframe from uniformly sampled video frames using the variance of the Laplacian as the sharpness score. Given the known category label, LLaMA-3 (Grattafiori et al., 2024) generates a set of foreground object nouns associated with the event. These nouns are used as open-vocabulary prompts for Florence-2 (Xiao et al., 2024) to detect and crop candidate regions from the selected keyframe. We then encode each candidate region and the full audio track with ImageBind (Girdhar et al., 2023), and select the region with the highest cross-modal cosine similarity as the visual box $\mathbf{v}_i^{\text{box}}$. This procedure is used only to construct training boxes and is not required during inference. For further details regarding the implementation of LLaMa prompts,please refer to Appendix B.

3.2.2 AUDITORY BOX EXTRACTION

The full audio track is first resampled to 48 kHz. We then use short-time root-mean-square (RMS) energy to identify non-silent intervals, retaining intervals whose energy is within 25 dB of the maximum, as shown in Figure 3. Intervals shorter than 0.5 seconds are discarded to remove brief transients. Each remaining audio segment and the selected keyframe are encoded by ImageBind (Girdhar et al., 2023). The segment with the highest cross-modal cosine similarity to the keyframe is selected as the auditory box $\mathbf{a}_i^{\text{box}}$.

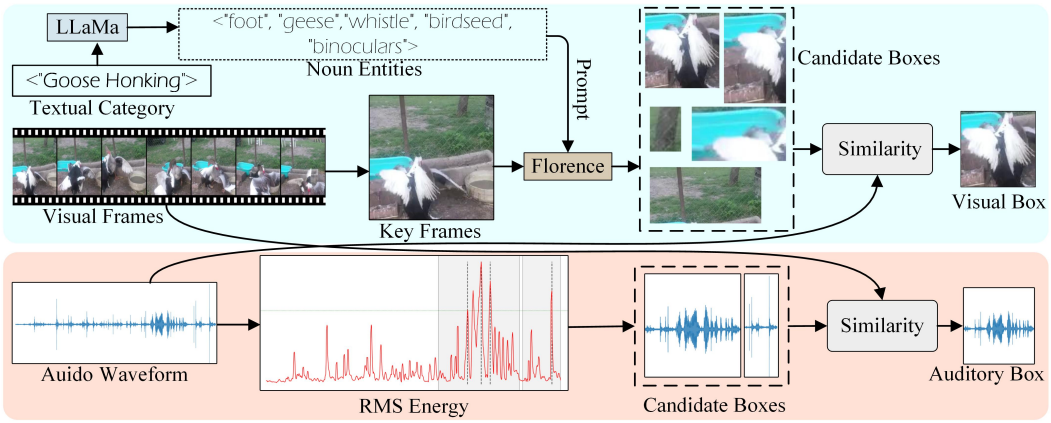

Figure 3: Visual box is grounded via LLM-generated entities and open-vocabulary detection, while auditory box is isolated via energy-based segmentation, both finalized via cross-modal similarity.

3.3 Cue-to-Context Subsumption Modeling

3.3.1 Distributional Feature Mapping

Given the box input $\mathbf{x}_i^{\text{box}}$ and the full input $\mathbf{x}_i^{\text{full}}$, we use the same Distribution Mapper to parameterize each input as a diagonal Gaussian distribution, as shown in Figure 4. For each input, the concatenated audio and visual features are first projected to 512 dimensions and then passed through two fully connected layers with LayerNorm, GELU, and dropout. A residual connection combines the outputs of the two layers to produce the hidden feature $\mathbf{h}_i^r$, where $r \in \{\text{box}, \text{full}\}$. The hidden feature is then processed by two separate LayerNorm operations and linear heads to predict the mean and standard deviation:

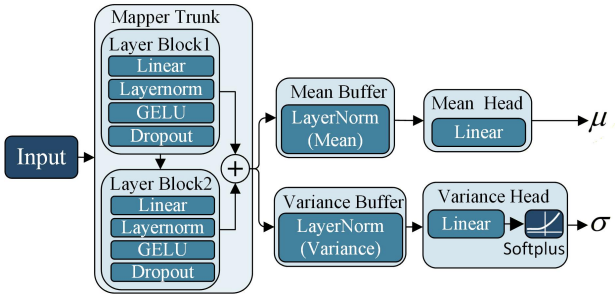

Figure 4: Architecture of the Distribution Mapper.

$$\boldsymbol{\mu}_i^r = \mathbf{W}_\mu \, \mathrm{LN}_\mu(\mathbf{h}_i^r) + \mathbf{b}_\mu, \qquad \boldsymbol{\sigma}_i^r = \mathrm{Softplus}\left(\mathbf{W}_\sigma \, \mathrm{LN}_\sigma(\mathbf{h}_i^r) + \mathbf{b}_\sigma\right) + \epsilon, \tag{1}$$

where $\epsilon = 10^{-6}$ ensures numerical stability and Softplus guarantees that every element of $\boldsymbol{\sigma}_i^r$ is positive. The resulting Gaussian embedding is defined as

$$q_i^r = \mathcal{N}\left(\boldsymbol{\mu}_i^r, \mathrm{diag}\left((\boldsymbol{\sigma}_i^r)^2\right)\right), \qquad r \in \{\text{box}, \text{full}\}. \tag{2}$$

Accordingly, the mapper produces $\boldsymbol{\mu}_i^{\text{box}}, \boldsymbol{\sigma}_i^{\text{box}}$ for the box input and $\boldsymbol{\mu}_i^{\text{full}}, \boldsymbol{\sigma}_i^{\text{full}}$ for the full input.

3.3.2 Subsumption Learning

In our method, we establish a directional subsumption relation between the box distribution and the full distribution. The box retains the localized event cues, while the full sample presents these cues under a particular context. We therefore regard the box distribution as the subsuming distribution and the full distribution as its contextualized realization. In the Gaussian space, this relation requires their means to remain close, while the box distribution is expected to maintain a broader spread than the full distribution. A straightforward approach might be applying a deterministic distance metric, such as Mean Squared Error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{k=1}^{d} \left[\left(\mu_{i,k}^{\text{full}} - \mu_{i,k}^{\text{box}}\right)^2 + \left(\sigma_{i,k}^{\text{full}} - \sigma_{i,k}^{\text{box}}\right)^2 \right], \tag{3}$$

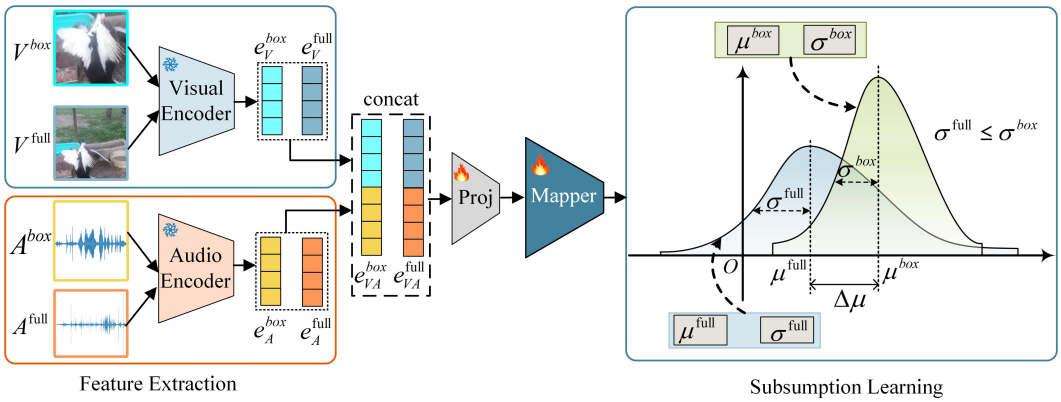

Figure 5: Overview of CCSL. Frozen encoders extract the box and full audio-visual features, e_{VA}^{box} and e_{VA}^{full}. The trainable projection and mapping modules parameterize them as $q^{\text{box}} = \mathcal{N}(\mu^{\text{box}}, (\sigma^{\text{box}})^2)$ and $q^{\text{full}} = \mathcal{N}(\mu^{\text{full}}, (\sigma^{\text{full}})^2)$. CCSL reduces $\Delta\mu$ and constrains σ^{full} relative to σ^{box}, modeling q^{full} as a contextual realization subsumed by q^{box}.

where N is the number of training samples and d is the dimension of the distributional parameters. However, this metric treats the two distributions symmetrically and directly pushes their standard deviations towards equality. The directional structure expected by subsumption is therefore lost. We instead build our objective upon the Kullback-Leibler divergence. Specifically, we measure the directional divergence from the full distribution to the box distribution:

$$\mathcal{L}_{\text{KL}} = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}\left(q_i^{\text{full}} \parallel q_i^{\text{box}}\right). \tag{4}$$

This direction is not interchangeable. The box distribution serves as the reference distribution, while the full distribution is constrained with respect to it. Nevertheless, directly minimizing this divergence still drives the two variances towards equality. For this reason, we introduce a margin $m > 1$ into the full variance, namely

$$q_i^{\text{full}} = \mathcal{N}\left(\boldsymbol{\mu}_i^{\text{full}}, m\,\text{diag}\left((\boldsymbol{\sigma}_i^{\text{full}})^2\right)\right). \tag{5}$$

The resulting Cue-to-Context Subsumption loss is formulated as

$$\mathcal{L}_{\text{CCS}} = \frac{1}{2Nd}\sum_{i=1}^{N}\sum_{k=1}^{d}\left[\log\frac{(\sigma_{i,k}^{\text{box}})^2}{m(\sigma_{i,k}^{\text{full}})^2} + \frac{m(\sigma_{i,k}^{\text{full}})^2 + \left(\mu_{i,k}^{\text{full}} - \mu_{i,k}^{\text{box}}\right)^2}{(\sigma_{i,k}^{\text{box}})^2} - 1\right]. \tag{6}$$

This objective draws the two means together and encourages the variance relation. The box distribution therefore preserves a broader range around the shared event cues, while the full distribution describes one realization under the context of the current sample. In this sense, the scaled directional divergence provides a probabilistic surrogate for cue-to-context subsumption. Finally, we introduce a variance regularization term to prevent the box distribution from expanding without control:

$$\mathcal{L}_{\text{reg}} = \frac{\alpha}{Nd}\sum_{i=1}^{N}\sum_{k=1}^{d}(\sigma_{i,k}^{\text{box}})^2, \tag{7}$$

where the regularization weight α is empirically set to 0.05.

3.3.3 Overall Learning Objective

Let $\mathbf{z}_c^t$ denote the projected textual feature of category c. Since recognition is performed from the full sample during inference, we use the full mean $\boldsymbol{\mu}_i^{\text{full}}$ to compute the category prediction. Specifically,

the Cross-Entropy loss is formulated as

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left((\boldsymbol{\mu}_i^{\mathrm{full}})^\top \mathbf{z}_{y_i}^t\right)}{\sum_{c\in\mathcal{S}}\exp\left((\boldsymbol{\mu}_i^{\mathrm{full}})^\top \mathbf{z}_c^t\right)}, \tag{8}$$

where $\mathcal{S}$ is the set of seen categories and y_i is the ground-truth label of the i-th training sample. Meanwhile, we align the box mean with the textual feature of its corresponding category:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{Nd}\sum_{i=1}^{N}\left\|\boldsymbol{\mu}_i^{\mathrm{box}} - \mathbf{z}_{y_i}^t\right\|_2^2, \tag{9}$$

where d is the feature dimension. This alignment anchors the localized event cues to their textual category, while the subsumption objective connects them to the corresponding full representation. Finally, we aggregate four objectives, namely the category prediction loss, box alignment loss, Cue-to-Context Subsumption loss, and variance regularization:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{CCS}} + \mathcal{L}_{\mathrm{reg}}. \tag{10}$$

All four terms are averaged over their corresponding samples and feature dimensions before summation.

During inference, only the full audio and visual features are required. We obtain $\boldsymbol{\mu}_j^{\mathrm{full}}$ from the full input and compare it with the textual features of all candidate categories. Under the AV-GZSL setting, the prediction is given by

$$\widehat{y}_j = \arg\min_{c\in\mathcal{S}\cup\mathcal{U}}\left[\left\|\boldsymbol{\mu}_j^{\mathrm{full}} - \mathbf{z}_c^t\right\|_2 + \beta\mathbb{I}(c\in\mathcal{S})\right], \tag{11}$$

where β is the calibrated stacking penalty applied to seen categories. Neither the box inputs nor the predicted standard deviations are required during inference.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed method on three benchmarks. VGGSound-GZSL (Mercea et al., 2022b) focus on diverse environmental audio-visual events, consisting of 93,752 videos across 276 classes, split into 138/69/69 for training, validation, and testing, respectively. UCF-GZSL (Mercea et al., 2022b) is an action-oriented benchmark comprising 6,816 clips from 51 classes, with a split of 30/12/9. ActivityNet-GZSL (Mercea et al., 2022b) covers complex daily human activities, containing approximately 20,000 videos spanning 200 categories, with a split of 99/51/50.

Evaluation Metrics. We adopt the mean class accuracy as the primary metric to evaluate performance. For the GZSL setting, we report the mean class accuracy on seen classes, denoted as S, and unseen classes, denoted as U. We compute the Harmonic Mean (HM) as the comprehensive performance indicator, formulated as:

$$HM = \frac{2\times S\times U}{S+U}. \tag{12}$$

Additionally, we report the standard Top-1 accuracy for the conventional ZSL setting, where the test search space is restricted to unseen classes.

Implementation Details. Following established protocols, we adopt a two-stage training strategy. In the first stage, the model is trained for at most 10 epochs, and the best epoch is selected on the validation set. In the second stage, we retrain the model on the union of the training and validation sets for the selected number of epochs. We evaluate three feature configurations, namely C3D (Tran et al., 2015) and VGGish (Hershey et al., 2017), SeLaVi (Asano et al., 2020), and CLIP (Radford et al., 2021) and CLAP (Wu et al., 2023). The model is optimized using Adam with a batch size of 64. The initial learning rates are 1×10^{-4} for VGGSound-GZSL and ActivityNet-GZSL, and 7×10^{-5} for UCF-GZSL. We set the variance margin m to 2.0 and the regularization weight α to 0.05. The boxes are constructed offline from the training samples and are not required during inference. Finally, calibrated stacking (Chao et al., 2016) is employed to mitigate the bias towards seen categories. All experiments are conducted on a single NVIDIA A100 GPU. For more implementation details,please refer to Appendix A.

Table 1: Comparison with the state-of-the-art methods on VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL. The best results are **bolded** and the second-best are underlined.

Methods	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑
SeLaVi:												
CJME	8.69	4.78	6.17	5.16	26.04	8.21	12.48	8.29	5.55	4.75	5.12	5.84
AVGZSLNet	18.05	3.48	5.83	5.28	52.52	10.90	18.05	13.65	8.93	5.04	6.44	5.40
TCaF	9.64	5.91	7.33	6.06	**58.60**	21.74	31.72	24.81	18.70	7.50	10.71	7.91
AVCA	14.90	4.00	6.31	6.00	51.53	18.43	27.15	20.01	24.86	8.02	12.13	9.13
DAAN	12.98	6.61	8.76	8.38	52.38	23.49	32.44	31.09	22.13	10.36	14.11	11.09
CCSL (Ours)	**19.27**	**11.22**	**14.18**	**14.37**	56.48	**26.71**	**36.27**	**32.80**	**27.16**	**14.29**	**18.73**	**12.81**
C3D & VGGish:												
Hyper-multiple	15.62	6.00	8.67	7.31	**74.26**	35.79	48.30	52.11	**36.98**	9.60	15.25	10.39
SACMA	13.42	6.00	8.29	6.46	60.15	27.46	37.71	29.07	17.12	7.32	10.25	7.52
MDST++	12.21	8.38	9.94	**10.32**	65.32	44.79	53.14	**55.22**	27.53	10.72	15.43	13.12
CCSL (Ours)	**22.66**	**13.47**	**16.90**	15.42	68.09	**49.97**	**57.64**	54.56	35.69	**16.44**	**22.51**	**16.47**
CLIP & CLAP:												
AVCA	32.47	6.81	11.26	8.16	34.90	38.67	36.69	38.67	24.04	19.88	21.76	20.88
ClipClap	29.68	11.12	16.18	11.53	77.14	43.91	55.97	46.96	**45.98**	20.06	27.93	22.76
LSC-AVZSL	31.23	12.00	17.34	13.34	80.59	49.95	61.67	51.22	39.52	25.20	30.77	26.77
EZ-AVGZL	33.28	20.15	25.10	21.06	78.32	46.35	58.24	48.27	38.52	12.85	19.27	13.01
EZ-AVOOD	**39.33**	11.84	18.21	13.28	83.53	48.01	60.97	50.92	41.56	21.06	27.95	25.20
CCSL (Ours)	35.64	**30.63**	**32.94**	**39.44**	**84.03**	**56.30**	**67.42**	**61.26**	44.75	**30.89**	**36.55**	**32.07**

4.2 Comparison with State-of-the-Art Methods

Table 1 shows that CCSL achieves the best HM performance among the compared methods across all three benchmarks. Under CLIP and CLAP features, CCSL improves HM by 5.75 to 7.84 percentage points over the strongest baselines. This advantage remains consistent under the SeLaVi setting and the setting using C3D and VGGish features. Notably, CCSL falls slightly short in seen accuracy on VGGSound-GZSL. Its seen accuracy is 35.64, compared with 39.33 for EZ-AVOOD. However, its unseen accuracy reaches 30.63, substantially exceeding the 11.84 achieved by EZ-AVOOD. This contrast shows that the HM improvement mainly stems from stronger recognition of unseen categories. More broadly, CCSL achieves a better balance between seen and unseen recognition, where strong performance on seen categories can conceal a substantial bias against unseen categories. The consistent HM gains across different benchmarks and feature settings support the effectiveness of using localized event cues to guide full-sample representations.

4.3 The Impact of Audio-Visual Boxes

4.3.1 Analysis of Box Proportions

We analyze the proportions of the visual and auditory boxes relative to their corresponding full inputs across the three datasets. Specifically, the visual proportion is computed as the box area divided by the keyframe area, while the auditory proportion is computed as the box duration divided by the full audio duration. As shown in Figure 6, both modalities cover a broad range of proportions, with noticeable concentrations at smaller ratios and sharp peaks near 1.0. The smaller ratios indicate that the selected event cues often occupy only part of the full input. Meanwhile, the peaks near 1.0 show that some selected boxes retain the entire spatial or temporal extent. UCF-GZSL and ActivityNet-GZSL further exhibit more visible concentrations in the intermediate range, particularly for the visual boxes.

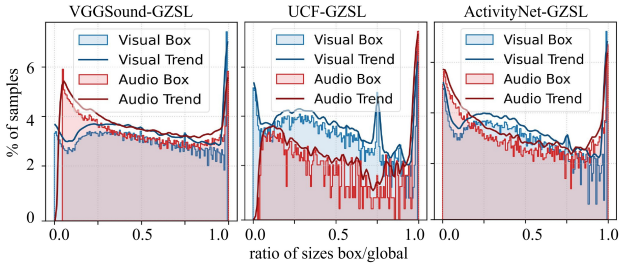

Figure 6: Distributions of the visual box-to-frame area ratio and auditory box-to-audio duration ratio across three datasets.

4.3.2 Impact of Additional Box Inputs

To investigate whether the performance gains merely come from the additional data provided by the boxes, we conduct a plug-and-play experiment on VGGSound-GZSL under the CLIP and CLAP setting. For each training instance, its paired visual and auditory boxes are assigned the same category label and appended as an additional audio-visual training sample, without modifying the architecture or learning objective of the baseline. As shown in Table 2, this direct augmentation brings little improvement to either AVCA or ClipClap-GZSL. The absolute HM changes do not exceed 0.25 percentage points. Naively appending the boxes provides no explicit relation between the localized cues and their corresponding full samples. These results indicate that the improvement does not stem from the additional box data alone. Instead, it comes from explicitly modeling the relation between the box and full representations.

Table 2: Impact of incorporating boxes into baseline models.

Methods	w/ boxes	S ↑	U ↑	HM ↑	ZSL ↑
AVCA	✗	32.47	6.81	11.26	8.16
AVCA	✓	30.48	6.82	11.15	8.17
ClipClap	✗	29.68	11.12	16.18	11.53
ClipClap	✓	27.22	11.26	15.93	11.95
CCS(ours)	✓	**35.64**	**30.63**	**32.94**	**39.44**

4.3.3 Distributional Dynamics During Training

We trace the average variance of the full distribution and the L2 distance between the full and box means over 10 training epochs. The variance is computed by averaging $(\boldsymbol{\sigma}^{\text{full}})^2$ over the samples and feature dimensions in each mini-batch. For the comparison without CCSL, the box branch is retained only to compute the mean distance and does not contribute to model optimization. As illustrated in Figure 7, without CCSL, the average full variance increases continuously during training. Conversely, CCSL rapidly reduces the variance and maintains it at a stable level. Although the mean distance increases in both settings after an initial decline, it remains consistently smaller with CCSL. These results show that CCSL regulates the distributional scale of the full representation and keeps its mean closer to the box mean during training.

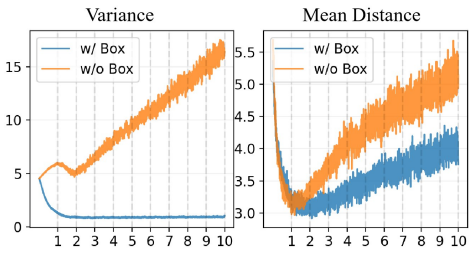

Figure 7: Comparison of the average full variance and box-to-full mean distance.

4.4 Qualitative Analysis

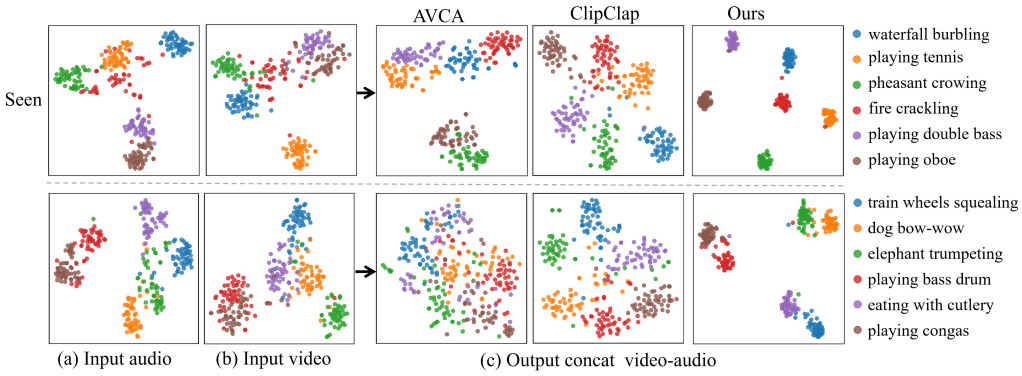

Figure 8: t-SNE visualizations of audio-visual representations on VGGSound-GZSL.

We further provide t-SNE visualizations on VGGSound-GZSL to compare the audio-visual representations learned by AVCA, ClipClap-GZSL, and CCSL. We sample up to 50 instances from each of six seen and six unseen categories. For CCSL, we visualize the full-sample mean representations, namely $\boldsymbol{\mu}^{\text{full}}$. As shown in Figure 8(a) and (b), the input audio and visual features exhibit substantial overlap among categories. After projection, AVCA and ClipClap-GZSL form partially separated

Table 3: Ablation study of visual and auditory boxes on three benchmarks.

Settings	V-Box	A-Box	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
			S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑
Global Only	✗	✗	34.17	19.63	24.94	20.82	79.82	44.91	57.48	47.16	39.61	18.38	25.11	22.63
w/ Visual Box	✓	✗	34.83	24.91	29.05	29.87	82.63	51.48	63.43	54.89	42.58	23.67	30.43	26.41
w/ Auditory Box	✗	✓	34.88	24.78	28.98	29.74	81.15	48.37	60.61	51.28	40.73	19.48	26.36	22.65
Ours (Full)	✓	✓	**35.64**	**30.63**	**32.94**	**39.44**	**84.03**	**56.30**	**67.42**	**61.26**	**44.75**	**30.89**	**36.55**	**32.07**

clusters for seen categories, while their unseen-category representations remain strongly overlapped. In contrast, CCSL produces more compact clusters with clearer separation for both seen and unseen categories. The visualization shows that CCSL learns a more structured audio-visual representation space, particularly for unseen categories.

4.5 Ablation Studies

4.5.1 Audio-visual Boxes

To examine the contribution of each box modality, we replace the removed box feature with the corresponding feature from the full input, while keeping the model architecture and learning objectives unchanged. As shown in Table 3, using both boxes consistently achieves the best HM and ZSL results on all three benchmarks. Removing the visual box causes a larger HM decrease on UCF-GZSL and ActivityNet-GZSL, whereas removing either box produces comparable degradation on VGGSound-GZSL. Moreover, the improvements in unseen accuracy are consistently larger than those in seen accuracy. These results show that the visual and auditory boxes provide complementary event cues, with the visual box contributing more strongly on UCF-GZSL and ActivityNet-GZSL.

4.5.2 Training Objectives

We conduct eight ablation experiments on VGGSound-GZSL to examine the contribution of each training objective. As shown in Table 4, the pretrained-only setting yields near-zero performance under the evaluated framework. The cross-entropy loss $\mathcal{L}_{\text{CE}}$ establishes the basic task-specific baseline, and adding the alignment loss $\mathcal{L}_{\text{align}}$ improves HM from 20.93 to 24.86. Directly adding the variance regularization $\mathcal{L}_{\text{reg}}$ reduces HM to 22.58, indicating that controlling the box variance alone is insufficient without modeling its relation to the full distribution. Replacing the proposed subsumption objective with the symmetric parameter-matching loss $\mathcal{L}_{\text{MSE}}$ brings only marginal changes, even when combined with $\mathcal{L}_{\text{reg}}$. In contrast, $\mathcal{L}_{\text{CCS}}$ raises HM to 30.75, mainly by increasing unseen accuracy from 19.82 to 27.55. Finally, combining $\mathcal{L}_{\text{CCS}}$ with $\mathcal{L}_{\text{reg}}$ achieves the best HM of 32.94 and ZSL accuracy of 39.44. These results show that the improvement comes from modeling the directional relation between the box and full distributions.

Table 4: Ablation study on different training objectives.

Training Objectives	S ↑	U ↑	HM ↑	ZSL ↑
None (Pre-trained Only)	0.32	0.41	0.36	1.28
$\mathcal{L}_{CE}$	25.28	17.85	20.93	21.06
$\mathcal{L}_{CE} + \mathcal{L}_{align}$	33.35	19.82	24.86	26.15
$\mathcal{L}_{CE} + \mathcal{L}_{align} + \mathcal{L}_{reg}$	30.40	17.96	22.58	25.20
$\mathcal{L}_{CE} + \mathcal{L}_{align} + \mathcal{L}_{MSE}$	33.25	19.88	24.88	26.10
$\mathcal{L}_{CE} + \mathcal{L}_{align} + \mathcal{L}_{MSE} + \mathcal{L}_{reg}$	33.47	20.01	25.05	27.85
$\mathcal{L}_{CE} + \mathcal{L}_{align} + \mathcal{L}_{CCS}$	<u>34.80</u>	<u>27.55</u>	<u>30.75</u>	<u>35.60</u>
$\mathcal{L}_{CE} + \mathcal{L}_{align} + \mathcal{L}_{CCS} + \mathcal{L}_{reg}$	**35.64**	**30.63**	**32.94**	**39.44**

5 Conclusion

In this work, we propose Cue-to-Context Subsumption Learning (CCSL) for AV-GZSL. CCSL models the box and full inputs as Gaussian distributions and introduces a directional subsumption objective to guide the full representation with localized event cues. Experiments on three benchmarks show consistent improvements in HM and unseen accuracy, while further analyses verify the contributions of the visual and auditory boxes and the proposed objective. However, CCSL depends on foundation models for visual box extraction and remains sensitive to box scale, since inaccurate or improperly sized boxes may weaken the subsumption relation or cause variance collapse. The multi-stage extraction process also introduces additional computation, motivating future work on more robust and efficient box extraction.

REFERENCES

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18794–18803, 2025.

Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pp. 52–68. Springer, 2016.

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.

Ying Cheng, Yang Li, Junjie He, and Rui Feng. Mixtures of experts for audio-visual learning. *Advances in Neural Information Processing Systems*, 37:219–243, 2024.

Yujie Dong, Shiming Chen, Bowen Duan, Weiping Ding, Yisong Wang, and Xinge You. Object-aware image augmentation for audio-visual zero-shot learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *Proceedings of the 32nd ACM International conference on multimedia*, pp. 7533–7541, 2024.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Jie Hong, Zeeshan Hayder, Junlin Han, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Hyperbolic audio-visual zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7873–7883, 2023.

David Kurzendörfer, Otniel-Bogdan Mercea, A Koepke, and Zeynep Akata. Audio-visual generalized zero-shot learning using pre-trained large multi-modal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2627–2638, 2024.

Wenrui Li, Penghong Wang, Xingtao Wang, Wangmeng Zuo, Xiaopeng Fan, and Yonghong Tian. Multi-timescale motion-decoupled spiking transformer for audio-visual zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Xiaoyong Li, Jing Yang, Yuling Chen, Wei Zhang, Xiaoli Ruan, Chengjiang Li, and Zhidong Su. Learning semantic consistency for audio-visual zero-shot learning. *Artificial Intelligence Review*, 58(7):212, 2025b.

Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7590–7598, 2023.

Liang Liu, Shuaiyong Li, and Yongqiang Zhu. Audio-visual semantic graph network for audio-visual event localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23957–23966, 2025a.

Yang Liu, Xun Zhang, Jiale Du, Xinbo Gao, and Jungong Han. Extremely simple out-of-distribution detection for audio-visual generalized zero-shot learning. *arXiv preprint arXiv:2503.22197*, 2025b.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 147298–147334, 2026a.

Ziyang Luo, Nian Liu, Fahad Shahbaz Khan, and Junwei Han. Aurora: Augmented understanding via structured reasoning and reinforcement learning for reference audio-visual segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7716–7724, 2026b.

Siteng Ma, Wenrui Li, Haocheng Tang, Yeyu Chai, Jisheng Chu, and Xingtao Wang. Semantic-guided pseudo-feature attention network for audio-visual zero-shot learning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12497–12501. IEEE, 2026a.

Siteng Ma, Xiaoyu Niu, Haocheng Tang, Zhe Yang, Jisheng Chu, and Wenrui Li. Fusion-regularized alignment modality-adaptive audio-visual network for audio-visual zero-shot learning. *Neurocomputing*, pp. 133693, 2026b.

Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3090–3099, 2021.

Otniel-Bogdan Mercea, Thomas Hummel, A Sophia Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European conference on computer vision*, pp. 488–505. Springer, 2022a.

Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10553–10563, 2022b.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Shentong Mo and Pedro Morgado. Audio-visual generalized zero-shot learning the easy way. In *European Conference on Computer Vision*, pp. 377–395. Springer, 2024.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025.

Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3251–3260, 2020.

Subin Park and Jung Uk Kim. Generate, analyze, and refine: Training-free sound source localization via mllm meta-reasoning. *arXiv preprint arXiv:2604.06824*, 2026.

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.

Hyeonggon Ryu, Seongyu Kim, Joon Son Chung, and Arda Senocak. Seeing speech and sound: Distinguishing and locating audio sources in visual scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13540–13549, 2025.

Li Shen, Yangzhu Wang, Xuyi Fan, Yuning Wei, and Huaxin Qiu. Av2ts: A multivariate time series modeling framework for audio-visual segmentation. *IEEE Transactions on Multimedia*, 2026.

Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. Wave: learning unified & versatile audio-visual embeddings with multimodal llm. In *International Conference on Learning Representations*, volume 2026, pp. 61596–61612, 2026.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Object-aware sound source localization via audio-visual scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8342–8351, 2025.

Ziyu Wang, Wenrui Li, Hongtao Chen, Jisheng Chu, Hengyu Man, and Xiaopeng Fan. Spiking temporal-enhanced network for zero-shot audio-visual learning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9137–9141. IEEE, 2026.

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

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Jing Yang, Xiaoyong Li, Yuankai Wu, Yuling Chen, Xiaoli Ruan, Chengjiang Li, and Qing Hou. Self-attention and cross-modal attention for audio-visual zero-shot learning. *Information Fusion*, pp. 103966, 2025.

Runlin Yu, Yipu Gong, Wenrui Li, Aiwen Sun, and Mengren Zheng. Discrepancy-aware attention network for enhanced audio-visual generalized zero-shot learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1112–1121, 2025.

Zihan Zhang, Jie Hong, Siyuan Fan, Yanghao Zhou, and Pengfei Fang. On aligning hierarchical standardized embedding for audio-visual generalized zero-shot learning. *arXiv preprint arXiv:2606.11602*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2025.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 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.

A Detailed Experimental Setups

A.1 Feature Dimensions

We employ three distinct sets of audio-visual encoders. Since these off-the-shelf models produce embeddings of varying sizes, the input dimensions for the concatenated audio-visual features and the textual category embeddings differ accordingly.

CLIP & CLAP: The visual features from CLIP are 512-D. The auditory features from CLAP are 512-D, resulting in a concatenated audio-visual input of 1024-D. The corresponding textual embeddings extracted via the text encoders are also 1024-D (i.e., 512-D + 512-D).

C3D & VGGish: The visual features extracted by C3D are 4096-D. The auditory features from VGGish are 512-D, resulting in a concatenated audio-visual input of 4608-D. The corresponding textual embeddings extracted via word2vec Mikolov et al. (2013) are 600-D (i.e., 300-D + 300-D).

SeLaVi: Both the visual and auditory features are 512-D, resulting in a concatenated audio-visual input of 1024-D. Similar to the C3D & VGGish setup, the textual embeddings are 600-D.

A.2 Network Architecture

Regardless of the initial input dimensions, the raw concatenated audio-visual features are first passed through a linear projection block. This block consists of Fully Connected layers combined with Layer Normalization and Dropout. The hidden dimension is generally set to 512 for CLIP&CLAP and SeLaVi, and 1024 for C3D&VGGish. To prevent overfitting on the large-scale VGGSound dataset, these dimensions are reduced to 128 and 256. These projected features are unified to a dimension of 512 and then fed into the distribution mapper, which ultimately outputs the parameters with a dimension of 64.

A.3 Hyperparameter Settings

For the large-scale VGGSound-GZSL, the dropout rate is set to 0.2. For the relatively smaller datasets, UCF-GZSL and ActivityNet-GZSL, the dropout rate is set to 0.1. We optimize the network using the Adam optimizer with weight decay. We employ a ReduceLROnPlateau scheduling strategy. We monitor the evaluation metric, and if the performance does not improve for 3 epochs, the learning rate is multiplied by a factor of 0.1.

B Details of Audio-Visual Box Extraction

We leverage LLaMA-3 Grattafiori et al. (2024) to generate a set of candidate noun entities for each category. As illustrated in Figure 9, for VGGSound-GZSL, we design a prompt to instruct the model: "Analyze the sound event of <**textual category**>. List the 1 to 4 most essential, visible physical objects that act as the direct sound source or are physically interacting to produce the sound." For UCF-GZSL, the prompt is adapted as "Analyze a video of a person performing the fine-grained action <textual category>. List the 2 to 4 most essential, discriminative physical objects or props that the person is directly manipulating." For ActivityNet-GZSL, the prompt is "Analyze the complex human activity of <textual category>. List the 3 to 6 most essential, discriminative physical objects that are actively used or interacted with during this activity."

To guarantee that the LLM outputs isolated noun entities rather than full descriptive sentences, we implement an automated validation and retry mechanism. We parse the generated text and reject any output containing overly long phrases. Upon a validation failure, the mechanism increases the sampling temperature and regenerates the response until a valid list of nouns is obtained.

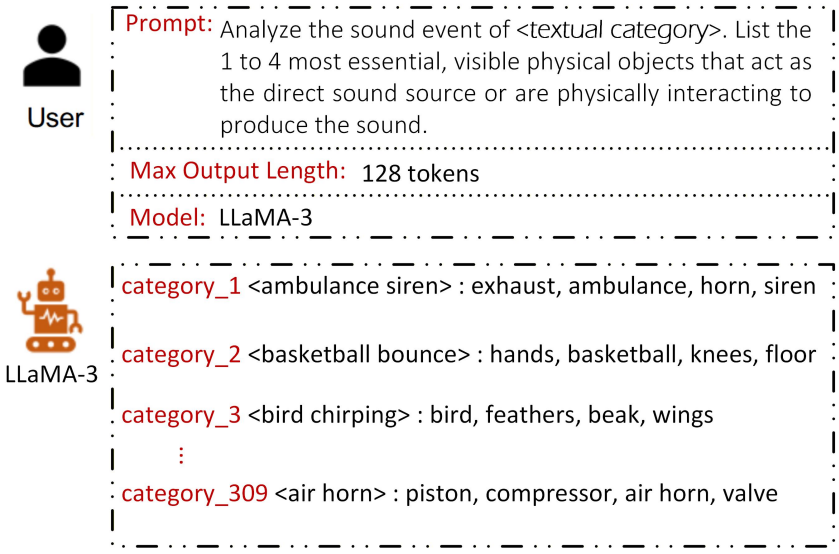

Figure 9: Illustration of the VGGSound-GZSL prompt design and generated noun entities using LLaMA-3.

C Additional Experimental Analyses

C.1 Hyperparameter Analysis

The margin m is a critical parameter controlling the intensity of the subsumption. We conduct extensive ablation experiments on it, as illustrated in Figure 10. We observe that when m ranges from 2.0 to 10.0, the HM metric fluctuates around its optimal peak. This indicates that compressing the global variance into a certain range is acceptable. Its exact numerical boundaries require further investigation. The proportion of core events is not fixed, but rather varies within a certain range. However, as we continue to increase m, the model experiences a sharp decline. This suggests that m exceeds the acceptable range, forcing the model to fixate on a concentrated point.

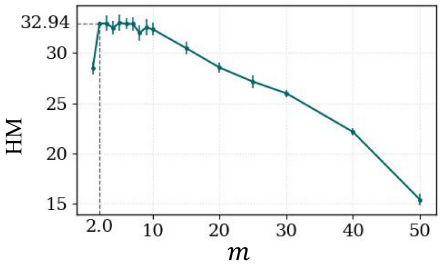

Figure 10: Hyperparameter analysis of margin m.

C.2 Additional Ablation Studies

The regularization weight α is used to constrain the box variance. Table 5 reports the sensitivity analysis of α on the VGGSound-GZSL dataset. The model achieves the best performance when α is set to 0.05. A smaller α provides insufficient regularization, allowing the box distribution to become excessively broad and weakening the subsumption constraint. In contrast, a larger α imposes a stronger penalty on the variance and forces the mapped distributions to become overly concentrated. This restricts their ability to capture the possible contextual variations associated with the localized event cues, thereby impairing the zero-shot generalization capability of the model. These results show that $\alpha = 0.05$ provides an appropriate balance between maintaining sufficient distributional coverage and avoiding excessive uncertainty.

Table 5: Ablation of α on VGGSound-GZSL.

α	VGGSound-GZSL			
	S	U	HM	ZSL
0.01	34.12	27.55	30.48	36.10
0.05	**35.64**	**30.63**	**32.94**	**39.44**
0.10	30.80	24.07	27.02	31.25
0.30	23.76	16.46	19.45	22.51

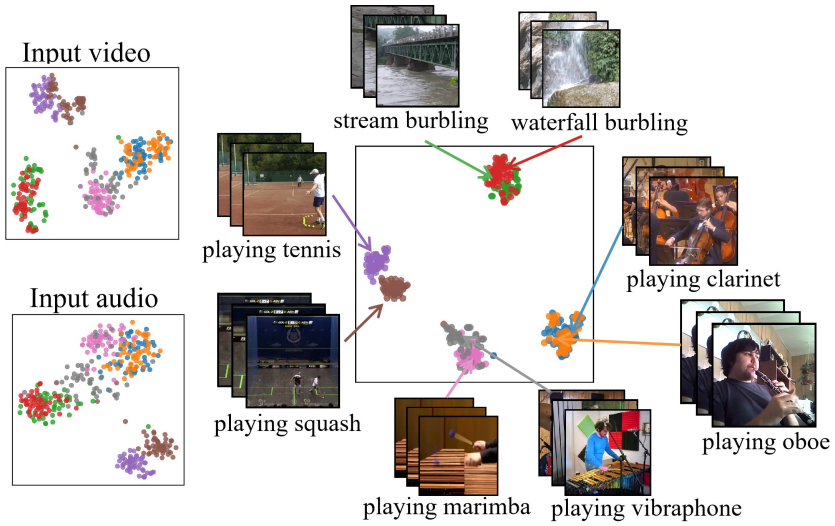

Figure 11: t-SNE visualization of highly similar categories.

C.3 Semantic Clustering of Related Categories

Figure 11 visualizes the features of highly similar categories using t-SNE. While the raw audio and video input features are loosely distributed, the learned joint semantic space exhibits compact intra-class clustering. However, this visualization also reveals a limitation of our method, as highly similar categories are mapped close to each other or even overlap without clear separation. This issue is caused by the adopted box subsumption constraints. Since similar categories possess similar boxes, the model is driven to group them together during training. This phenomenon also explains why the accuracy improvement on seen classes remains limited. Future work will focus on refining these constraints to enhance fine-grained inter-class separability.

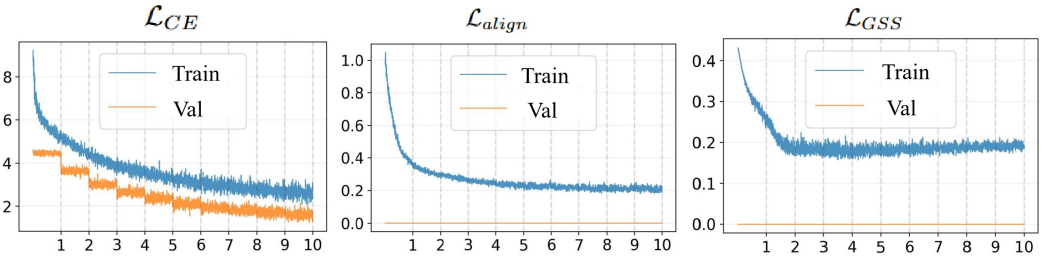

Figure 12: Convergence curves of the proposed loss functions..

C.4 Convergence Analysis

Convergence of Loss Functions. Figure 12 illustrates the convergence curves of loss functions. The cross-entropy loss $\mathcal{L}_{CE}$ exhibits a steady decrease on both the training and validation sets. Meanwhile, the training curves for the alignment loss $\mathcal{L}_{align}$ and the subsumption learning loss $\mathcal{L}_{CCS}$ quickly converge to a stable plateau. This demonstrates that our multimodal alignment and box subsumption constraints can be effectively optimized.

Convergence of Mean and Variance. Figure 13 illustrates the convergence of statistics parameters during training to verify the stability. We track the global variance and mean alongside their relative metrics to the box constraints. These include the variance ratio $\frac{\sigma^2}{(\sigma^{box})^2}$ and the mean distance $||\mu_{global} - \mu_{box}||$. As depicted in Figure 13(a), the global variance is constrained to a stable value to prevent representation collapse. The variance ratio shown in Figure 13(b) exhibits a steady decline and quickly plateaus. Furthermore, although the global mean in Figure 13(c) maintains an

upward trend as the feature centers evolve, the mean distance in Figure 13(d) demonstrates a decrease. This confirms that the feature distributions are converging towards their corresponding box targets without causing training instability.

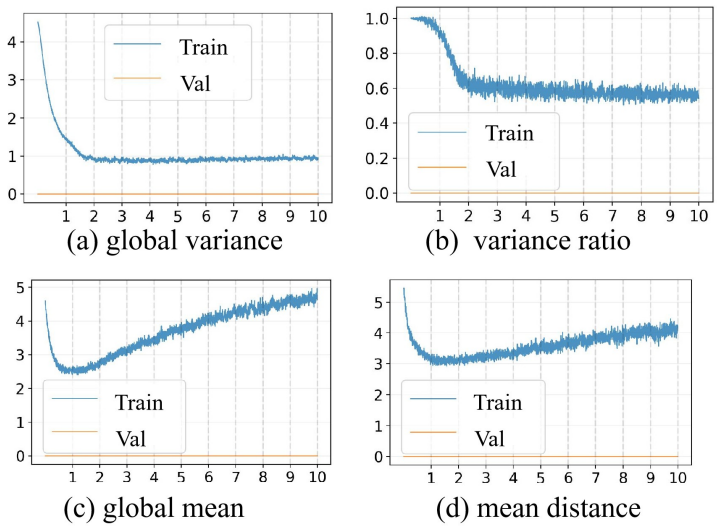

Figure 13: Convergence curves of statistical parameters during training.

D BROADER IMPACTS

This work reduces the reliance on human annotations for audio-visual recognition, facilitating cost-effective applications in multimedia retrieval, accessibility, and automated sensitive content monitoring on short-video platforms. However, since our framework relies on pre-trained foundation models such as LLaMA-3 and Florence, it may propagate their latent cultural biases. Furthermore, deploying such recognition systems must be strictly regulated to prevent privacy violations surveillance.